Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics

Daniel Sorensen
Daniel Gianola

Springer

Statistics for Biology and Health

Series Editors
K. Dietz, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Daniel Sorensen Daniel Gianola

Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics

Springer

Daniel Sorensen
Department of Animal Breeding
and Genetics
Danish Institute of Agricultural Sciences
DK-8830 Tjele
Denmark
sorensen@inet.uni2.dk

Daniel Gianola
Department of Animal Science
Department of Dairy Science
Department of Biostatistics and Medical
Informatics
University of Wisconsin-Madison
Madison, WI 53706
USA
gianola@calshp.cals.wisc.edu

Series Editors
K. Dietz
Institut für Medizinische Biometrie
Universität Tübingen
Westbahnhofstrasse 55
D-72070 Tübingen
GERMANY

M. Gail
National Cancer Institute
Rockville, MD 20892
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
FRANCE

J. Samet
School of Public Health
Department of Epidemiology
Johns Hopkins University
615 Wolfe Street
Baltimore, MD 21205-2103
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

Library of Congress Cataloging-in-Publication Data
Sorensen, Daniel.
Likelihood, Bayesian and MCMC methods in quantitative genetics / Daniel Sorensen, Daniel Gianola.
p. cm. — (Statistics for biology and health)
Includes bibliographical references and index.
ISBN 0-387-95440-6 (alk. paper)
1. Genetics—Statistical methods. 2. Monte Carlo method. 3. Markov processes. 4. Bayesian statistical decision theory. I. Gianola, Daniel, 1947– II. Title. III. Series.
QH438.4.S73 S675 2002
575.6′07′27—dc21 2002019555

ISBN 0-387-954406 Printed on acid-free paper.

© 2002 Springer-Verlag New York, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10866246

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface

Statistical genetics results from the merger of genetics and statistics into a coherent quantitative theory for predicting and interpreting genetic data. Based on this theory, statistical geneticists developed techniques that made notable contributions to animal and plant breeding practices in the second half of the last century as well as to advances in human genetics.

There has been an enormous research impetus in statistical genetics over the last 10 years. Arguably, this was stimulated by major breakthroughs in molecular genetics, by the advent of automatic data-recording devices, and by the possibility of applying computer-intensive statistical methods to large bodies of data with relative ease. Data from molecular biology and biosensors are characterized by their massive volume. Often, intricate distributions need to be invoked for appropriate modeling. Data-reduction techniques are needed for accounting for the involved nature of these data and for extracting meaningful information from the observations. Statistical genetics plays a major role in this process through the development, implementation, and validation of probability models for inference. Many of these models can be daunting and, often, cannot be fitted via standard methods. Fortunately, advances in computing power and computer-based inference methods are making the task increasingly feasible, especially in connection with likelihood and Bayesian inference.

Two important breakthroughs in computational statistics have been the bootstrap and Markov chain Monte Carlo (MCMC) methods. In this book we focus on the latter. MCMC was introduced into the statistical literature in the late 1980s and early 1990s, and incorporation and adaptation of the methods to the needs of quantitative genetic analysis was relatively rapid,

particularly in animal breeding. Also, MCMC is having a major impact in applied statistics (especially from a Bayesian perspective), opening the way for posing models with an enormous amount of flexibility. With MCMC, it is possible to arrive at better descriptions of the perceived underlying structures of the data at hand, free from the strictures of standard methods of statistical analysis.

The objective of this book is to present the main ideas underlying likelihood and Bayesian inference and MCMC methods in a manner that is accessible to numerate biologists, giving step-by-step derivations and fully worked-out examples. Most of these examples are from quantitative genetics and, although not exclusively, we focus on normal or generalized linear models.

Most students and researchers in agriculture, biology, and medicine lack the background needed for understanding the foundations of modern biometrical techniques. This book has been written with this particular readership in mind. A number of excellent books describing MCMC methods have become available in recent years. However, the main ideas are presented typically in a technically demanding style, as these books have been written by and addressed to statisticians. The statistician often has the mathematical background needed to "fill in the blanks". What is tedious detail to a statistician, so that it can be omitted from a derivation, can cause considerable consternation to a reader with a different background. In particular, biologists need a careful motivation of each model from a subject matter perspective, plus a detailed treatment of all the algebraic steps needed to carry out the analysis. Cavalier statements such as "it follows immediately", or "it is easy to show", are encountered frequently in the statistical literature and cause much frustration to biological scientists, even to numerate ones. For this reason, we offer considerably more detail in the developments than what may be warranted for a more mathematically apt audience. We do not apologize for this, and hope that this approach will be viewed sympathetically by the scientific community to which we belong. Nevertheless, some mathematical and statistical prerequisites are needed in order to be able to extract maximum benefit from the material presented in this book. These include a beginning course in differential and integral calculus, an exposure to elementary linear algebra (preferably with a statistical bent), an understanding of probability theory and of the concepts of statistical inference, and a solid grounding in the applications of mixed effects linear models. Most students of quantitative genetics and animal breeding acquire this preparation during the first two years of their graduate education, so we do not feel that the requirements are especially stringent. Some applied statisticians reading this book may be caught by the quantitative genetics jargon. However, we attempt to relate biological to statistical parameters and we trust that the meaning will become clear from the context.

The book is organized into four parts. Part I (Chapters 1 and 2) presents a review of probability and distribution theory. Random variables and their distributions are introduced and illustrated. This is followed by a discussion on functions of random variables. Applied and theoretical statisticians can skip this part of the book safely, although they may find some of the examples interesting.

The first part lays the background needed for introducing methods of inference, which is the subject of the seven chapters in Part II. Chapters 3 and 4 cover the classical theory of likelihood inference. Properties of the maximum likelihood estimator and tests of hypotheses based on the Neyman–Pearson theory are discussed. An effort has been made to derive, in considerable detail, many of the important asymptotic results and several examples are given. The problems encountered in likelihood inference under the presence of nuisance parameters are discussed and illustrated. Chapter 4 ends with a presentation of models for which the likelihood does not have a closed form. Bayesian inference is the subject of chapters 5-8 in Part II . Chapter 5 provides the essential ingredients of the Bayesian approach. This is followed by a chapter covering in fair detail the analysis of the linear model. Chapter 7 discusses the role of the prior distribution in Bayesian analysis. After a short tour of Bayesian asymptotics, the concepts of statistical information and entropy are introduced. This is followed by a presentation of Bayesian analysis using prior distributions conveying vague prior knowledge, perhaps the most contentious topic of the Bayesian paradigm. The chapter ends with an overview of a technically difficult topic called reference analysis. Chapter 8 deals briefly with hypothesis testing from a Bayesian perspective. Chapter 9, the final one of this second part, provides an introduction to the expectation–maximization (EM) algorithm, a topic which has had far-reaching influences in the statistical genetics literature. This algorithm is extremely versatile, and is so inextricable from the statistical structure of a likelihood or Bayesian problem that we opted to include it in this part of the book.

The first two parts of the book described above provide the basis for positing probability models. The implementation and validation of models via MCMC requires some insight on the subtleties on which this technique is based. This is presented in Part III, whose intent is to explain this remarkable computational tool, within the constraints imposed by the authors' limited mathematics. After an introduction to discrete Markov chains in Chapter 10, the MCMC procedures are discussed in a detailed manner in Chapter 11. An inquisitive reader should be able to follow the derivation of the acceptance probability of various versions of the celebrated Metropolis–Hastings algorithm, including reversible jump. An overview of methods for analyzing MCMC output is the subject of Chapter 12.

Part IV gives a presentation of some of the models that are being used in quantitative genetics at present. The treatment is mostly Bayesian and the models are implemented via MCMC. The classical Gaussian mixed

model for single- and multiple-trait analyses is described in Chapter 13. Extensions are given for robust analyses using t distributions. The Bayesian MCMC implementation of this robust analysis requires minor changes in a code previously developed for analyzing Gaussian models, illustrating the remarkable versatility of the MCMC techniques. Chapter 14 discusses analyses involving ordered categorical traits based on the threshold model of Sewall Wright. This chapter also includes a Bayesian MCMC description of a model for joint analysis of categorical and Gaussian responses. Chapter 15 deals with models for the analysis of longitudinal data, and the book concludes with Chapter 16, which introduces segregation analysis and models for the detection of quantitative trait loci.

Although this book can be used as a text, it cannot claim such status fully. A textbook requires carefully chosen exercises, and probably a more linear development than the one presented here. Hence, these elements will need to be provided by the instructor, should this book be considered for classroom use. We have decided not to discuss software issues, although some reasonably powerful public domain programs are already available. The picture in this area is changing too rapidly, and we felt that many of our views or recommendations in this respect would probably be rendered obsolete at the time of publication.

The book evolved from cooperation between the two authors with colleagues from Denmark and Wisconsin leading to a series of papers in which the first applications in animal breeding of Bayesian hierarchical models computed via MCMC methods were reported. Subsequently, we were invited to teach or coteach courses in Likelihood and Bayesian MCMC analysis at Ames (USA), Armidale (Australia), Buenos Aires (Argentina), Edinburgh (Scotland), Guelph (Canada), Jokioinen (Finland), Liège (Belgium), Lleida (Spain), Madison (USA), Madrid (Spain), Milan (Italy), Montecillo (Mexico), Piracicaba (Brazil), Ribeirao Preto (Brazil), Toulouse (France), Uppsala (Sweden), Valencia (Spain), and Viçosa (Brazil). While in the course of these teaching experiences, we thought it would be useful to amalgamate some of our ideas in book form. What we hope you will read is the result of several iterations, starting from a monograph written by Daniel Sorensen and entitled "Gibbs Sampling in Quantitative Genetics". This was published first in 1996 as Internal Report No. 82 by the Danish Institute of Agricultural Sciences (DIAS).

Colleagues, friends, and loved ones have contributed in a variety of ways toward the making of this book. Carlos Becerril, Agustín Blasco, Rohan Fernando, Bernt Guldbrandtsen (who also made endless contributions with LaTeX related problems), Larry Schaeffer, and Bruce Walsh worked through a large part of the manuscript. Specific chapters were read by Anders Holst Andersen, José Miguel Bernardo, Yu-mei Chang, Miguel Pérez Enciso, Davorka Gulisija, Shyh-Forng Guo, Mark Henryon, Bjorg Heringstad, Just Jensen, Inge Riis Korsgaard, Mogens Sandø Lund, Nuala

Sheehan, Mikko Sillanpää, Miguel Angel Toro, and Rasmus Waagepetersen. We acknowledge their valuable suggestions and corrections. However, we are solely responsible for the mistakes that evaded scrutiny, as no book is entirely free of errors. Some of the mistakes find a place in the book by what one may mercifully call random accidents. Other mistakes may reflect incomplete knowledge of the topic on our side. We would be grateful if we could be made aware of these errors.

We wish to thank colleagues at the Department of Animal Breeding and Genetics, DIAS, and at the Departments of Animal Sciences and of Dairy Science of the University of Wisconsin-Madison for providing an intellectually stimulating and socially pleasant atmosphere. We are in special debt to Bernt Bech Andersen for much support and encouragement, and for providing a rare commodity: intellectual space.

We acknowledge John Kimmel from Springer-Verlag for encouragement and patience. Tony Orrantia, also from Springer-Verlag, is thanked for his sharp professional editing.

DG wishes to thank Arthur B. Chapman for his influential mentoring and for his views on the ethics of science, the late Charles R. Henderson for his pioneering work in linear models in animal breeding, and my colleagues and friends Jean-Louis Foulley, Rohan Fernando, and Sotan Im, from whom I learned much. DG had to fit the book into a rather hectic schedule of lecturing and research, both at home and overseas. This took much time away from Graciela, Magdalena, and Daniel Santiago, but they always gave me love, support and encouragement. I also wish to thank Gorgias and Alondra (my parents), the late Tatu, Morocha, and Héctor, and Chiquita, Mumu, Arturo, and Cecilia for their love.

This book was written "at work", at home, in airports and in hotels, on week-ends and on vacation. Irrespective of place, DS received consistent support from Maiken, Jon, and Elsebeth. They accepted that I was unavailable, and put up with moments of frustration (often in good spirit) when things did not work out. I was influenced by and am in debt to my early teachers in Reading and Edinburgh, especially Robert Curnow, Bill Hill, and Alan Robertson. Brian Kennedy introduced me to mixed linear model theory while I was a post-doc in Davis, California, and later in Guelph. I have learned much from him. To my parents I owe unveiling for me at an early age, that part of life that thrives on the top of trees, in worlds of reason and poetry, where it finds its space and achieves its splendor.

This page intentionally left blank

Contents

This page intentionally left blank

Part I

Review of Probability and Distribution Theory

This page intentionally left blank

1

Uncertainty, Random Variables, and Probability Distributions

1.1 Introduction

Suppose there is a data set consisting of observations on body weight taken on beef animals and that there are questions of scientific or practical importance to be answered from these data. The questions, for example, might be:

(1) Which are potential factors to be included in a statistical model for analysis of the records?

(2) What is the amount of additive genetic variance in the reference beef cattle population?

Answers to these questions are to be obtained from a finite amount of information contained in the data at hand. Because information is finite, there will be uncertainty associated with the outcome of the analyses. In this book, we are concerned with learning about the state of nature from experimental or observational data, and with measuring the degree of uncertainty at different stages of the learning process using probability models. This uncertainty is expressed via probabilities; arguably, probability can be viewed as the mathematics of uncertainty.

This chapter reviews probability models encountered often in quantitative genetic analysis, and that are used subsequently in various parts of this book. The treatment is descriptive and informal. It is assumed that the reader will have had an introductory course in probability theory or in mathematical statistics at the level of, for example, Hogg and Craig (1995), or Mood et al. (1974). The concept of a random variable is defined

and this is followed by an overview of univariate discrete and continuous distributions. Multivariate distributions are presented in the last section of the chapter, with considerable emphasis on the multivariate normal distribution, because of the important role it plays in statistical genetic analysis. For additional information, readers should consult treatises such as Johnson and Kotz (1969, 1970a,b, 1972).

1.2 Univariate Discrete Distributions

We start by providing a succinct and intuitive introduction to the concepts of random variable and probability which is adequate for our purposes.

Informally, a locus is a location in the genome where a gene resides. The word gene has become a rather elusive concept over recent years (see Keller, 2000, for a discussion). Here, we refer to the gene as the Mendelian "factor" transmitted from parent to offspring. Consider a locus at which there are two possible variants or alleles: A_1 and A_2. Envisage the random experiment consisting of generating genotypes from alleles A_1 and A_2. There are four possible outcomes for this experiment, represented by the four possible genotypes: A_1A_1, A_1A_2, A_2A_1, and A_2A_2. Any particular outcome cannot be known in advance; thus the use of the word "random". The set of all possible outcomes of the experiment constitutes its sample space, denoted by Ω. "Possible" means "conceptually possible" and often compromises between some perception of physical reality and mathematical convenience. In this simple case, the sample space consists of the four possible genotypes; that is, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Each possible outcome ω is a sample point in Ω. Here, $\omega_1 = A_1A_1$, $\omega_2 = A_1A_2$, $\omega_3 = A_2A_1$, and $\omega_4 = A_2A_2$.

Typically one is interested not so much on the experiment and on the totality of its outcomes, but on some consequences of the experiment. For example, imagine that individuals carrying allele A_2 suffer from a measurable disease. Let $X(\omega)$ be the function of the sample point ω, defined as follows: $X(\omega) = 0$ if $\omega = \omega_1$, $X(\omega) = 1$ if $\omega = \omega_2$, ω_3 or ω_4, where 0 and 1 indicate absence or presence of disease, respectively.

The same gene could have an additive effect on another trait, such that allele A_2 confers an increase in some measurement. A different function $Y(\omega)$ of ω could represent this situation defined as follows: $Y(\omega) = 0$ if $\omega = \omega_1$, $Y(\omega) = 1/2$ if $\omega = \omega_2$, ω_3, $Y(\omega) = 1$ if $\omega = \omega_4$. These two functions map the sample space to the real line $\mathbb{R}$. That is, the functions $X(\omega)$ and $Y(\omega)$, with domain Ω, make some real number correspond to each outcome of the experiment. The functions $X(\omega)$ and $Y(\omega)$ are known as random variables. Some authors prefer the term random quantities (i.e., De Finetti, 1975a; Bernardo and Smith, 1994), because the function in question is not a variable, but a deterministic mapping from Ω to $\mathbb{R}$. However, in this book, the usual term, random variable, will be adopted.

When attention focuses at the level of the random variable, this allows in some sense, to ignore the (often abstract) sample space Ω. One can define a new (concrete) sample space S in $\mathbb{R}$, to be the range of possible values that the random variable can take, so that no further mapping is required. In this case the random variable becomes the identity mapping, and can be thought of as the variable that describes the result of the random experiment. In the case of $X(\omega)$, the sample space is $S = \{0, 1\}$, whereas in the case of $Y(\omega)$ the sample space is $S = \left\{0, \frac{1}{2}, 1\right\}$. The random variable $X(\omega)$ will be written simply as X.

To distinguish between a random variable and its realized values, upper case letters are often used for the former whereas lower case letters are employed for the latter. This notation becomes awkward when bold face letters denote matrices (upper case) or vectors (lower case), so this convention is not followed consistently in this book. However, it will be clear from the context whether one is discussing a random variable or its realized values.

The sample spaces of the random variables X and Y consist of a countable number of values. Such spaces are called discrete and X and Y are discrete random variables.

The random variable X (whether discrete or continuous) in turn induces a probability P_X on S. This is a function that to each event (subset) $G \subseteq S$, assigns a number in the closed interval $[0, 1]$, defined by

$$P_X(G) = \Pr(\{\omega \in \Omega : X(\omega) \in G\}) = \Pr(X \in G). \tag{1.1}$$

The probability P_X is called the distribution of X, which for a discrete random variable, is completely determined by specifying $\Pr(X = x)$ for all x.

The mathematical model sketched above is quite divorced from the intuitive understanding(s) of the concept of probability. The interpretation of probability of the event G usually adopted in this book, is that it measures the subjective degree of plausibility or belief in G, conditional on information currently available to the subject. An alternative interpretation of probability is as the long run frequency of occurrence of G with independent replications of the random experiment carried out under identical conditions.

By definition, a probability is a number between 0 and 1; both 0 and 1 are included as valid numbers. A probability cannot be negative or larger than 1. Thus,

$$0 \leq \Pr(X = x) \leq 1, \text{ for all } x \in S.$$

Suppose that the sample space $S = \{x_1, x_2, \ldots\}$ is countable. The probability function (or probability mass function, abbreviated p.m.f.) of a discrete random variable X is defined by

$$p(x) = \begin{cases} \Pr(X = x), & \text{if } x \in S, \\ 0, & \text{otherwise.} \end{cases}$$

In distribution theory, the values that X can take are often denoted mass points, and $p(x)$ is the probability mass associated with the mass point x. The distribution (1.1) reduces to the p.m.f. when the subset G becomes a mass point at $X = x$. The support is the set of values of x for which the probability function is non-zero. The probability function gives a full description of the randomness or uncertainty associated with X. It can be used to calculate the probability that X takes certain values, or that it falls in a given region.

To illustrate, suppose that an individual is the progeny from crossing randomly two lines, and that each line consists of genotypes A_1A_2 only. Let the random variable X now represent the number of A_2 alleles in the progeny from this cross. From Mendel's laws of inheritance, the probability distribution of X is

Genotype	A_2A_2	A_2A_1	A_1A_1
x	2	1	0
$\Pr(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The probability that the number of A_2 alleles in a randomly drawn individual is 2, say, is

$$\Pr(\{\omega \in \Omega : X(\omega) = 2\}) = \frac{1}{4},$$

abbreviated hereinafter as

$$\Pr(X = 2) = p(2) = \frac{1}{4}. \tag{1.2}$$

In the above example, we have

$$\Pr(X = 0) = \frac{1}{4};\ \Pr(X = 1) = \frac{1}{2};\ \Pr(X = 2) = \frac{1}{4},$$

for each of the three values that X can take. Similarly,

$$\Pr(X = 1 \text{ or } X = 2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

To every random variable X there is an associated function called the cumulative distribution function (c.d.f.) of X. It is denoted by $F(x)$ and is defined by

$$F(x) = \Pr(X \leq x), \text{ for all } x. \tag{1.3}$$

The c.d.f. is also a special case of (1.1), when G is the half-open interval $(-\infty, x]$. The notation emphasizes that the c.d.f. is a function of x. For the example above, for $x = 1$,

$$F(1) = p(0) + p(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}. \tag{1.4}$$

Since $F(\cdot)$ is defined for all values of x, not just those in $S = \{0, 1, 2\}$, one can compute for $x = 1.5$ and for $x = 1.99$ say, $F(1.5) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$ and $F(1.99) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$, respectively. However, for $x = 2$, $F(2) = 1$. Thus, $F(x)$ is a step-function. In this example, the entire c.d.f. is

$$\begin{aligned} F(x) &= 0, \text{ for } x < 0, \\ F(x) &= \frac{1}{4}, \text{ for } 0 \leq x < 1, \\ F(x) &= \frac{3}{4}, \text{ for } 1 \leq x < 2, \\ F(x) &= 1, \text{ for } x \geq 2. \end{aligned}$$

In general, one writes

$$F(x) = \sum_{t \leq x} p(t), \tag{1.5}$$

where t is a "dummy" variable as used in calculus. Another notation that will be used in this book is

$$F(x) = \sum_{t} I(t \leq x)\, p(t), \tag{1.6}$$

where $I(t \leq x)$ is the indicator function that takes the value 1 if the argument is satisfied, in this case, $t \leq x$, and zero otherwise.

The distribution function $F(x)$ has the following properties:

- If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- $\Pr(x_1 < X \leq x_2) = \Pr(X \leq x_2) - \Pr(X \leq x_1) = F(x_2) - F(x_1)$.
- $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

Example 1.1 *Point mass*

Consider a point mass at i (such that the total probability mass is concentrated on i). Then,

$$F(x) = \begin{cases} 0, & x < i, \\ 1, & x \geq i. \end{cases}$$

This property that F is only 0 or 1 is a characteristic of point masses. ■

In the following section, some of the most widely used discrete distributions in genetics are discussed. These are the Bernoulli, binomial and the Poisson distributions.

1.2.1 The Bernoulli and Binomial Distributions

A random variable X that has a Bernoulli probability distribution can take either 1 or 0 as possible values (more generally, it can have two modalities

only) with probabilities θ and $1-\theta$, respectively. The quantity θ is called a parameter of the distribution of X and, in practice, this may be known or unknown. The p.m.f. of X is

$$\Pr(X=x|\theta)=Br(x|\theta)\begin{cases}\theta^{x}(1-\theta)^{1-x}, & \text{for } x=0,1,\\ 0, & \text{otherwise,}\end{cases} \tag{1.7}$$

where $0\leq\theta\leq 1$ and Br is an abbreviation for Bernoulli.

In general, the notation $Br(\theta)$, say, will be used to specify a distribution, in this case the Bernoulli distribution with parameter θ, whereas $Br(x|\theta)$ refers to the corresponding p.m.f. of X. The notation for the p.m.f. $\Pr(X=x|\theta)$ is used to stress the dependence on the parameter θ.

Often a random variable X having a Bernoulli p.m.f. is referred to as saying that X has a Bernoulli distribution. This phraseology will be adopted in this book for other random variables (discrete or continuous) having a named p.m.f. or probability density function (abbreviated p.d.f. and defined in Section 1.3).

Suppose the frequency of allele A in a population is 0.34 (the gene frequency is given by the number of A alleles found, divided by twice the number of individuals in a diploid organism). Then one may postulate that a randomly drawn allele is a Bernoulli random variable with parameter $\theta=0.34$. If θ is known, one can specify the probability distribution of X exactly. In this case, there will be uncertainty about the values that X can take, but not about the value of θ. If θ is unknown, there are two (intimately related) consequences. First, the Bernoulli random variables in the sequence $X_1, X_2, \ldots$ are no longer independent (the concept of independence is discussed below and in Section 1.4). Second, there will be an extra source of uncertainty associated with X, and this will require introducing an additional probability distribution for θ in order to calculate the probability distribution of X correctly (accounting for the uncertainty about θ). This is a central problem in statistical inference, as will be seen later in this book. Typically, (1.7) cannot be specified free of error about the parameter, because θ must be inferred from previous considerations, or from a finite set of data at hand. Clearly, it is somewhat disturbing to use an estimated value of θ and then proceed as if (1.7) were the "true" distribution. In this book, procedures for taking such "errors" into account will be described.

The mean of the Bernoulli distribution is obtained as follows:

$$E(X|\theta)=0\times\Pr(X=0|\theta)+1\times\Pr(X=1|\theta)=\theta, \tag{1.8}$$

where $E(\cdot)$ denotes expected or average value. Similarly,

$$E\left(X^2|\theta\right)=0^2\times\Pr(X=0|\theta)+1^2\times\Pr(X=1|\theta)=\theta.$$

Then, by definition of the variance of a random variable, abbreviated as Var hereinafter,

$$Var(X|\theta)=E\left(X^2|\theta\right)-E^2(X|\theta)=\theta-\theta^2=\theta(1-\theta). \tag{1.9}$$

A random variable X has a binomial distribution (the term process will be used interchangeably) if its p.m.f. has the form

$$\Pr(X = x|\theta, n) = Bi(x|\theta, n) = \begin{cases} \binom{n}{x} \theta^x (1-\theta)^{n-x}, & x = 0, 1, ..., n, \\ 0, & \text{otherwise}, \end{cases} \tag{1.10}$$

where $0 \leq \theta \leq 1$ and $n \geq 1$. Here the parameters of the process are θ and n, and inferences about X are conditional on these parameters. In a sequence of n identical and independent Bernoulli trials, each with "success" probability θ, define the random variables $Y_1, Y_2, \ldots, Y_n$ as follows:

$$Y_i = \begin{cases} 1, & \text{with probability } \theta, \\ 0, & \text{with probability } 1-\theta. \end{cases}$$

Then the random variable $X = \sum_{i=1}^{n} Y_i$ has the binomial distribution $Bi(\theta, n)$. The mean and variance of X are readily seen to be equal to $n\theta$ and to $n\theta(1-\theta)$, respectively.

The binomial distribution thus stems from consideration of n mutually independent Bernoulli random variables all having the same parameter θ. Mutual independence means that knowledge of the value of some subset of the n Bernoulli random variables does not alter the state of uncertainty about the remaining ones. For example, in the two-variable case, if two coins are tossed by two different individuals, having observed the outcome of one of the tosses will not modify the state of uncertainty about the second toss.

Example 1.2 *Number of females in a sibship of size 3*
Let the random variable X denote the total number of females observed in three randomly chosen births, so $n = 3$ and the value of X ranges from 0 to 3. Assume that θ, the probability of a female birth, is known and equal to $\theta = 0.49$. Then

$$\begin{aligned} \Pr(X = 0|\theta, n) &= Bi(0|0.49, 3) = 0.1327, \\ \Pr(X = 1|\theta, n) &= Bi(1|0.49, 3) = 0.3823, \\ \Pr(X = 2|\theta, n) &= Bi(2|0.49, 3) = 0.3674, \\ \Pr(X = 3|\theta, n) &= Bi(3|0.49, 3) = 0.1176. \end{aligned}$$

The events are mutually exclusive and exhaustive because, for example, if $X = 0$, then all other values of the random variable are excluded. Hence, the probability that X takes at least one of these four values is equal to 1. Further, the probability that X is equal to 0 or 2 is

$$\begin{aligned} \Pr(X = 0 \text{ or } X = 2|\theta, n) &= \Pr(X = 0|\theta, n) + \Pr(X = 2|\theta, n) \\ &= 0.1327 + 0.3674 = 0.5001. \end{aligned}$$

When events are disjoint (mutually exclusive), the probability that either one or another event of interest occurs is given by the sum of their elementary probabilities, as in the above case. ■

Example 1.3 *Simulating a binomial random variable*
A simple way of simulating a binomial random variable $X \sim Bi(\theta, n)$ is to generate n independent standard uniforms and to set X equal to the number of uniform variates that are less than or equal to θ (Gelman et al., 1995). ■

1.2.2 The Poisson Distribution

When θ is small and n is large, an approximation to $Bi(x|\theta, n)$ is obtained as follows. Let $n\theta = \lambda$ be a new parameter. Expanding the binomial coefficient in (1.10) one obtains, for $X = k$,

$$\begin{aligned} Bi(k|\theta, n) &= \frac{n(n-1)...(n-k+1)}{k!}\theta^k(1-\theta)^{n-k} \\ &= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n} \\ &\times \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned} \tag{1.11}$$

Suppose that $n \to \infty$ and $\theta \to 0$, such that $n\theta = \lambda$ remains constant. Taking limits, when $n \to \infty$,

$$\begin{aligned} \lim_{n\to\infty} Bi(k|\theta, n) &= \lim_{n\to\infty}\left[\frac{n-1}{n}\cdots\frac{n-k+1}{n}\left(1-\frac{\lambda}{n}\right)^{-k}\right] \\ &\times \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n \\ &= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n. \end{aligned} \tag{1.12}$$

Recall the binomial expansion

$$(a+b)^n = \sum_{x=0}^{n}\binom{n}{x}b^x a^{n-x}.$$

Put $a = 1$ and $b = -\lambda/n$. Then

$$\begin{aligned} \left(1-\frac{\lambda}{n}\right)^n &= \sum_{x=0}^{n}\binom{n}{x}\left(\frac{-\lambda}{n}\right)^x \\ &= \sum_{x=0}^{n}\frac{n}{n}\frac{(n-1)}{n}\cdots\frac{(n-x+1)}{n}\frac{(-\lambda)^x}{x!}, \end{aligned}$$

so

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=\sum_{x=0}^{\infty}\frac{(-\lambda)^x}{x!}. \tag{1.13}$$

A Taylor series expansion of $\exp(V)$ about 0 yields

$$\exp(V)=1+V+\frac{V^2}{2!}+\frac{V^3}{3!}+\cdots$$

so if the series has an infinite number of terms, this is expressible as

$$\exp(V)=\sum_{x=0}^{\infty}\frac{V^x}{x!}.$$

Making use of this in (1.13), with $V=-\lambda$, leads directly to

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=\exp(-\lambda).$$

Finally, employing this in (1.12) gives the result

$$\lim_{n\to\infty}Bi(k|\theta,n)=\frac{\lambda^k\exp(-\lambda)}{k!}. \tag{1.14}$$

This gives an approximation to the binomial probabilities for situations where n is "very large" and the probability θ is "very small". This may be useful, for example, for calculating the probability of finding a certain number of mutants in a population when the mutation rate is low.

In fact, approximation (1.14) plays an important role in distribution theory in statistics: a random variable X is said to have a Poisson distribution with parameter λ (this being strictly positive, by definition) if its probability function is

$$\Pr(X=x|\lambda)=Po(x|\lambda)=\begin{cases}\lambda^x\exp(-\lambda)/x!, & \text{for } x=0,1,2,\ldots,\infty,\\ 0, & \text{otherwise,}\end{cases} \tag{1.15}$$

where Po is an abbreviation for Poisson. Here, X is a discrete random variable, and its sample space is countably infinite. This distribution could feature in models for quantitative trait loci (QTL) detection, and may be suitable, for example, for describing the occurrence of family sizes in animals with potentially large sibships, such as insects or fish. It has been used in animal breeding in the analysis of litter size in pigs and sheep.

Example 1.4 *Exposure of mice to carcinogens*

In order to illustrate the relationship between the binomial and the Poisson distributions, suppose there is a 2% chance that a mouse exposed to a certain carcinogen will develop cancer. What is the probability that two

or more mice will develop cancer within a group of 60 experimental mice exposed to the causative agent? Let the random variable X describe the number of mice that develop cancer, and suppose that the sampling model $Bi(x|0.02, 60)$ is tenable. Using (1.10), the probability that two or more mice will have a carcinoma is computed to be

$$\begin{aligned}\Pr(X \geq 2|\theta, n) &= 1 - \Pr(X \leq 1|\theta, n) \\ &= 1 - \Pr(X = 0|\theta, n) - \Pr(X = 1|\theta, n) \\ &= 0.3381.\end{aligned}$$

However, given that $\theta = 0.02$ and $n = 60$ seem (in a rather arbitrary manner) reasonably small and large, respectively, the above probability is approximated using the Poisson distribution $Po(1.2)$, with its parameter arrived at as $\lambda = 0.02 \times 60 = 1.2$. The probability that k mice will develop cancer is given by

$$\Pr(X = k|\lambda = 1.2) = \frac{e^{-1.2}(1.2)^k}{k!}.$$

Therefore,

$$\begin{aligned}\Pr(X \geq 2|\lambda = 1.2) &= 1 - \Pr(X \leq 1|\lambda = 1.2) \\ &= 1 - \Pr(X = 0|\lambda = 1.2) - \Pr(X = 1|\lambda = 1.2) \\ &= 1 - \frac{1 + 1.2}{e^{1.2}} = 0.3374,\end{aligned}$$

so the approximation is satisfactory, at least in this case. ■

1.2.3 Binomial Distribution: Normal Approximation

Another useful approximation to the binomial distribution is based on the central limit theorem, to be discussed later in this chapter. Let X be a binomial variable, defined as $X = \sum_{i=1}^{n} U_i$, where the U_i's are independent and identically distributed (for short, i.i.d.) $Br(\theta)$ random variables. Then, as given in (1.8) and in (1.9), $E(U_i) = \theta$ and $Var(U_i) = \theta(1 - \theta)$. Therefore, the algebra of expectations yields $E(X) = \sum_{i=1}^{n} E(U_i) = n\theta$ and $Var(X) = \sum_{i=1}^{n} Var(U_i) = n\theta(1 - \theta)$. The formula for the variance is a consequence of the independence assumption made about the Bernoulli variates, so the variance of a sum is the sum of the variances. Now, the random variable

$$Z = \frac{X - n\theta}{[n\theta(1 - \theta)]^{1/2}} \tag{1.16}$$

has expectation and variance equal to 0 and 1, respectively. By virtue of the central limit theorem, for large n, Z is approximately distributed as an $N(0, 1)$ random variable, which stands for a standard, normally distributed

random variable (it is assumed that the reader is somewhat familiar with the normal or Gaussian distribution, but it will be discussed subsequently).

If $X \sim Bi(\theta, n)$, given fixed numbers x_1 and x_2, $x_1 < x_2$, as $n \to \infty$, it follows, from the approximation above, that

$$\begin{aligned}\Pr(x_1 < X \leq x_2) &= \Pr\left(\frac{x_1 - n\theta}{[n\theta(1-\theta)]^{1/2}} < \frac{X - n\theta}{[n\theta(1-\theta)]^{1/2}} \leq \frac{x_2 - n\theta}{[n\theta(1-\theta)]^{1/2}}\right) \\ &= \Pr(z_1 < Z \leq z_2) = \Pr(Z \leq z_2) - \Pr(Z \leq z_1) \\ &\approx \Phi(z_2) - \Phi(z_1), \end{aligned} \tag{1.17}$$

where $z_i = (x_i - n\theta)/[n\theta(1-\theta)]^{1/2}$, and $\Phi(\cdot)$ is the c.d.f. of a standard normal random variable.

For the mouse example, the probability that two or more mice will develop cancer can be approximated as

$$\begin{aligned}\Pr(X \geq 2) &= \Pr\left(\frac{X - n\theta}{[n\theta(1-\theta)]^{1/2}} \geq \frac{2 - n\theta}{[n\theta(1-\theta)]^{1/2}}\right) \\ &= \Pr(Z \geq 0.7377) \\ &= 1 - \Pr(Z < 0.7377) \approx 0.230.\end{aligned}$$

Here the normal approximation is not adequate. In general, this approximation does not work well, unless $\min[n\theta, n(1-\theta)] \geq 5$, and the behavior is more erratic when θ is near the boundaries. In the example, $n\theta$ is equal to $60(0.02) = 1.2$. A better approximation is obtained using the continuity correction (e.g., Casella and Berger, 1990). Instead of approximating $\Pr(X \geq 2)$, one approximates $\Pr(X \geq 2 - 0.5)$; this yields 0.391, which is a little closer to the exact result 0.338 obtained before.

For a recent discussion about the complexities associated with this approximation in the context of calculation of the confidence interval for a proportion, the reader is referred to Brown et al. (2001) and to Henderson and Meyer (2001).

1.3 Univariate Continuous Distributions

The variable Z, defined in (1.16) is an example of a continuous random variable, that is, one taking any of an infinite number of values. In the case of Z these values exist in the real line, $\mathbb{R}$; the sample space is here continuous. More precisely, a random variable is (absolutely) continuous if its c.d.f. $F(x)$, defined in (1.18), is continuous at every real number x.

The probability that the continuous random variable will assume any particular value is zero because the possible values that X can take are not countable. A typical example of a random variable that can be regarded

as continuous is the body weight of an animal. In practice, however, measurements are taken on a discrete scale only, with the degree of discreteness depending on the resolution of the instrument employed for recording. However, much simplicity can be gained by treating body weight as if it were continuous.

In the discrete case, as given in (1.5), the distribution function is defined in terms of a sum. On the other hand, the c.d.f. of a continuous random variable X on the real line is given by the integral

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p(t)\, dt, \quad \text{for} -\infty < x < \infty, \tag{1.18}$$

where $p(\cdot)$ is the p.d.f. of this random variable. If dt is an infinitesimally small increment in the variable, so that the density is continuous and fairly constant between t and $t + dt$, the probability that the random variable takes a value in the interval between t and $t + dt$ is nearly $p(t)\, dt$ (Bulmer, 1979). If $p(t) > p(u)$ then values of X are more likely to be seen close to t than values of X close to u.

As a little technical aside, the p.d.f. is not uniquely defined. This is because one can always change the value of a function at a finite number of points, without changing the integral of the function over the interval. This lack of uniqueness does not pose problems for probability calculations but can lead to subtle complications in certain statistical applications. However, if the distribution function is differentiable then a natural choice for $p(x)$ is

$$\frac{d}{dx} F(x) = p(x). \tag{1.19}$$

Any probability statement about the random variable X can be deduced from either $F(x)$ or $p(x)$; the p.d.f. encapsulates all the necessary information needed to make statements of uncertainty about X. This does not mean that X is characterized by its c.d.f.. Two different random variables can have the same c.d.f. or p.d.f.. As a trivial example, consider the Gaussian random variable X with mean zero and variance 1. Then the random variable $-X$ is also Gaussian with the same mean and variance.

Properties of the p.d.f. are:

- $p(x) \geqslant 0$ for all x.
- $\int_{-\infty}^{\infty} p(x)\, dx = 1$.

(Any positive function p satisfying $\int_{-\infty}^{\infty} p(x)\, dx = 1$ may be termed a p.d.f.).

- for $x \in (a, b)$, $\Pr(a < X < b) = \int_a^b p(x)\, dx = F(b) - F(a)$.

(Because of the lack of uniqueness mentioned above, it is possible to find another density function p^*, such that $\int_a^b p(x)\,dx = \int_a^b p^*(x)\,dx$, for all a and b but for which $p(x)$ and $p^*(x)$ differ at a finite number of points).

By virtue of the continuity of X,

$$\begin{aligned}&\Pr(a < X < b) = \Pr(a < X \leq b)\\ &= \Pr(a \leq X < b) = \Pr(a \leq X \leq b),\end{aligned} \tag{1.20}$$

so $<$ and $\leq$ can be used indiscriminately in this case.

- The probability that X takes a particular value is,

$$\Pr(X = x) = \lim_{\varepsilon \to 0} \int_{a-\varepsilon}^{a+\varepsilon} p(x)\,dx = 0.$$

- As noted above, the p.d.f. describes the probability that X takes a value in a neighborhood of a point x. Thus if Δ is small and p is continuous,

$$\Pr(x \leq X \leq x + \Delta) = \int_x^{x+\Delta} p(x)\,dx \cong \Delta\, p(x). \tag{1.21}$$

Consider the continuous random variable X that has density $p(x)$ with support on the real line, $\mathbb{R}$. Let A be some interval in $\mathbb{R}$. Notationally, we say that $A \subseteq \mathbb{R}$ and, using the concept of the indicator function introduced in (1.6), the probability that the random variable X belongs in A can be written

$$\begin{aligned}\Pr(X \in A) &= \int_A p(x)\,dx\\ &= \int I(x \in A)\, p(x)\,dx\\ &= E[I(x \in A)].\end{aligned} \tag{1.22}$$

In this book we shall often encounter integrals of the form

$$\int g(x)\, p(x)\,dx.$$

This is the expectation of g with respect to a probability on $\mathbb{R}$. It will be referred to as "integrating g over the distribution of X". We may also use, loosely, "integrating g, or taking the expectation of g, with respect to p".

Often it suffices to identify the p.d.f. of a random variable X only up to proportionality; this is particularly convenient in the context of a Bayesian analysis, as will be seen repeatedly in this book. A function $f(x)$, such that $p(x) \propto f(x)$, is called the kernel of $p(x)$. That is, if it is known that $p(x) = Cf(x)$, the constant C can be determined using the fact that a p.d.f. integrates to 1. Thus, for a random variable X taking any value on the real line

$$\int_{-\infty}^{\infty} p(x)\,dx = C\int_{-\infty}^{\infty} f(x)\,dx = 1$$

so

$$C = \frac{1}{\int_{-\infty}^{\infty} f(x)\,dx}.$$

Equivalently,

$$p(x) = \frac{f(x)}{\int_{-\infty}^{\infty} f(x)\,dx}. \tag{1.23}$$

If the random variable takes values within the interval (*lower*, *upper*), say, the integration limits above must be changed accordingly.

Example 1.5 *Kernel and integration constant of a normal distribution*
Suppose X has a normal distribution with mean μ and variance σ^2. Then its density function is

$$p\left(x|\mu,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

The kernel is

$$f(x) = \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

and the constant of integration is

$$C = \frac{1}{\int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx} = \left(2\pi\sigma^2\right)^{-\frac{1}{2}}. \tag{1.24}$$

To prove (1.24), denote the integral in the denominator as I. Using the substitution $u = (x-\mu)/\sigma$, since $dx = \sigma\,du$, I can be written

$$I = \sigma\int_{-\infty}^{\infty} \exp\left[-\frac{u^2}{2}\right] du.$$

Now write the square of I as a double integral

$$\begin{aligned} I^2 &= \sigma^2\int_{-\infty}^{\infty} \exp\left[-\frac{u^2}{2}\right] du \int_{-\infty}^{\infty} \exp\left[-\frac{v^2}{2}\right] dv \\ &= \sigma^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{\left(u^2+v^2\right)}{2}\right] du\,dv \end{aligned}$$

and make a transformation to polar coordinates

$$\begin{aligned} u &= r\cos(\theta), \\ v &= r\sin(\theta). \end{aligned}$$

If the angle θ is expressed in radians, then it can take values between 0 and 2π. The variable r takes values between 0 and ∞. From calculus, the absolute value of the two-dimensional Jacobian of the transformation is the absolute value of the determinant

$$\begin{aligned} \det\left[\frac{\partial(u,v)}{\partial(r,\theta)}\right] &= \det\begin{bmatrix} \partial u/\partial r & \partial u/\partial\theta \\ \partial v/\partial r & \partial v/\partial\theta \end{bmatrix} \\ &= \det\begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix} \\ &= r\left[\cos^2(\theta)+\sin^2(\theta)\right] = r. \end{aligned}$$

Using this, together with the result

$$u^2+v^2 = r^2\left[\cos^2(\theta)+\sin^2(\theta)\right] = r^2$$

one obtains

$$\begin{aligned} I^2 &= \sigma^2\int_0^\infty\int_0^{2\pi} r\exp\left(-\frac{r^2}{2}\right)d\theta dr \\ &= \sigma^2\int_0^\infty r\exp\left(-\frac{r^2}{2}\right)dr\int_0^{2\pi}d\theta \\ &= 2\pi\sigma^2\int_0^\infty r\exp\left(-\frac{r^2}{2}\right)dr. \end{aligned}$$

The integral in the preceding expression can be evaluated from the family of gamma integrals (Box and Tiao, 1973). For $a>0$ and $p>0$:

$$\int_0^\infty r^{p-1}\exp\left(-ar^2\right)dr = \frac{1}{2}a^{-p/2}\Gamma\left(\frac{p}{2}\right), \tag{1.25}$$

where $\Gamma(\cdot)$ is the gamma function (which will be discussed later). Setting $p=2$ and $a=\frac{1}{2}$, it follows that

$$\int_0^\infty r\exp\left(-\frac{r^2}{2}\right)dr = \frac{1}{2}\times 2\times\Gamma(1) = 1,$$

because $\Gamma(1)=0!=1$. Therefore, $I^2=2\pi\sigma^2$, and, finally,

$$C = \frac{1}{\int_0^\infty \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx} = \frac{1}{I} = \frac{1}{\sqrt{2\pi\sigma^2}}$$

yielding the desired result. ■

1.3.1 The Uniform, Beta, Gamma, Normal, and Student-t Distributions

Uniform Distribution

A random variable X has a uniform distribution over the interval $[a, b]$ (in view of (1.20) we shall allow ourselves to be not quite consistent throughout this book when defining the support of distributions of continuous random variables whose sample space is truncated) if its p.d.f. is

$$p(x|a,b) = Un(x|a,b) = \begin{cases} \frac{1}{b-a}, & \text{for } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases} \tag{1.26}$$

The c.d.f. is

$$F(x|a,b) = \begin{cases} 0, & \text{if } x < a, \\ \int_a^x Un(x|a,b)\,dx = \frac{1}{b-a}\int_a^x dx = \frac{x-a}{b-a}, & \text{for } a \le x \le b, \\ 1, & \text{for } x > b. \end{cases}$$

The average value of X is found by multiplying the value x times its "infinitesimal probability" $p(x)\,dx$, and then summing over all possible such values. This corresponds to the integral

$$E(X|a,b) = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}. \tag{1.27}$$

Likewise the average value of X^2 is

$$E\left(X^2|a,b\right) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3-a^3}{3(b-a)} = \frac{(a+b)^2 - ab}{3}.$$

Then the variance can be found to be, after algebra,

$$\begin{aligned} Var(X|a,b) &= E\left(X^2|a,b\right) - E^2(X|a,b) \\ &= \frac{(a+b)^2 - ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \end{aligned} \tag{1.28}$$

Example 1.6 *Simulation of discrete random variables*
Suppose one wishes to generate a discrete random variable with p.m.f.

$$\Pr(X = x_i) = p_i, \quad i = 1, 2, 3, 4,$$

with $\sum_{i=1}^4 p_i = 1$. To accomplish this, generate a random number U from a uniform distribution $Un(0,1)$ and set

$$X = \begin{cases} x_1, & if\ U < p_1, \\ x_2, & if\ p_1 \le U < p_1 + p_2, \\ x_3, & if\ p_1 + p_2 \le U < p_1 + p_2 + p_3, \\ x_4, & if\ p_1 + p_2 + p_3 \le U < 1. \end{cases}$$

To confirm that X has the desired distribution, first note that $p(u) = 1$. Then

$$\begin{aligned}\Pr(X = x_3) &= \Pr(p_1 + p_2 \leq U < p_1 + p_2 + p_3) \\ &= \Pr(U \leq p_1 + p_2 + p_3) - \Pr(U \leq p_1 + p_2) \\ &= \int_0^{p_1+p_2+p_3} du - \int_0^{p_1+p_2} du = p_3.\end{aligned}$$

This method of generating the random variable is known as the inverse transform method. For continuous random variables, the general algorithm is as follows. If one wishes to generate a random variable that has c.d.f. F, then:
(a) first generate $U \sim Un(0,1)$; and
(b) then set $X = F_X^{-1}(U)$;
and X has the desired c.d.f. $F_X(\cdot)$. To see this, note that

$$\begin{aligned}\Pr(X \leq x) &= \Pr\left(F_X^{-1}(U) \leq x\right) \\ &= \Pr(U \leq F_X(x)) \\ &= \int_0^{F_X(x)} du = F_X(x).\end{aligned}$$

The equality in the second line follows because $F_X(x)$ is a monotone function. ■

Simulating Uniform Random Numbers

The uniform distribution plays an important role in the generation of pseudorandom numbers with a computer. Ripley (1987) and Ross (1997) discuss how uniform random "deviates" can be produced in an electronic computer; Fishman (1973) presents a readable discussion. A summary of the basic ideas follows.

The most common algorithm for generating "random" numbers actually produces a nonrandom sequence of values. Each number depends on the previous one, which implies that all numbers are determined by the initial value in the sequence. However, if a generator is designed appropriately (in some sense which is beyond the scope of this book), the numbers appear as "independent" draws from a uniform distribution. A popular method is known as the "linear congruential generator". It has the form

$$Z_i \equiv aZ_{i-1} + c \mod(m), \tag{1.29}$$

where the "seed" Z_0 is an integer $0 \leq Z_0 < m$, the multiplier a is an integer $1 < a < m$, c is an integer increment $(0 \leq c < m)$, and modulus m is an integer $0 < m$. The modulus notation means that Z_i is the remainder obtained when $aZ_{i-1} + c$ is divided by m. This remainder can be expressed

as

$$Z_i = aZ_{i-1} + c - \left\lfloor \frac{aZ_{i-1}+c}{m} \right\rfloor m,$$

where

$$\left\lfloor \frac{aZ_{i-1}+c}{m} \right\rfloor = \text{largest integer in the quantity,}$$

note that $aZ_{i-1}+c$ is integer. For example, if the quantity within brackets is $\frac{12}{5}$, the largest integer is 2; if the quantity is $\frac{1}{6}$, the largest integer is 0. The arithmetic guarantees that each element in the sequence $\{Z_i\}$ is an integer in the interval $[0, m-1]$. Then, when $\{Z_i\}$ displays "sufficiently random" behavior, a pseudorandom sequence in the unit interval can be obtained as

$$X_i = \frac{Z_i}{m}.$$

Since $Z_i < m$, the sequence can contain at most m distinct numbers and, as soon as a number is repeated, the entire sequence repeats itself. Thus, m should be as large as possible to ensure a large quantity of distinct numbers. The parameters a, c and Z_0 must be selected so that as many as possible of the m numbers occur in a cycle. It can be shown (Cohen, 1986) that the sequence $\{Z_i\}$ is periodic: there exists a $P \leq m$ such that $Z_i = Z_{i+P}$ for all $i \geq 0$. A large P (therefore, a large m) is desirable for achieving apparent randomness. A generator is said to have full period when $P = m$; however, a full period does not guarantee that a generator is adequate. For example, old generators, such as RANDU (formerly used in the IBM Scientific Subroutine Package; Cohen, 1986) had $a = 65,539$, $c = 0$ and $m = 2^{31}$; it had a maximal period of 2^{29}, and yet failed some tests. On the other hand, the full-period LCG generator (Cohen, 1986) has $a = 69,069$, $c = 1$ and $m = P = 2^{32}$ and has failed tests as well. Marsaglia and Zaman (1993) describe "Kiss", a generator that has displayed excellent performance from a statistical point of view (Robert, 1994).

The dependency between Z_i and Z_0 can be deduced as follows. Consider (1.29) and write

$$\begin{aligned} Z_1 &\equiv aZ_0 + c \bmod (m), \\ Z_2 &\equiv aZ_1 + c \bmod (m), \\ Z_3 &\equiv aZ_2 + c \bmod (m). \end{aligned}$$

Expressing Z_3 in terms of Z_2, and this in terms of Z_1, one obtains the identity

$$Z_3 \equiv a^3 Z_0 + (a^2 + a + 1)c \bmod (m)$$

and, more generally

$$Z_n \equiv a^n Z_0 + (a^{n-1} + a^{n-2} + ... + a + 1)c \bmod (m).$$

Note that

$$(a^{n-1} + a^{n-2} + ... + a + 1)(a - 1) = a^n - 1.$$

Hence

$$Z_n \equiv a^n Z_0 + \frac{a^n - 1}{(a-1)} c \mod (m) \tag{1.30}$$

and, consequently,

$$X_n \equiv a^n X_0 + \frac{a^n - 1}{(a-1)m} c \mod (m)$$

where $X_0 = Z_0/m$.

Example 1.7 *A hypothetical random number generator*
In order to illustrate, an example from Fishman (1973) will be modified slightly. Consider a hypothetical generator with $a = 3$, $c = 0, Z_0 = 4$, and $m = 7$. The pseudo-random uniform deviates, given in the table below, can be computed with (1.29) or with (1.30):

i	$Z_i = aZ_{i-1} + c \left\lfloor \frac{aZ_{i-1}+c}{m} \right\rfloor m$	$X_i = \frac{Z_i}{m}$
0	4	$4/7 = 0.5714$
1	$3 \times 4 - \left\lfloor \frac{3\times 4}{7} \right\rfloor \times 7 = 12 - 1 \times 7 = 5$	$5/7 = 0.7143$
2	$3 \times 5 - \left\lfloor \frac{3\times 5}{7} \right\rfloor \times 7 = 15 - 2 \times 7 = 1$	$1/7 = 0.1429$
3	$3 \times 5 - \left\lfloor \frac{3\times 5}{7} \right\rfloor \times 7 = 15 - 2 \times 7 = 1$	$3/7 = 0.4286$
4	$3 \times 3 - \left\lfloor \frac{3\times 3}{7} \right\rfloor \times 7 = 9 - 1 \times 7 = 2$	$2/7 = 0.2857$
5	$3 \times 2 - \left\lfloor \frac{3\times 2}{7} \right\rfloor \times 7 = 6 - 0 \times 7 = 6$	$6/7 = 0.8571$
6	$3 \times 6 - \left\lfloor \frac{3\times 6}{7} \right\rfloor \times 7 = 18 - 2 \times 7 = 4$	$4/7 = 0.5714$

The sequence repeats itself after six steps. If the parameters stay as before and $m = 8$, a string of $4s$ is obtained, illustrating a very poor generator of pseudo-random numbers. ■

Beta Distribution

The beta distribution can be used as a model for random variables that take values between zero and one, such as probabilities, a gene frequency, or the coefficient of heritability. A random variable X has a beta distribution if its p.d.f. has the form

$$p(x|a,b) = Be(x|a,b) = \begin{cases} Cx^{a-1}(1-x)^{b-1}, & \text{for } x \in [0,1], \\ 0, & \text{otherwise,} \end{cases} \tag{1.31}$$

where $a > 0$ and $b > 0$ are parameters of this distribution. The value of the integration constant is $C = \Gamma(a+b)/\Gamma(a)\Gamma(b)$, where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \quad \alpha > 0, \tag{1.32}$$

and $\Gamma(\cdot)$ is the gamma function, as seen in (1.25). Note that, if $\alpha = 1$,

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}\Big|_0^\infty = 1.$$

Suppose $\alpha > 1$. The integral in (1.32) can be evaluated by parts. Recall the basic calculus result

$$\int u\, dv = uv - \int v\, du.$$

Put $u = x^{\alpha-1}$ and $dv = e^{-x}$; then $du = (\alpha-1)\, x^{\alpha-2}$ and $v = -e^{-x}$. The integral sought is then

$$\begin{aligned}\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx &= -x^{\alpha-1} e^{-x}\Big|_0^\infty + \int_0^\infty (\alpha-1)\, x^{\alpha-2} e^{-x}\, dx \\ &= (\alpha-1) \int_0^\infty x^{\alpha-2} e^{-x}\, dx = (\alpha-1)\,\Gamma(\alpha-1)\end{aligned} \tag{1.33}$$

with the preceding following from the definition of the gamma function in (1.32). Thus, provided α is integer and greater than one,

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\ldots 3 \times 2 \times 1 \times \Gamma(1) = (\alpha-1)! \tag{1.34}$$

Consider now the beta density in (1.31), and write it explicitly as

$$p(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}. \tag{1.35}$$

The expected value and variance of the beta distribution can be deduced by using the fact that the constant of integration is such that

$$C^{-1} = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{1.36}$$

The mean value of a beta distributed random variable is then

$$\begin{aligned} E(X|a,b) &= C\int_0^1 x\left[x^{a-1}(1-x)^{b-1}\right]dx \\ &= C\int_0^1 x^{a+1-1}(1-x)^{b-1}\,dx \\ &= C\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)\,a\Gamma(a)}{\Gamma(a)(a+b)\Gamma(a+b)} = \frac{a}{(a+b)}. \end{aligned} \tag{1.37}$$

Using a similar development, it can be established that

$$\begin{aligned} Var(X|a,b) &= E\left(X^2|a,b\right) - E^2(X|a,b) \\ &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned} \tag{1.38}$$

Example 1.8 *Variation in gene frequencies*

Consider the following, patterned after Wright (1968). Suppose n alleles are sampled from some population, and that the outcome of each draw is either A or a. In the absence of correlation between outcomes, the probability distribution of X, the number of A alleles observed in the sample of size n, can be calculated using the binomial distribution (1.10); the probability θ corresponds to the frequency of allele A in the population. Wright refers to this set of n draws as "clusters". If this sampling scheme were to be repeated a large number of times, an excess of clusters consisting largely or entirely of A or of a would indicate correlation between outcomes resulting, for example, from family aggregation. The distribution of correlated draws can be derived by assuming that θ varies according to some distribution. Because θ varies between 0 and 1, the beta distribution is convenient here. The parameterization of Wright is adopted subsequently. Let

$$c = a + b$$

and

$$\overline{\theta} = \frac{a}{a+b}$$

be the mean of the distribution. The beta density (1.35) can then be written as

$$p\left(\theta|\overline{\theta},c\right) = \frac{\Gamma(c)}{\Gamma\left(\overline{\theta}c\right)\Gamma\left[c\left(1-\overline{\theta}\right)\right]}\theta^{c\overline{\theta}-1}(1-\theta)^{c\left(1-\overline{\theta}\right)-1}. \tag{1.39}$$

If gene frequencies vary at random in the clusters according to this beta distribution, the between-cluster variance of gene frequencies is given by

$$Var\left(\theta|\overline{\theta},c\right)=\frac{\overline{\theta}\left(1-\overline{\theta}\right)}{1+c}.$$

This variability can accommodate "extra-binomial" variation and account for the possible correlations between alleles that have been drawn. This example will be elaborated further later. ■

Gamma, Exponential, and Chi-square Distributions

The gamma distribution arises in quantitative genetic studies of variance components and of heritability through Bayesian methods. For example, Wang et al. (1994) analyzed a selection experiment for increased prolificacy in pigs in which the initial state of uncertainty about genetic and environmental variance components was represented by "inverted" gamma distributions (these are random variables whose reciprocals are distributed according to gamma distributions). A random variable X has a gamma distribution $Ga\left(a,b\right)$ if its p.d.f. has the form

$$p\left(x|a,b\right)=Ga\left(x|a,b\right)=Cx^{a-1}\exp\left[-bx\right],\quad x>0, \tag{1.40}$$

where $C=b^a/\Gamma\left(a\right)$; the "shape" parameter a and the "inverse scale" parameter b are positive. A special case of interest arises when $a=1$; then X is said to have an exponential distribution with parameter b and density

$$p\left(x|b\right)=b\exp\left(-bx\right),\quad x\geq 0. \tag{1.41}$$

The exponential distribution has been employed for modeling survival or length of productive life in livestock (Famula, 1981).

Another special case of interest is when $a=v/2$ and $b=1/2$; then X is said to possess a central chi-square distribution with positive parameter v (often known as degrees of freedom) and density

$$p\left(x|v\right)=Cx^{(v/2)-1}\exp\left(-x/2\right),\quad \text{for } v>0 \text{ and } x>0, \tag{1.42}$$

where the integration constant is

$$C=\frac{\left(1/2\right)^{v/2}}{\Gamma\left(v/2\right)}. \tag{1.43}$$

The chi-square process will be revisited in a section where quadratic forms on normal variables are discussed.

The kth ($k = 1, 2, ...$) moment from the origin of the gamma distribution is, by definition,

$$\begin{aligned} E\left(X^k|a,b\right) &= \int_0^\infty x^k \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\, dx \\ &= \frac{b^a}{\Gamma(a)} \int_0^\infty x^{k+a-1} \exp(-bx)\, dx. \end{aligned}$$

Let $X = Y^2$, so $dx = 2y\, dy$. Then

$$E\left(X^k|a,b\right) = \frac{2b^a}{\Gamma(a)} \int_0^\infty y^{2(k+a)-1} \exp\left(-by^2\right) dy.$$

Using the gamma integral in (1.25) with $p = 2(k + a)$,

$$\begin{aligned} E\left(X^k|a,b\right) &= \frac{2b^a}{\Gamma(a)} \int_0^\infty y^{2(k+a)-1} \exp\left(-by^2\right) dy \\ &= \frac{b^a}{\Gamma(a)} b^{-\frac{2(k+a)}{2}} \Gamma\left(\frac{2(k+a)}{2}\right). \end{aligned} \tag{1.44}$$

The mean of the distribution follows, by putting $k = 1$,

$$E(X|a,b) = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b} \tag{1.45}$$

because $\Gamma(a+1) = a\Gamma(a)$, as seen in (1.33). Similarly,

$$E\left(X^2|a,b\right) = \frac{\Gamma(a+2)}{b^2\Gamma(a)} = \frac{a(a+1)}{b^2}.$$

The variance of the distribution is thus,

$$Var(X|a,b) = \frac{a}{b^2}. \tag{1.46}$$

The coefficient of variation (ratio between the standard deviation and the mean of the distribution) is equal to $1/\sqrt{a}$, so it depends on only one of the parameters.

The Normal and Student-t Distributions

The normal distribution, without doubt, is the most important random process in statistical genetics. It has played a central role in model development (e.g., the infinitesimal model of inheritance; for a thorough discussion, see

Bulmer, 1980); selection theory (Pearson, 1903; Gianola et al., 1989), estimation of dispersion parameters (Henderson, 1953; Patterson and Thompson, 1971; Searle et al., 1992); prediction of genetic values (Henderson, 1963, 1973, 1975), inference about response to selection (Sorensen et al., 1994), and evaluation of hypotheses (Edwards, 1992; Rao, 1973). The density of a normally distributed random variable X with mean μ and variance σ^2, i.e., $N\left(x|\mu,\sigma^2\right)$, has been presented already, and is

$$p\left(x|\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\quad \text{for}-\infty<x<\infty, \tag{1.47}$$

with $-\infty<\mu<\infty$ and $\sigma>0$ being the parameters and their corresponding spaces. Since $p(\mu+\varepsilon)=p(\mu-\varepsilon)$, it is seen that the distribution is symmetric about μ, so the mean and median are equal, with $E\left(X|\mu,\sigma^2\right)=\mu$. The parameter σ is the standard deviation of X. The density has a maximum at $x=\mu$ (so the mean is equal to the mode and to the median) and, at this value, $p(\mu)=\left(\sigma\sqrt{2\pi}\right)^{-1}$. Furthermore, the density has inflection points (where the curve changes from concave to convex) at $\mu\pm\sigma$.

In order to derive the mean and variance, use is made of what is called the moment generating function. Let X be a random variable having some distribution, and consider a Taylor series expansion of $\exp(tX)$ about $X=0$, where t is a dummy variable. Then

$$\begin{aligned}\exp(tX) &= 1+\left\{t\exp(tX)\right\}_{X=0}X+\left\{t^2\exp(tX)\right\}_{X=0}\frac{X^2}{2!}+\cdots\\ &\quad\cdots+\left\{t^k\exp(tX)\right\}_{X=0}\frac{X^k}{k!}+\\ &= 1+tX+\frac{t^2X^2}{2!}+\cdots+\frac{t^kX^k}{k!}+\cdots\end{aligned}$$

The moment generating function is the average value of $\exp(tX)$:

$$E\left[\exp(tX)\right]=1+tE(X)+\frac{t^2E\left(X^2\right)}{2!}+\cdots+\frac{t^kE\left(X^k\right)}{k!}+\cdots \tag{1.48}$$

which is a linear combination of all moments of the distribution. Differentiation of (1.48) with respect to t, and setting $t=0$, yields $E(X)$. Likewise, differentiation of (1.48) twice with respect to the dummy variable, and setting $t=0$, gives $E\left(X^2\right)$; in general, k differentiations (followed by putting $t=0$) will produce $E\left(X^k\right)$. With this technique, moments can be found for most distributions, provided the moment generating function exists.

Example 1.9 *Finding the mean and variance of a normal process*
For the normal distribution, the moment generating function is

$$M(t) = \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx. \tag{1.49}$$

The exponents in the integral can be combined, by adding and subtracting terms and by completing a square, as follows:

$$\begin{aligned} tx - \frac{(x-\mu)^2}{2\sigma^2} &= tx - \left(\frac{x^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2} - \frac{2\mu x}{2\sigma^2}\right) + \left(\mu t + \frac{\sigma^2 t^2}{2}\right) \\ &\quad - \left(\mu t + \frac{\sigma^2 t^2}{2}\right) \\ &= \left(\mu t + \frac{\sigma^2 t^2}{2}\right) + \frac{2\sigma^2 tx}{2\sigma^2} \\ &\quad - \left[\left(\frac{x^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2} - \frac{2\mu x}{2\sigma^2}\right) - \left(\frac{2\sigma^2\mu t}{2\sigma^2} + \frac{\sigma^4 t^2}{2\sigma^2}\right)\right] \\ &= \left(\mu t + \frac{\sigma^2 t^2}{2}\right) \\ &\quad - \frac{1}{2\sigma^2}\left[x^2 + \mu^2 + \sigma^4 t^2 - 2\mu x - 2\sigma^2 tx - 2\sigma^2\mu t\right] \\ &= \left(\mu t + \frac{\sigma^2 t^2}{2}\right) - \frac{\left[x - (\mu + \sigma^2 t)\right]^2}{2\sigma^2}. \end{aligned}$$

Employing this in (1.49), one obtains

$$\begin{aligned} M(t) &= \frac{\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{\left[x - (\mu + \sigma^2 t)\right]^2}{2\sigma^2}\right\} dx \\ &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \end{aligned} \tag{1.50}$$

as the integral evaluates to $\sigma\sqrt{2\pi}$, the reciprocal of the integration constant of the normal distribution. The moments of the distribution can now be obtained by successive differentiation of $M(t)$ with respect to t, and then evaluating the result at $t = 0$. For example, the first moment is

$$\begin{aligned} E(X) &= M'(0) = \left[\frac{d}{dt}\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)\right]_{t=0} \\ &= \left[(\mu + \sigma^2 t)\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)\right]_{t=0} = \mu, \end{aligned}$$

yielding the mean of the distribution. Similarly, the second moment is

$$\begin{aligned} E\left(X^2\right) &= M''(0) = \left\{\left[\sigma^2 + \left(\mu + \sigma^2 t\right)^2\right] \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)\right\}_{t=0} \\ &= \sigma^2 + \mu^2. \end{aligned}$$

It follows that the variance of the distribution is $E\left(X^2\right) - E^2(X) = \sigma^2$, as anticipated. ■

The normal and Student-t distributions are intimately related to each other. If a random variable X has a $t\left(\mu, \sigma^2, \nu\right)$ distribution, its density function is

$$p\left(x|\mu, \sigma^2, \nu\right) = \frac{\Gamma\left[(\nu+1)/2\right]}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma}} \left[1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2}. \tag{1.51}$$

It can be seen that this distribution is symmetric about μ, its mean. Two additional positive parameters here are: the degrees of freedom ν and the scale of the distribution σ. The variance of a t process is

$$Var\left(X|\mu, \sigma^2, \nu\right) = \frac{\nu}{(\nu-2)}\sigma^2.$$

Thus, when ν is 2 or less the variance is infinite. When the degrees of freedom parameter goes to infinity, the process tends toward a normal $N\left(\mu, \sigma^2\right)$ one; when $\nu = 1$ the distribution is called Cauchy (whose mean and higher moments do not exist).

Derivation of density (1.51) requires consideration of the distribution of pairs of random variables. As shown in more detail later on in this chapter, if given μ and S_i^2, X_i is normally distributed

$$X_i|\mu, S_i^2 \sim N\left(\mu, S_i^2\right), \tag{1.52}$$

and if S_i^2 follows a scaled inverted chi-square distribution with parameters ν and σ^2:

$$S_i^2|\nu, \sigma^2 \sim \nu\sigma^2\chi_\nu^{-2}, \tag{1.53}$$

then the density

$$p\left(x_i|\mu, \sigma^2, \nu\right) = \int p\left(x_i|\mu, S_i^2\right) p\left(S_i^2|\nu, \sigma^2\right) dS_i^2 \tag{1.54}$$

is that of the $t\left(\mu, \sigma^2, \nu\right)$ distribution (1.51). The t distribution can therefore be interpreted as a mixture of normal distributions with a common mean and a variance that varies according to a scaled inverted chi-square.

This way of viewing the t distribution leads to a straightforward method for drawing Monte Carlo samples. First draw S^{2*} from (1.53); second, draw x_i^* from $\left[X_i|\mu, S^{*2}\right]$. Then x_i^* is a draw from (1.54) and repeating this procedure generates an i.i.d. sample from (1.54). This is known as the method of composition (Tanner, 1996).

1.4 Multivariate Probability Distributions

Results presented for a single random variable generalize in a fairly direct way to multivariate situations, that is, to settings in which it is desired to assign probabilities to events involving several random variables jointly. For example, consider the joint distribution of genotypes at two loci for a quantitative trait. A question of interest might be if genotype Bb, say, at locus B appears more frequently in association with genotype aa at locus A than with either genotype AA or Aa. Here one can let X be a random variable denoting genotype at locus A, and Y denote the genotype at locus B. The joint distribution of X and Y is called a bivariate distribution. Whenever the random process involves two or more random variables, one speaks of multivariate probability distributions.

We start this section by introducing the concept of independence. Let $(X_1, X_2, \ldots, X_n)$ be a random vector whose elements X_i $(i = 1, 2, \ldots, n)$ are one-dimensional random variables. Then

$$X_1, X_2, \ldots, X_n$$

are called mutually independent if, for every $(x_1, x_2, \ldots, x_n)$,

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2) \ldots p(x_n). \tag{1.55}$$

The term $p(x_i)$ is a marginal probability density. It is obtained by integrating (adding if the random variable is discrete) over the distribution of the remaining $(n-1)$ random variables

$$p(x_i) = \int p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_n, \tag{1.56}$$

where the integral is understood to be of dimension $(n-1)$. Mutual independence implies pairwise independence; however, pairwise independence does not imply mutual independence.

A generalization of the concept of i.i.d. random variables is that of exchangeable random variables, an idea due to De Finetti (1975b). The random variables $X_1, X_2, \ldots, X_n$ are exchangeable, if any permutation of any subset of size k of the random variables, for any $k \leq n$, has the same distribution. That is, for exchangeable random variables, any permutation of the indexes will lead to the same joint probability density. In practice, in a modeling context, the idea of exchangeability is associated with ignorance or symmetry: the less one knows about a problem, the more confidently one can make claims of exchangeability. For example, in rolling a die on which no information is available, one is prepared to assign equal probabilities to all six outcomes. Then the probability of obtaining a 1 and a 5, in two independent throws, should be the same as that of obtaining any other two possible outcomes. Note that random variables that are not independent

can be exchangeable. For example, if $X_1, X_2, \ldots, X_n$ are i.i.d. as Bernoulli $Br(\theta)$, then, given $\sum_{i=1}^{n} X_i = t$, $X_1, X_2, \ldots, X_n$ are exchangeable, but not independent.

For simplicity, the remainder of this section motivates developments using a bivariate situation. Analogously to the univariate case, the joint c.d.f. of two random variables X and Y is defined to be

$$F(x, y) = \Pr(X \leq x, Y \leq y), \tag{1.57}$$

where $\Pr(X \leq x, Y \leq y)$ means "probability that X takes a value smaller than or equal to x and Y takes a value smaller than or equal to y". If the two random variables are discrete, their joint probability distribution is given by

$$\Pr(X = x, Y = y) = p(x, y). \tag{1.58}$$

When the two random variables are continuous, a density function must be introduced, as in the univariate case, because the probability that $X = x$ and $Y = y$ is 0. The joint p.d.f. of random variables (X, Y), that take value in $\mathbb{R}^2$, is, by definition,

$$p(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \tag{1.59}$$

provided the distribution function $F(\cdot)$ is differentiable (in this case at least twice). The joint p.d.f. $p(x, y)$ is defined for all (x, y) in $\mathbb{R}^2$. The joint probability that (X, Y) belong to a subset A of $\mathbb{R}^2$ is given by the two-fold integral

$$\begin{aligned} \Pr[(X, Y) \in A] &= \int\int_A p(x, y)\, dx\, dy \\ &= \int\int I[(x, y) \in A]\, p(x, y)\, dx\, dy. \end{aligned}$$

For example, if $A = (a_1, b_1) \times (a_2, b_2)$, then the integral above is

$$\begin{aligned} \Pr[(X, Y) \in A] &= \int\int I[x \in (a_1, b_1), y \in (a_2, b_2)]\, p(x, y)\, dx\, dy \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} p(x, y)\, dx\, dy. \end{aligned}$$

Conditional distributions play an important role when making inferences about nonobservable quantities given information on observable quantities. The conditional probability distribution of X given $Y = y$ is the function of X given by

$$p(x|y) = \frac{p(x, y)}{p(y)}, \quad p(y) > 0. \tag{1.60}$$

In the case $p(y) = 0$, the convention is that $p(x|y) = p(x)$. Expression (1.60) implies that

$$p(x, y) = p(x|y)\, p(y)$$

regardless of whether $p(y) > 0$. In the case of continuous random variables, notice that (1.60) integrates to 1:

$$\begin{aligned} \int p(x|y)\, dx &= \frac{\int p(x, y)\, dx}{p(y)} \\ &= \frac{p(y)}{p(y)} = 1. \end{aligned}$$

Because $p(x|y)$ is a function of X at the fixed value $Y = y$, one can also write

$$p(x|y) \propto p(x, y) \tag{1.61}$$

as the denominator does not depend on x. This result holds both for discrete and continuous random variables, and highlights the fact that a joint distribution must be the starting point in a multivariate analysis. Again, when X and Y are statistically independent, $p(x, y) = p(x)p(y)$ and

$$p(x|y) = \frac{p(x)\, p(y)}{p(y)} = p(x). \tag{1.62}$$

In the case of n mutually independent random variables $X_1, X_2, \ldots, X_n$, the conditional distribution of any subset of the coordinates, given the value of the rest, is the same as the marginal distribution of the subset.

In the discrete two-dimensional case, the marginal distribution of X is given by

$$\Pr(X = x) = p(x) = \sum_y p(x, y) = \sum_y p(x|y)\, p(y). \tag{1.63}$$

Above, the sum is over all possible values of Y, each weighted by its marginal probability $\Pr(Y = y)$. For instance, if one were interested in finding the marginal distribution of genotypes at locus A from the joint distribution of genotypes at the two loci, application of (1.63) yields

$$\Pr(X = x) = p(x|BB)p(BB) + p(x|Bb)p(Bb) + p(x|bb)p(bb)$$

where $x = AA$, Aa, or aa. This gives the total probability that $X = x$ (Aa, say) as the sum of the probabilities of the three ways in which this can occur, i.e., when $X = Aa$ jointly with Y being either BB, Bb, or bb. Further, the conditional probability that $Y = BB$ given that $X = Aa$ can

be written, making use of (1.60), as

$$\begin{aligned}&\Pr(Y = BB|X = Aa)\\ &= \frac{p(Aa|BB)p(BB)}{p(Aa|BB)p(BB) + p(Aa|Bb)p(Bb) + p(Aa|bb)p(bb)}\\ &= \frac{p(BB|Aa)p(Aa)}{p(Aa|BB)p(BB) + p(Aa|Bb)p(Bb) + p(Aa|bb)p(bb)}.\end{aligned}$$

In the continuous case, the counterpart of (1.63) is the two-dimensional analogue of (1.56), and is obtained by integrating over the distribution of Y:

$$p(x) = \int p(x, y)\, dy = \int p(x|y)\, p(y)\, dy. \tag{1.64}$$

Results (1.60) and (1.62) are fairly general, and apply to either scalar or vector variates. The random variables can be either discrete or continuous, or one can be discrete while the other continuous. For example, suppose Y is the genotype at a molecularly marked locus (discrete random variable) and X is the breeding value at a QTL for growth rate in pigs (continuous). If the observation that $Y = y$ alters our knowledge about the distribution of breeding values, then one can exploit this stochastic dependence in a marker-assisted selection program for faster growth. Otherwise, if knowledge of the marked locus does not contribute information about growth rate, we would be in the situation depicted by (1.62).

Often, a joint density can be identified only up to proportionality. As in the univariate case, one can write $p(x, y) = Cf(x, y)$, where C is a constant (i.e., it does not depend on X and Y), and $f(x, y)$ is known as the kernel of the joint density. Hence,

$$p(x, y) = \frac{f(x, y)}{\int\int f(x, y)\, dx\, dy}. \tag{1.65}$$

It follows that

$$C^{-1} = \int\int f(x, y)\, dx\, dy. \tag{1.66}$$

Before reviewing standard multivariate distributions, a number of examples are discussed to illustrate some of the ideas presented so far.

Example 1.10 *Aitken's integral and the bivariate normal distribution*
A useful result in multivariate normal theory is Aitken's integral (e.g., Searle, 1971). Here assume that the random variables X and Y possess what is called a bivariate normal distribution, to be discussed in more detail below. Such a process can be represented as

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right),$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, and

$$\mathbf{V} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

is the 2×2 positive-definite variance–covariance matrix of the distribution. The joint density of X and Y is

$$p(x, y|\mu_X, \mu_Y, \mathbf{V}) = \frac{1}{2\pi |\mathbf{V}|^{\frac{1}{2}}}$$
$$\times \exp\left[-\frac{1}{2} \begin{pmatrix} X - \mu_X, & Y - \mu_Y \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}\right]. \tag{1.67}$$

The kernel of the density is the expression $\exp[\cdot]$, and Aitken's integral is

$$\int\int \exp\left[-\frac{1}{2} \begin{pmatrix} X - \mu_X, & Y - \mu_Y \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}\right] dx\, dy$$
$$= \quad 2\pi |\mathbf{V}|^{\frac{1}{2}}$$

which follows because $p(x, y|\mu_X, \mu_Y, \mathbf{V})$ integrates to 1. Then $C^{-1} = 2\pi |\mathbf{V}|^{\frac{1}{2}}$, and the bivariate normal density can be represented as

$$p(x, y|\mu_X, \mu_Y, \mathbf{V}) =$$
$$\frac{\exp\left[-\frac{1}{2} \begin{pmatrix} X - \mu_X, & Y - \mu_Y \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}\right]}{\int\int \exp\left[-\frac{1}{2} \begin{pmatrix} X - \mu_X, & Y - \mu_Y \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}\right] dx\, dy}.$$

■

Example 1.11 *A discrete bivariate distribution*

This example is elaborated after Casella and Berger (1990). Mastitis is a serious disease of the mammary gland in milk producing animals, and it has important economic consequences in dairy farming. Suppose that in a breed of sheep, the disease appears in three mutually exclusive and exhaustive modalities: absent, subclinical, or clinical. Let Y be a random variable denoting the disease status. Also, let X be another random variable taking two values only, $X = 1$ or $X = 2$, representing ewes that are born as singles or twins, respectively. From knowledge of the population, it is known that the bivariate random vector (X, Y) has a joint probability distribution given by:

	$Y = 1$	$Y = 2$	$Y = 3$	$p(x)$
$X = 1$	1/10	2/10	2/10	5/10
$X = 2$	1/10	1/10	3/10	5/10
$p(y)$	2/10	3/10	5/10	

The entries in this table give the joint probabilities of the six possible events. The element in the first row and column represents

$$\Pr(X = 1, Y = 1) = 1/10.$$

The last column gives the marginal probability distribution of X, and the last row presents the marginal probability distribution of Y. For example, the marginal probability of an animal born as a single ($X = 1$) is obtained as

$$\begin{aligned}\Pr(X = 1) = \Pr(X = 1, Y = 1) + \Pr(X = 1, Y = 2) \\ + \Pr(X = 1, Y = 3) = \frac{5}{10}.\end{aligned}$$

This can also be obtained from

$$\begin{aligned}\Pr(X = 1) &= \sum_{i=1}^{i=3} \Pr(X = 1|Y = y_i) \Pr(Y = y_i) \\ &= \left(\frac{1}{2}\frac{2}{10}\right) + \left(\frac{2}{3}\frac{3}{10}\right) + \left(\frac{2}{5}\frac{5}{10}\right) = \frac{5}{10}.\end{aligned}$$

The random variables X and Y are not independent because

$$p(x, y) \neq p(x)\, p(y).$$

For example,

$$p(1, 3) = \frac{1}{5} \neq \frac{1}{2}\frac{1}{2} = \Pr(X = 1) \Pr(Y = 3).$$

Note that there are some (X, Y) values for which their joint probability is equal to the product of their marginals. For example,

$$p(1, 1) = \frac{1}{10} = \frac{1}{2}\frac{1}{5} = \Pr(X = 1) \Pr(Y = 1).$$

However, this does not ensure independence, as seen from the case of $p(1, 3)$. All values must be checked. ■

Example 1.12 *Two independent continuous random variables*
Suppose the following function is a suitable density arising in the description of uncertainty about the pair of continuous random variables X and Y

$$p(x, y) = \begin{cases} 6xy^2, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The joint sample space can be viewed as a square of unit length. We check first whether this function is suitable as a joint p.d.f., by integrating it over

the entire sample space

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\,dx\,dy &= \int_{0}^{1}\int_{0}^{1} 6xy^2 dx\,dy \\ &= \int_{0}^{1} 3y^2 dy = y^3\big|_0^1 \\ &= 1.\end{aligned}$$

It is not essential that the function integrates to 1; the only requirement is that the integral must be finite. For example, if it integrates to K, say, then the joint p.d.f. would be $6xy^2/K$. When a density function does not integrate to a finite value, the corresponding distribution is said to be improper. Improper distributions arise in certain forms of Bayesian analysis, as will be seen later on. Employing (1.64), the marginal density of Y is obtained by integrating the joint density $p(x, y)$ over the distribution of X

$$p(y) = \int_0^1 p(x,y)\,dx = \int_0^1 6xy^2\,dx = 6y^2\frac{x^2}{2}\bigg|_0^1 = 3y^2.$$

It can be verified that the resulting density integrates to 1. We say that X has been "integrated out" of the joint density; the resulting marginal p.d.f. of Y is, therefore, not a function of X. This is important in a Bayesian context, where "nuisance parameters" are eliminated by integration. For example, if X were a nuisance parameter in a Bayesian model, the marginal density of the parameter of interest (Y in this case) would not depend on X. Likewise, the marginal density of X is

$$p(x) = \int_0^1 p(x,y)\,dy = \int_0^1 6xy^2\,dy = 2x,$$

and this integrates to 1 as well. The density function of the conditional distribution of Y given X is, using (1.60),

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{6xy^2}{2x} = 3y^2, \text{ for } 0 < y < 1.$$

Similarly, the conditional density of X given Y is

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{6xy^2}{3y^2} = 2x.$$

Hence, the conditional distribution $[Y|X]$ does not depend on X, and the conditional distribution $[X|Y]$ does not depend on Y (in general, the notation $[A|B, C]$ is employed to refer to the conditional distribution of random

variable A given random variables B and C). This is because X and Y are independent random variables: their joint density is obtained by multiplication of the marginal densities

$$p(x,y) = p(x)\,p(y) = 2x3y^2 = 6xy^2.$$

Because of independence, $p(y|x) = p(y)$ and $p(x|y) = p(x)$, as already seen.

We can calculate the regression curve, or regression function, of Y on X, which is denoted by $E(Y|X=x)$. This function, using the assumptions made in the preceding example, is

$$E(Y|X=x) = \int_0^1 yp(y|x)\,dy = \int_0^1 y3y^2\,dy = \frac{3}{4}. \tag{1.68}$$

This is a constant because Y is independent of X, so $E(Y|x) = E(Y)$ here. The uncertainty about the distribution of Y, once we know that $X = x$, can be assessed through the variance of the distribution $[Y|X]$, that is, by $Var(Y|X=x)$. By definition

$$Var(Y|X=x) = E\left(Y^2|x\right) - E^2(Y|x).$$

For the example above, having computed $E(Y|x) = E(Y) = 3/4$, we need now

$$E\left(Y^2|x\right) = \int_0^1 y^2\,p(y|x)\,dy = \int_0^1 y^2\,p(y)\,dy = E\left(Y^2\right) = \frac{3}{5}$$

so

$$Var(Y|X=x) = 3/5 - [3/4]^2 = 3/80,$$

which is equal to the variance of the distribution of Y in this case. As expected (because of the independence noted), the variance does not depend on X. Here, it is said that the dispersion is homoscedastic or constant throughout the regression line. ■

Example 1.13 *A continuous bivariate distribution*

Two random variables have as joint p.d.f.

$$p(x,y) = \begin{cases} x+y, & 0 \le x \le 1, 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

The marginal densities of X and Y are

$$\begin{aligned} p(x) &= \int_0^1 p(x,y)\,dy = x + \frac{1}{2}, \\ p(y) &= y + \frac{1}{2}. \end{aligned}$$

The probability that X is between $1/2$ and 1 can be found to be equal to

$$\int_{\frac{1}{2}}^{1} (x + 1/2)\, dx = 5/8.$$

The probability that X is between $1/2$ and 1 and that Y is between 0 and $1/2$ is

$$\int_{\frac{1}{2}}^{1} \int_{0}^{\frac{1}{2}} (x + y)\, dx\, dy = 1/4.$$

The variables X and Y are not independent, as $p(x, y) \neq p(x)\, p(y)$. The conditional density of Y given x is

$$p(y|x) = \frac{p(y, x)}{p(x)} = \frac{y + x}{x + \frac{1}{2}}.$$

The regression function of Y on X is

$$E(Y|X = x) = \int_{0}^{1} yp(y|x)\, dy = \int_{0}^{1} \frac{y(y + x)}{x + \frac{1}{2}} dy = \frac{3x + 2}{3(1 + 2x)},$$

which is a decreasing, nonlinear function of X. The variance of the conditional distribution $[Y|X]$ can be found to be equal to

$$Var(Y|X = x) = \frac{\left(1 + 6x + 6x^2\right)}{\left[18(1 + 2x)^2\right]}, \quad 0 \leq x \leq 1.$$

Here the conditional distribution is not homoscedastic. For example, for $x = 0$ and $x = 1$, the conditional variance is equal to $1/18$ and $13/162$, respectively. ■

1.4.1 The Multinomial Distribution

The binomial distribution is generalized to the multinomial distribution as follows. Let $C_1, C_2, \ldots, C_k$ represent k mutually exclusive and exhaustive classes or outcomes. Imagine an experiment consisting of n independent trials and, in each trial, only one of these k distinct outcomes is possible. The probability of occurrence of outcome i is p_i $(i = 1, 2, \ldots, k)$ on every trial. Let X_i be a random variable corresponding to the number (counts) of times that the ith outcome occurred in the n trials. When $k = 2$, this is a binomial experiment and X_1 counts the number of "successes" and $X_2 = n - X_1$ counts the number of "failures" in the n independent trials.

The p.m.f. of the random vector $(X_1, X_2, \ldots, X_k)$ is

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k | X_1 + X_2 + \cdots + X_k = n) \\ = \frac{n!}{x_1!\, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \end{aligned} \tag{1.69}$$

with $\sum_{i=1}^k x_i = n$, $\sum_{i=1}^k p_i = 1$. Therefore an alternative way of writing (1.69) is

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k | X_1 + X_2 + \cdots + X_k = n) \\ = \frac{n! p_1^{x_1} p_2^{x_2} \ldots (1 - p_1 - \cdots - p_{k-1})^{(n - x_1 - \cdots - x_{k-1})}}{x_1! x_2! \ldots (n - x_1 - \cdots - x_{k-1})!}. \end{aligned}$$

Observe that the k random variables $(X_1, X_2, \ldots, X_k)$ are not independent; any $k-1$ of them determines the kth. The mean and (co)variance of this distribution are

$$\begin{aligned} E(X_i) &= np_i, \\ Var(X_i) &= np_i(1 - p_i), \\ Cov(X_i, X_j) &= -np_i p_j, \quad i \neq j. \end{aligned}$$

The n trials can be divided into two classes: X_i counts belonging to outcome i, and the remaining events corresponding to "non-i". The marginal distribution of the random variable X_i is binomial with p.m.f.

$$Bi(x_i | p_i, n). \tag{1.70}$$

Further, given X_i, the vector $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ is multinomially distributed

$$\mathbf{X}_{-i} | X_i \sim Mu(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_k, n - X_i), \tag{1.71}$$

where $q_j = p_j/(1 - p_i)$.

Let

$$\mathbf{X}_{-i,-j} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k).$$

Then the p.m.f. of the distribution $[\mathbf{X}_{-i,-j} | X_i, X_j]$ is,

$$Mu(\mathbf{x}_{-i,-j} | r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_{j-1}, r_{j+1}, \ldots, r_k, n - X_i - X_j) \tag{1.72}$$

where $r_s = p_s/(1 - p_i - p_j)$.

As in the binomial case, a generalization of (1.16) exists. As $n \to \infty$, the vector of observed responses $\mathbf{X}$ will tend to a multivariate normal distribution (to be discussed below) with mean vector

$$E(\mathbf{X}) = \{np_i\}, \quad i = 1, \ldots, k, \tag{1.73}$$

and covariance matrix

$$\mathbf{V} = n \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{bmatrix}. \tag{1.74}$$

Summing the elements of any row of (1.74), the ith, say, yields

$$np_i(1-p_1-p_2-\cdots-p_k) = 0.$$

Hence $\mathbf{V}$ has rank $k-1$. A generalized inverse of $\mathbf{V}$ can readily be shown (Stuart and Ord, 1991) to be equal to

$$\mathbf{V}^- = \frac{1}{n} \begin{bmatrix} \widetilde{\mathbf{V}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\widetilde{\mathbf{V}}^{-1}$ is a $(k-1) \times (k-1)$ matrix with diagonal elements $(1/p_i) + (1/p_k)$ $(i = 1, 2, \ldots, k-1)$ and off-diagonals $1/p_k$. Therefore the exponent of the limiting singular multivariate normal distribution has the following form:

$$(\mathbf{x} - n\mathbf{p})' \mathbf{V}^- (\mathbf{x} - n\mathbf{p}), \tag{1.75}$$

where $\mathbf{p}' = (p_1, \ldots, p_k)$. For large n, the random variable defined in (1.75) has a central chi-square distribution with $tr(\mathbf{V}^-\mathbf{V}) = (k-1)$ degrees of freedom (Searle, 1971). This gives the basis for the usual test of goodness of fit.

Example 1.14 *Generating samples from the multinomial distribution*
A general procedure for simulating multivariate random variables is based on the composition or conditional distribution method (Devroye, 1986). Let

$$\mathbf{X} = (X_1, \ldots, X_d)$$

denote a d-dimensional random vector with p.d.f. $f(\mathbf{x})$. Then,

$$f(\mathbf{x}) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \ldots f(x_d|x_1, \ldots, x_{d-1}).$$

If the marginal distributions and all the conditional distributions are known, this method allows reducing a multivariate generation problem to d univariate generations.
As an example, consider generating a multinomial random variable

$$(X_1, X_2, X_3, X_4) \sim Mu(p_1, p_2, p_3, p_4, n)$$

with $\sum_{i=1}^4 p_i = 1$ and $\sum_{i=1}^4 X_i = n$. Using (1.70), (1.71), and (1.72), one proceeds as follows. Generate

$$X_1 \sim Bi(p_1, n),$$

$$X_2|X_1 = x_1 \sim Bi\left(\frac{p_2}{1-p_1}, n - x_1\right),$$

$$X_3|X_1 = x_1, X_2 = x_2 \sim Bi\left(\frac{p_3}{1-p_1-p_2}, n - x_1 - x_2\right).$$

Set

$$X_4 = n - x_1 - x_2 - x_3.$$

The vector (x_1, x_2, x_3, x_4) is a realized value from $\mathbf{X}$. If at any time in the simulation $n = 0$, use the convention that a $Bi(p, 0)$ random variable is identically zero (Gelman et al., 1995). ■

1.4.2 The Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the beta distribution. For example, as discussed in Chapter 11, Example 11.7, the Dirichlet is a natural model for the distribution of gene frequencies at a locus with more than two alleles. The random vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_k),$$

$X_1, X_2, \ldots, X_k \geq 0$, $\sum_{i=1}^{k} X_i = 1$, follows the Dirichlet distribution of dimension k, with parameters

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k), \quad \alpha_j > 0,$$

if its probability density $Di_k(\mathbf{x}|\boldsymbol{\alpha})$ is

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}. \tag{1.76}$$

A simple and efficient way to sample from (1.76) is first to draw k independent gamma random variables with scale parameter α_i and unit scale: $Ga(y_i|\alpha_i, 1)$. Then form the ratios

$$x_i = \frac{y_i}{\sum_{j=1}^{k} y_i}, \quad i = 1, \ldots, k.$$

The vector $(x_1, x_2, \ldots, x_k)$ is a realized value from (1.76).

1.4.3 The d-Dimensional Uniform Distribution

A $d \times 1$ random vector $\mathbf{x}$ is uniformly distributed on $[0, 1]^d$ if

$$p(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in [0, 1]^d, \\ 0, & \text{otherwise.} \end{cases}$$

From (1.56), each scalar element of $\mathbf{x}$ is distributed as $[0, 1]$. Often, multidimensional uniform distributions are assigned as prior distributions to some of the parameters of a Bayesian model.

1.4.4 The Multivariate Normal Distribution

In this subsection, the multivariate normal distribution is introduced, followed by a presentation of the marginal and conditional distributions induced by this process, of its moment generating function, of the distribution of linear combinations of normal variates, and by a simple derivation of the central limit theorem. The subsection concludes with examples that illustrate applications of the multivariate normal distribution in quantitative genetics.

Density, Mean Vector, and Variance–Covariance Matrix

Let $\mathbf{y}$ denote a random vector of dimension n (the notational distinction between a random variable and its realized value is omitted here). This vector is said to have an n-dimensional multivariate normal distribution if its p.d.f. is

$$p(\mathbf{y}|\mathbf{m},\mathbf{V}) = |2\pi\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m})\right] \tag{1.77}$$

where

$$\mathbf{m} = E(\mathbf{y}|\mathbf{m},\mathbf{V}) = \int \mathbf{y}p(\mathbf{y}|\mathbf{m},\mathbf{V})\,d\mathbf{y} \tag{1.78}$$

is the mean vector, and

$$\begin{aligned}\mathbf{V} &= E[(\mathbf{y}-\mathbf{m})(\mathbf{y}-\mathbf{m})'] = \int (\mathbf{y}-\mathbf{m})(\mathbf{y}-\mathbf{m})'p(\mathbf{y}|\mathbf{m},\mathbf{V})\,d\mathbf{y} \\ &= \int \mathbf{y}\mathbf{y}'p(\mathbf{y}|\mathbf{m},\mathbf{V})\,d\mathbf{y} - \mathbf{m}\mathbf{m}'\end{aligned} \tag{1.79}$$

is the variance–covariance matrix of the distribution, assumed to be nonsingular. All integrals are n-dimensional, and taken over $\mathbb{R}^n$, the entire n-dimensional space. The notation $d\mathbf{y}$ is used for $dy_1 dy_2 \ldots dy_n$. From the density in (1.77) it follows that

$$\int \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m})\right] d\mathbf{y} = |2\pi\mathbf{V}|^{1/2}.$$

This is a generalization of Aitken's integral seen in Example 1.10

Marginal and Conditional Distributions

Partition the vector of random variables as $\mathbf{y} = [\mathbf{y}_1', \mathbf{y}_2']'$, with the corresponding partitions of $\mathbf{m}$ and $\mathbf{V}$ being

$$\begin{aligned}\mathbf{m} &= [\mathbf{m}_1', \mathbf{m}_2']' \\ \mathbf{V} &= \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}.\end{aligned}$$

For any arbitrary partition, it can be shown that all marginal distributions are normal. For example, the marginal density of $\mathbf{y}_1$ is

$$p(\mathbf{y}_1|\mathbf{m}_1,\mathbf{V}_{11}) = \int_{-\infty}^{\infty} p(\mathbf{y}_1,\mathbf{y}_2|\mathbf{m},\mathbf{V})\,d\mathbf{y}_2$$
$$= (2\pi)^{-\frac{n_1}{2}}|\mathbf{V}_{11}|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{y}_1-\mathbf{m}_1')'\mathbf{V}_{11}^{-1}(\mathbf{y}_1-\mathbf{m}_1)\right], \tag{1.80}$$

where n_1 is the order of $\mathbf{y}_1$. The conditional distribution

$$[\mathbf{y}_1|\mathbf{y}_2,\mathbf{m},\mathbf{V}]$$

is normal as well, with density

$$p(\mathbf{y}_1|\mathbf{y}_2,\mathbf{m},\mathbf{V}) = (2\pi)^{-\frac{n_1}{2}}|Var(\mathbf{y}_1|\mathbf{y}_2)|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}[\mathbf{y}_1-E(\mathbf{y}_1|\mathbf{y}_2)]'Var^{-1}(\mathbf{y}_1|\mathbf{y}_2)[\mathbf{y}_1-E(\mathbf{y}_1|\mathbf{y}_2)]\right\}. \tag{1.81}$$

This holds for any partition of $\mathbf{y}$, irrespective of whether the components are scalars or vectors. The mean vector and covariance matrix of this conditional distribution are

$$E(\mathbf{y}_1|\mathbf{y}_2,\mathbf{m},\mathbf{V}) = \mathbf{m}_1+\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{y}_2-\mathbf{m}_2) \tag{1.82}$$

and

$$Var(\mathbf{y}_1|\mathbf{y}_2,\mathbf{m},\mathbf{V}) = \mathbf{V}_{11}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}, \tag{1.83}$$

respectively. The variance–covariance matrix does not depend on $\mathbf{y}_2$, so this conditional distribution is homoscedastic. This is an important feature of the multivariate normal distribution. Another important fact to be noticed is that marginal and conditional normality is arrived at assuming bivariate normality as point of departure. However, marginal normality does not imply joint normality.

A sufficient condition for independence in the multivariate normal distribution is that the variance–covariance matrix is diagonal. For example, suppose that

$$\mathbf{V} = \mathbf{I}\sigma^2,$$

where σ^2 is a positive scalar, and that $E(y_i) = m$ $(i = 1, 2, ..., n)$. Then

$$(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-m)^2$$

and

$$|2\pi\mathbf{V}|^{-\frac{1}{2}} = \left|2\pi\mathbf{I}\sigma^2\right|^{-\frac{1}{2}} = \left(2\pi\sigma^2\right)^{-\frac{n}{2}}|\mathbf{I}| = \left(2\pi\sigma^2\right)^{-\frac{n}{2}}.$$

Using this in (1.77) yields

$$p(\mathbf{y}|\mathbf{m},\mathbf{V}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{\sum_{i=1}^{n}(y_i-m)^2}{2\sigma^2}\right] = \prod_{i=1}^{n} p\left(y_i|m,\sigma^2\right)$$

which is the product of the densities of n i.i.d. normal variates, each with mean m and variance σ^2.

If the matrix $\mathbf{V}$ is singular, then $\mathbf{y}$ is said to have a singular or degenerate normal distribution. If the rank of $\mathbf{V}$ is $k < n$, the singular density can be written as

$$\frac{(2\pi)^{-k/2}}{(\lambda_1\lambda_2\ldots\lambda_k)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-}(\mathbf{y}-\mathbf{m})\right],$$

where $\mathbf{V}^{-}$ is a generalized inverse of $\mathbf{V}$ and $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the nonzero eigenvalues of $\mathbf{V}$. Details can be found, for example, in Searle (1971), Rao (1973), Anderson (1984), and Mardia et al. (1979).

Moment Generating Function

Let $\mathbf{y} \sim N(\mathbf{m},\mathbf{V})$. The formula for the univariate case in (1.49) extends to the multivariate normal situation, after similar algebra (Searle, 1971), to

$$M(\mathbf{t}) = E[\exp(\mathbf{t}'\mathbf{y})] = \frac{1}{|2\pi\mathbf{V}|^{1/2}}$$
$$\times \int_{-\infty}^{\infty} \exp[\mathbf{t}'\mathbf{y}] \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m})\right] d\mathbf{y}$$
$$= \exp\left(\mathbf{t}'\mathbf{m} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right), \tag{1.84}$$

where $\mathbf{t}$ is a dummy vector of order n. Differentiation of the moment generating function with respect to $\mathbf{t}$ yields

$$\frac{\partial M(\mathbf{t})}{\partial \mathbf{t}} = \exp\left(\mathbf{t}'\mathbf{m} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right)(\mathbf{m}+\mathbf{V}\mathbf{t})$$

and putting $\mathbf{t} = \mathbf{0}$ gives directly $E(\mathbf{y}) = \mathbf{m}$. An additional differentiation gives

$$\frac{\partial^2 M(\mathbf{t})}{\partial \mathbf{t}\,\partial \mathbf{t}'} = \exp\left(\mathbf{t}'\mathbf{m} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right)\left[\mathbf{V} + (\mathbf{m}+\mathbf{V}\mathbf{t})(\mathbf{m}+\mathbf{V}\mathbf{t})'\right]$$

which, for $\mathbf{t} = \mathbf{0}$, leads to $E(\mathbf{y}\mathbf{y}') = \mathbf{V} + \mathbf{m}\mathbf{m}'$. The covariance matrix is then $E(\mathbf{y}\mathbf{y}') - E(\mathbf{y})E(\mathbf{y}') = \mathbf{V}$.

Linear Functions of Normally Distributed Random Variables

Another important property of the multivariate Gaussian distribution is that a linear transformation of a normally distributed vector is normal as well. Let $x = \alpha + \boldsymbol{\beta}'\mathbf{y}$ be a random variable resulting from a linear combination of the multivariate normal vector $\mathbf{y}$, where α is a known scalar and $\boldsymbol{\beta}$ is a vector, also known. The moment generating function of x is then

$$E\left\{\exp\left[\left(\alpha+\boldsymbol{\beta}'\mathbf{y}\right)t\right]\right\} = \exp\left(\alpha t\right) E\left[\exp\left(t\boldsymbol{\beta}'\mathbf{y}\right)\right] = \exp\left(\alpha t\right) M\left(\mathbf{t}^*\right),$$

where $\mathbf{t}^* = t\boldsymbol{\beta}$. Making use of (1.84) in the preceding expression gives

$$\begin{aligned} E\left\{\exp\left[\left(\alpha+\boldsymbol{\beta}'\mathbf{y}\right)t\right]\right\} &= \exp\left(\alpha t\right)\exp\left(\mathbf{t}^{*\prime}\mathbf{m}+\frac{\mathbf{t}^{*\prime}\mathbf{V}\mathbf{t}^*}{2}\right) \\ &= \exp\left[t\left(\alpha+\boldsymbol{\beta}'\mathbf{m}\right)+\frac{t^2\left(\boldsymbol{\beta}'\mathbf{V}\boldsymbol{\beta}\right)}{2}\right]. \end{aligned}$$

This is the moment generating function of a normal random variable with mean $\alpha + \boldsymbol{\beta}'\mathbf{m}$ and variance $\boldsymbol{\beta}'\mathbf{V}\boldsymbol{\beta}$. The property is important in quantitative genetics under additive inheritance: here, the additive genetic value of an offspring is equal to the average value of the parents, plus a residual. If all these terms follow a multivariate normal distribution, it follows that the additive value in the progeny is normally distributed as well.

Central Limit Theorem

As stated, the normal distribution has played a central role in statistics and quantitative genetics. An example is the so-called infinitesimal model (Fisher, 1918; Bulmer, 1971). Here, $\mathbf{y}$ in (1.77) represents a vector of additive genetic values and $\mathbf{V}$ is a function of additive genetic relationships between subjects, or of twice the coefficients of coancestry (Malécot, 1969). This matrix of coancestries enters when quantifying the process of genetic drift, in estimation of genetic variance–covariance components, and in inferences about additive genetic values and functions thereof in animal and plant breeding.

The infinitesimal model posits that the additive genetic value for a quantitative trait is the result of the sum of values at each of an infinite number loci. If the population in question is in joint equilibrium at all loci as a result of random mating without selection over many generations, the contributions from the different loci will be statistically independent from each other. This will be true even under the presence of linkage, since linkage slows down the approach to equilibrium but does not change the equilibrium ultimately attained. The celebrated central limit theorem leads to the result that the additive genetic value follows a normal distribution, approximately, irrespective of the distribution of effects at individual loci. Here, borrowing from Bulmer (1979), it is shown that this is so.

Let $Y = \sum_{i=1}^{n} Y_i$ be the additive genetic value of an individual, and let Y_i be the value at locus i. The mean and variance at locus i are μ_i and σ_i^2, respectively, so that $E(Y) = \sum_{i=1}^{n} \mu_i = \mu$ and $Var(Y) = \sum_{i=1}^{n} \sigma_i^2 = \sigma^2$, say, as the effects at different loci are mutually independent of each other. Consider the moment generating function of the standardized variate $Z = (Y - \mu)/\sigma$:

$$M_Z(t) = E[\exp(tZ)] = E\left\{\exp\left[\sum_{i=1}^{n} tZ_i\right]\right\},$$

where $Z_i = (Y_i - \mu_i)/\sigma$ and $\sum_i Z_i = Z$. Further,

$$\begin{aligned} M_Z(t) &= \int \exp\left[\sum_{i=1}^{n} tZ_i\right] p(z_1, z_2, ..., z_n)\, d\mathbf{z} \\ &= \prod_{i=1}^{n} E\left\{\exp\left[\frac{t(Y_i - \mu_i)}{\sigma}\right]\right\} = \prod_{i=1}^{n} M_i\left(\frac{t}{\sigma}\right), \end{aligned} \tag{1.85}$$

where $M_i(t)$ is the moment generating function of $(Y_i - \mu_i)$. The preceding follows because the n random variables Y_i are mutually independent. Using (1.48), one can write

$$\begin{aligned} M_i\left(\frac{t}{\sigma}\right) &= E\left[\frac{t(Y_i - \mu_i)}{\sigma}\right] \\ &= 1 + \frac{tE(Y_i - \mu_i)}{\sigma} + \frac{t^2 E(Y_i - \mu_i)^2}{2\sigma^2} + \cdots \\ &\quad \cdots + \frac{t^k E(Y_i - \mu_i)^k}{k!\sigma^k} + \cdots \\ &\approx 1 + \frac{t^2\sigma_i^2}{2\sigma^2} + \cdots. \end{aligned} \tag{1.86}$$

This is so, because for large n, third- and higher-order moments (from the mean) for small individual loci effects are small, relative to σ^3, σ^4, etc., so the corresponding terms can be ignored. Then, employing (1.86) in (1.85) gives

$$\log[M_Z(t)] = \sum_{i=1}^{n} \log\left[M_i\left(\frac{t}{\sigma}\right)\right] \approx \sum_{i=1}^{n} \log\left(1 + \frac{t^2\sigma_i^2}{2\sigma^2}\right). \tag{1.87}$$

Now consider an expansion about 0 of the function

$$\log(1 + x) = \log(1) + x - x^2 + 2x^3 + \cdots \approx x.$$

The approximation results from the fact that for values of x near 0, higher-order terms can be neglected. Using this in (1.87) gives

$$M_Z(t) \approx \exp\sum_{i=1}^{n} \frac{t^2\sigma_i^2}{2\sigma^2} = \exp\left(\frac{t^2}{2}\right) \tag{1.88}$$

which is the moment generating function of a normally distributed variable with mean 0 and variance 1; this can be verified by inspection of (1.50). Thus, approximately, $Z \sim N(0,1)$ and, since $Y = \mu + Z\sigma$ is a linear combination of Z, it follows that $Y \sim \left(\mu, \sigma^2\right)$ is approximately normal as well. (A little more formally, it is the c.d.f. of Z that converges to the c.d.f. of the $N(0,1)$, as $n \to \infty$. This is known as convergence in distribution).

As pointed out by Bulmer (1979), the central limit theorem explains why many observed distributions "look" normal. To the extent that random variation is the result of a large number of independent factors acting additively, each making a small contribution to the total variation, the resulting distribution should be close to a normal one.

There are extensions of the central limit theorem for variables that are independent but not identically distributed, and for dependent random variables. The latter is particularly relevant for the study of time series and Markov processes. A good starting point is the book of Lehmann (1999).

Example 1.15 *A tetravariate normal distribution*

Consider two genetically related individuals and suppose that a measurement, e.g., stature, is taken on each of them. Assume that the infinitesimal additive genetic model described above operates. The expected genetic value of a child, given the genetic values of the parents, is equal to the average of the genetic values of the father and mother. Also, assume that the population from which the two individuals are drawn randomly, is homogeneous in every possible respect. Let the model for the measurements be (we relax the distinction between random variables and their realized values unless the setting requires maintaining it)

$$\begin{aligned} y_1 &= a_1 + e_1, \\ y_2 &= a_2 + e_2. \end{aligned}$$

Further, suppose that $E(a_i) = E(e_i) = 0$ $(i = 1, 2)$, so that $E(y_i) = 0$, and that $Cov(a_i, e_i) = 0$, implying that $Var(y_i) = \sigma_a^2 + \sigma_e^2$, where σ_a^2 is the additive genetic variance and σ_e^2 is the residual or environmental variance. Also, let $Cov(e_1, e_2) = 0$; thus, $Cov(y_1, y_2) = Cov(a_1, a_2) = r_{12}\sigma_a^2$ where r_{12} is the additive genetic relationship between individuals 1 and 2. The additive relationship is equal to twice the probability that a randomly chosen allele drawn from a locus from individual 1 is identical by descent (i.e., it is a biochemical copy from a common ancestor) to a randomly chosen allele taken from the same locus of individual 2 (Malécot, 1969). For example, the average additive genetic relationship between two full-sibs is equal to 1/2, whereas that between an individual and itself, in the absence of inbreeding, is equal to 1. Under multivariate normality, if pairs of such individuals are drawn at random from the population, the joint distribution of measurements and of additive genetic values can be

represented as

$$\begin{bmatrix} y_1 \\ y_2 \\ a_1 \\ a_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2+\sigma_e^2 & r_{12}\sigma_a^2 & \sigma_a^2 & r_{12}\sigma_a^2 \\ r_{12}\sigma_a^2 & \sigma_a^2+\sigma_e^2 & r_{12}\sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & r_{12}\sigma_a^2 & \sigma_a^2 & r_{12}\sigma_a^2 \\ r_{12}\sigma_a^2 & \sigma_a^2 & r_{12}\sigma_a^2 & \sigma_a^2 \end{bmatrix} \right).$$

In what follows, use is made of results (1.77) through (1.83). First, the mean and variance of the conditional distribution $[y_1|a_1]$ is derived, with the dependence on the parameters suppressed in the notation. Because of the assumption of joint normality, this distribution must be normal as well. The mean and variance are given by

$$\begin{aligned} E\left(y_1|a_1\right) &= 0+\sigma_a^2\left(\sigma_a^2\right)^{-1}(a_1-0)=a_1, \\ Var\left(y_1|a_1\right) &= \sigma_a^2+\sigma_e^2-\sigma_a^2\left(\sigma_a^2\right)^{-1}\sigma_a^2=\sigma_e^2, \end{aligned}$$

and these are calculated with (1.82) and (1.83), respectively. The expected value and variance of this conditional distribution can also be deduced directly from the model, and also hold in the absence of normality, provided that the covariance between the additive genetic value and the residual deviation of the same individual is null. This can be verified simply by fixing the additive genetic value, and then taking the expectation and the variance under this assumption.

Consider now the conditional distribution $[y_1|a_1,a_2]$. Intuitively, it is clear that this must be the same as $[y_1|a_1]$ because, given the additive genetic value of individual 1, knowledge of the genetic value of individual 2 should not provide any additional information about the stature of individual 1. Having fixed a_1, the only remaining term in the model is the residual e_1, and this is independent of a_2. Anyhow, the distribution $[y_1|a_1,a_2]$ must be normal, and application of (1.82) and (1.83) yields

$$E\left(y_1|a_1,a_2\right)=\begin{bmatrix} \sigma_a^2 & r_{12}\sigma_a^2 \end{bmatrix}\begin{bmatrix} \sigma_a^2 & r_{12}\sigma_a^2 \\ r_{12}\sigma_a^2 & \sigma_a^2 \end{bmatrix}^{-1}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}=a_1$$

and

$$\begin{aligned} Var\left(y_1|a_1,a_2\right) &= \sigma_a^2+\sigma_e^2 \\ -\begin{bmatrix} \sigma_a^2 & r_{12}\sigma_a^2 \end{bmatrix}&\begin{bmatrix} \sigma_a^2 & r_{12}\sigma_a^2 \\ r_{12}\sigma_a^2 & \sigma_a^2 \end{bmatrix}^{-1}\begin{bmatrix} \sigma_a^2 \\ r\sigma_a^2 \end{bmatrix}=\sigma_e^2. \end{aligned}$$

Because the mean and variance are sufficient to identify fully the desired normal distribution, it follows that $[y_1|a_1,a_2]=[y_1|a_1]$, as expected. Further, given the genotypic values, the observations are conditionally independent. Thus, the joint density of the two measurements, given the two genetic values, can be written as

$$\begin{aligned} p\left(y_1,y_2|a_1,a_2\right) &= p\left(y_1|a_1,a_2\right)p\left(y_2|a_1,a_2\right) \\ &= p\left(y_1|a_1\right)p\left(y_2|a_2\right). \end{aligned}$$

Exploiting situations of conditional independence is a key issue in the Bayesian analysis of hierarchical models. This is discussed in Chapter 6. We turn attention now to the distribution $[y_1|a_2]$. Again, this conditional probability distribution is normal, with mean

$$E\left(y_1|a_2\right) = 0 + r_{12}\sigma_a^2\left(\sigma_a^2\right)^{-1} a_2 = r_{12}a_2$$

and variance:

$$\begin{aligned} Var\left(y_1|a_2\right) &= \sigma_a^2 + \sigma_e^2 - r_{12}\sigma_a^2\left(\sigma_a^2\right)^{-1} r_{12}\sigma_a^2 \\ &= \sigma_a^2\left(1 - r_{12}^2\right) + \sigma_e^2. \end{aligned}$$

The variance of this distribution is smaller than that of the marginal distribution of y_1, but is larger than the variance of $[y_1|a_1]$; this is because r_{12} is larger than 0, although it cannot exceed 1. Conversely, the distribution $[a_2|y_1]$ has mean and variance

$$E\left(a_2|y_1\right) = \frac{r_{12}\sigma_a^2}{\sigma_a^2 + \sigma_e^2}y_1,$$

$$Var\left(a_2|y_1\right) = \sigma_a^2\left(1 - r_{12}^2\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}\right) = \sigma_a^2\left(1 - r_{12}^2h^2\right),$$

where $h^2 = \sigma_a^2/\left(\sigma_a^2 + \sigma_e^2\right)$ is called "heritability" in quantitative genetics. This parameter measures the fraction by which measurable differences between parents for a given trait are expected to be recovered in the next generation, under the supposition of additive Mendelian inheritance. The preceding mean and variance would be the parameters of the normal distribution used to infer the unobservable genetic value of individual 2 using a measurement (y_1) on related individual 1. Letting $r_{12} = 1$, one obtains the formulas needed to make inferences about the genetic value of animal 1 using his measurement or phenotypic value (remember that in genetics, "phenotype" means whatever can be observed in an individual, e.g., a blood group, or the testosterone level of a bull at a specified age).

Another conditional distribution of interest might be $[y_1|y_2]$. This process can be useful in a situation where, for example, one wishes to make probability statements about the phenotype of individual 1 based on a measurement obtained on relative 2. This distribution is normal with mean

$$\begin{aligned} E\left(y_1|y_2\right) &= 0 + r_{12}\sigma_a^2\left(\sigma_a^2 + \sigma_e^2\right)^{-1} y_2 \\ &= r_{12}h^2y_2 \end{aligned}$$

and variance

$$\begin{aligned} Var\left(y_1|y_2\right) &= \sigma_a^2 + \sigma_e^2 - r_{12}\sigma_a^2\left(\sigma_a^2 + \sigma_e^2\right)^{-1} r_{12}\sigma_a^2 \\ &= \left(\sigma_a^2 + \sigma_e^2\right)\left(1 - r_{12}^2h^4\right). \end{aligned}$$

It is seen that even for individuals that are as closely related as full-sibs ($r_{12} = 1/2$), and for heritability values as large as 1/2, knowledge of y_2 produces a reduction in variance of only 1/16, so not much knowledge about y_1 is gained in this situation. ■

Example 1.16 *Decomposing the joint distribution of additive genetic values*

Consider the following genealogy (also called pedigree in animal breeding):

Individual	Father	Mother
1	–	–
2	–	–
3	1	2
4	1	2
5	1	3
6	4	3
7	5	6

Let a_i $(i = 1, 2, \ldots, 7)$ be the additive genetic value of individual i. Assume a model of gene transmission that allows writing the regression of the additive genetic value of a child (a_o) on the additive genetic values of its father (a_f) and of its mother (a_m) as

$$a_o = \frac{1}{2}a_f + \frac{1}{2}a_m + e_o,$$

where e_o, often known as the Mendelian sampling term, is distributed independently of similar terms in any other ancestors. The joint distribution of the additive genetic values in the pedigree can be factorized as follows:

$$\begin{aligned} p\left(a_1, a_2, \ldots, a_7\right) &= p\left(a_7 | a_1, a_2 \ldots, a_6\right) p\left(a_1, a_2 \ldots, a_6\right) \\ &= p\left(a_7 | a_5, a_6\right) p\left(a_1, a_2 \ldots, a_6\right). \end{aligned}$$

The equality in the second line follows, because under the Mendelian inheritance model, given the additive genetic values of its parents, the additive genetic value a_7 is conditionally independent of the additive genetic values of non-descendants of 7. Similarly $p\left(a_1, a_2 \ldots, a_6\right)$ can be factorized as

$$\begin{aligned} p\left(a_1, a_2 \ldots, a_6\right) &= p\left(a_6 | a_1, a_2, \ldots, a_5\right) p\left(a_1, a_2, \ldots, a_5\right) \\ &= p\left(a_6 | a_3, a_4\right) p\left(a_1, a_2, \ldots, a_5\right). \end{aligned}$$

Continuing in this way with the remaining terms, we obtain finally

$$\begin{aligned} p\left(a_1, a_2, \ldots, a_7\right) = p\left(a_7 | a_5, a_6\right) p\left(a_6 | a_3, a_4\right) p\left(a_5 | a_1, a_3\right) \\ p\left(a_4 | a_1, a_2\right) p\left(a_3 | a_1, a_2\right) p\left(a_1\right) p\left(a_2\right). \end{aligned}$$

This is an important decomposition that will be encountered again in (16.1) from Chapter 16 in the context of segregation analysis, where genetic values

are modeled as discrete random variables. Incidentally, note that offspring are conditionally independent, given their parents, and therefore

$$p\left(a_5, a_6 | a_1, a_3, a_4\right) = p\left(a_5 | a_1, a_3\right) p\left(a_6 | a_3, a_4\right).$$

However, given a_1, a_3, a_4, and a_7 (the child of 5 and 6), then a_5 and a_6 are no longer conditionally independent; they are correlated negatively. ■

Example 1.17 *A multivariate normal sampling model*

Imagine there is a data set consisting of two measurements (traits) taken on each of n subjects. For each trait, a model having the following form is adopted:

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{a}_j + \mathbf{e}_j, \quad j = 1, 2, \tag{1.89}$$

where $\boldsymbol{\beta}_j$ and $\mathbf{a}_j$ are, formally, location vectors containing the effects of factors affecting variation of the responses, and $\mathbf{X}_j$ and $\mathbf{Z}_j$ are known incidence matrices relating these parameters to the data vectors $\mathbf{y}_j$, each of order n. The term $\mathbf{e}_j$ is a residual representing random variation about $\mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_j$. For example, (1.89) could represent a mixed effects model for quantitative genetic analysis (Henderson, 1973) in which case $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_2$) and $\mathbf{a}_1$ ($\mathbf{a}_2$) would be fixed and random effects, respectively, on trait 1 (trait 2), respectively. The records on traits 1 and 2 for individual i can be put in a two-dimensional vector $\mathbf{y}_i^* = [y_{i1}, y_{i2}]'$, $(i = 1, 2, ..., n)$. These n vectors are assumed to be conditionally independent, given the location vectors, and assumed to follow a bivariate normal distribution. The joint density of all data $\mathbf{y}^* = [\mathbf{y}_1^{*\prime}, \mathbf{y}_2^{*\prime}, ..., \mathbf{y}_n^{*\prime}]'$ is then expressible as

$$\begin{aligned} &p\left(\mathbf{y}^* | \boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_0\right) \\ &\propto \left|\mathbf{R}_0\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(y_{i_1} - m_{i_1}, y_{i_2} - m_{i_2}\right)\mathbf{R}_0^{-1}\begin{pmatrix} y_{i_1} - m_{i_1} \\ y_{i_2} - m_{i_2}\end{pmatrix}\right], \end{aligned} \tag{1.90}$$

where $m_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \mathbf{z}_{i1}'\mathbf{a}_1$, $m_{i2} = \mathbf{x}_{i2}'\boldsymbol{\beta}_2 + \mathbf{z}_{i2}'\mathbf{a}_2$, and $\mathbf{x}_{i1}', \mathbf{x}_{i2}', \mathbf{z}_{i1}', \mathbf{z}_{i2}'$ are the ith rows of the incidence matrices $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}_1, \mathbf{Z}_2$, respectively. Here the dispersion structure will be taken to be

$$Var\begin{bmatrix} \mathbf{y}_{i1} | \boldsymbol{\beta}_1, \mathbf{a}_1 \\ \mathbf{y}_{i2} | \boldsymbol{\beta}_2, \mathbf{a}_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \mathbf{R}_0 \ \forall i, \tag{1.91}$$

so $\mathbf{R}_0 = \{r_{ij}\}$ is the variance–covariance matrix between the two traits measured on the same individual. Let

$$\mathbf{S}_e = \begin{bmatrix} \sum_{i=1}^{n}\left(y_{i1} - m_{i1}\right)^2 & \sum_{i=1}^{n}\left(y_{i1} - m_{i1}\right)\left(y_{i2} - m_{i2}\right) \\ \sum_{i=1}^{n}\left(y_{i1} - m_{i1}\right)\left(y_{i2} - m_{i2}\right) & \sum_{i=1}^{n}\left(y_{i2} - m_{i2}\right)^2 \end{bmatrix} \tag{1.92}$$

be a matrix of sums of squares and cross-products of residuals. Then

$$\begin{aligned}
&\sum_{i=1}^{n} \left[y_{i_1} - m_{i1}, y_{i_2} - m_{i2}\right] \mathbf{R}_0^{-1} \begin{bmatrix} y_{i_1} - m_{i1} \\ y_{i_2} - m_{i2} \end{bmatrix} \\
= \; & tr \left\{ \sum_{i=1}^{n} \left[y_{i1} - m_{i1}, y_{i2} - m_{i2}\right] \mathbf{R}_0^{-1} \begin{bmatrix} y_{i1} - m_{i1} \\ y_{i2} - m_{i2} \end{bmatrix} \right\} \\
= \; & tr \left(\mathbf{R}_0^{-1} \mathbf{S}_e\right),
\end{aligned}$$

where $tr\,(\cdot)$ means "trace" (sum of diagonal elements) of a matrix (Searle, 1982). Matrices can be commuted cyclically (preserving comformability) under the tr operator. With this notation, the joint probability density of all data is

$$p\left(\mathbf{y}^* | \boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_0\right) = (2\pi)^{-\frac{n}{2}} \left|\mathbf{R}_0\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{R}_0^{-1} \mathbf{S}_e\right)\right], \tag{1.93}$$

which is a useful representation in multivariate analysis (Anderson, 1984).■

Example 1.18 *Computing conditional multivariate normal distributions*
Conditional distributions are important in prediction of genetic values using mixed effects models (Henderson, 1963, 1973; Searle et al., 1992). These distributions also play a key role in a Gibbs sampling-based Bayesian analysis. The algorithm will be discussed in detail later, and at this point it suffices to state that its implementation requires constructing all possible conditional distributions from a joint distribution of interest. The example here will illustrate how some conditional distributions that arise in connection with a mixed effects linear model can be computed under Gaussian assumptions.
Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{1.94}$$

where the random vectors $\mathbf{a}$ and $\mathbf{e}$ have a joint multivariate normal distribution with null mean vector and covariance matrix

$$Var \begin{bmatrix} \mathbf{a} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_a^2 & 0 \\ 0 & \mathbf{I}\sigma_e^2 \end{bmatrix}.$$

Above, $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices, $\mathbf{A}$ is a known matrix (e.g., of additive relationships between individuals), and σ_a^2 and σ_e^2 are variance components. Then $E\left(\mathbf{y}|\boldsymbol{\beta}\right) = \mathbf{X}\boldsymbol{\beta}$ and $Var\left(\mathbf{y}|\boldsymbol{\beta}, \sigma_a^2, \sigma_e^2\right) = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{I}\sigma_e^2$. Let

$$\begin{aligned}
\boldsymbol{\theta}' &= \left[\boldsymbol{\beta}', \mathbf{a}'\right]', \quad \mathbf{W} = [\mathbf{X}, \mathbf{Z}], \\
\boldsymbol{\Sigma} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{bmatrix} \text{ and } \mathbf{C} = \mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma},
\end{aligned}$$

where $k = \sigma_e^2/\sigma_a^2$. The linear set of equations

$$\mathbf{C}\,\widehat{\boldsymbol{\theta}} = \mathbf{W}'\mathbf{y} = \mathbf{r}$$

has solution $\widehat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{r}$, assuming the inverse of the coefficient matrix $\mathbf{C}$ exists. This system is called Henderson's mixed model equations. It will be shown later on, that in a certain Bayesian setting (Lindley and Smith, 1972; Gianola and Fernando, 1986) the posterior distribution of $\boldsymbol{\theta}$ when σ_a^2 and σ_e^2 are known, is multivariate normal with parameters

$$\boldsymbol{\theta}|\sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}, \mathbf{C}^{-1}\sigma_e^2\right). \tag{1.95}$$

Now, partition $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$ arbitrarily, so $\boldsymbol{\theta}_1$ can be a scalar or a vector. What then is the distribution $\left[\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right]$? From multivariate normal theory, because the posterior distribution $\left[\boldsymbol{\theta}|\sigma_a^2, \sigma_e^2, \mathbf{y}\right]$ is normal with parameters (1.95), it follows that the desired conditional posterior distribution must be normal as well. A useful way of arriving at the parameters of this conditional distribution is presented here. Given the above partition, one can write the joint posterior distribution of $\boldsymbol{\theta}$ as

$$\left[\begin{array}{c} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{array}\middle|\, \sigma_a^2, \sigma_e^2, y\right] \sim N\left(\left[\begin{array}{c} \widehat{\boldsymbol{\theta}}_1 \\ \widehat{\boldsymbol{\theta}}_2 \end{array}\right], \left[\begin{array}{cc} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{array}\right]\sigma_e^2\right), \tag{1.96}$$

where

$$\left[\begin{array}{cc} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{array}\right]^{-1} = \left[\begin{array}{cc} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{array}\right] = \mathbf{C}^{-1}.$$

Now define $\mathbf{r}' = [\mathbf{r}_1', \mathbf{r}_2']$, such that the partition is consistent with that of $\boldsymbol{\theta}$. Using (1.82), the expected value of the distribution $\left[\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right]$ is

$$E\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right) = \widehat{\boldsymbol{\theta}}_1 + \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\left(\boldsymbol{\theta}_2 - \widehat{\boldsymbol{\theta}}_2\right). \tag{1.97}$$

From the mixed model equations, after inverting $\mathbf{C}$, one has

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_1 &= \mathbf{C}^{11}\mathbf{r}_1 + \mathbf{C}^{12}\mathbf{r}_2, \\ \widehat{\boldsymbol{\theta}}_2 &= \mathbf{C}^{21}\mathbf{r}_1 + \mathbf{C}^{22}\mathbf{r}_2. \end{aligned}$$

Employing these expressions for $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ in (1.97) above, we get:

$$\begin{aligned} E\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right) &= \mathbf{C}^{11}\mathbf{r}_1 + \mathbf{C}^{12}\mathbf{r}_2 + \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1} \\ &\quad \times \left(\boldsymbol{\theta}_2 - \mathbf{C}^{21}\mathbf{r}_1 - \mathbf{C}^{22}\mathbf{r}_2\right) \\ &= \left[\mathbf{C}^{11} - \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\mathbf{C}^{21}\right]\mathbf{r}_1 \\ &\quad + \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\boldsymbol{\theta}_2 \\ &= \mathbf{C}_{11}^{-1}\left[\mathbf{r}_1 + \mathbf{C}_{11}\mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\boldsymbol{\theta}_2\right] \\ &= \mathbf{C}_{11}^{-1}\left(\mathbf{r}_1 - \mathbf{C}_{12}\boldsymbol{\theta}_2\right). \end{aligned} \tag{1.98}$$

In the above derivation, use is made of the standard matrix algebra result for partitioned inverses (e.g., Searle, 1982):

$$\mathbf{C}^{11} - \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\mathbf{C}^{21} = \mathbf{C}_{11}^{-1}$$

and

$$\mathbf{C}_{11}\mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1} = -\mathbf{C}_{12}.$$

The variance of the conditional posterior distribution $\left[\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right]$ is

$$\begin{aligned} Var\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right) &= \left[\mathbf{C}^{11} - \mathbf{C}^{12}\left(\mathbf{C}^{22}\right)^{-1}\mathbf{C}^{21}\right]\sigma_e^2 \\ &= \mathbf{C}_{11}^{-1}\sigma_e^2. \end{aligned} \tag{1.99}$$

Therefore, one can write

$$\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left[\mathbf{C}_{11}^{-1}\left(\mathbf{r}_1 - \mathbf{C}_{12}\boldsymbol{\theta}_2\right), \mathbf{C}_{11}^{-1}\sigma_e^2\right]. \tag{1.100}$$

Results (1.98) and (1.99) are useful in the implementation of a Gibbs sampler in a hierarchical or mixed effects linear model. Even if $\boldsymbol{\theta}$ has a large number of elements and $\mathbf{C}$ is, therefore, a very large matrix (difficult to invert by brute force methods), the mean and variance of $\left[\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right]$ involve the inverse of a matrix having order equal to the number of elements in $\boldsymbol{\theta}_1$. For example, if $\boldsymbol{\theta}_1$ is a scalar, only the reciprocal of the appropriate scalar element is needed. ■

1.4.5 Quadratic Forms on Normal Variables: the Chi-square Distribution

Let the random vector $\mathbf{y}$ have the distribution

$$\mathbf{y} \sim N\left(\mathbf{m}, \mathbf{V}\right)$$

and consider the random variable $\mathbf{y}'\mathbf{Q}\mathbf{y}$. This is called a quadratic form on vector $\mathbf{y}$; the matrix of constants $\mathbf{Q}$ can be taken to be symmetric, without loss of generality. Then, provided that $\mathbf{QV}$ is idempotent one has that

$$\mathbf{y}'\mathbf{Q}\mathbf{y} \sim \chi^2\left(rank\left(\mathbf{Q}\right), \frac{1}{2}\mathbf{m}'\mathbf{Q}\mathbf{m}\right). \tag{1.101}$$

(Matrix $\mathbf{A}$ is idempotent if $\mathbf{AA} = \mathbf{A}$. If $\mathbf{A}$ is idempotent, $rank\left(\mathbf{A}\right) = tr\,\mathbf{A}$). Expression (1.101) means that the quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$ has a noncentral chi-square distribution with integer degrees of freedom equal to $rank(\mathbf{Q})$, and where $\frac{1}{2}\mathbf{m}'\mathbf{Q}\mathbf{m}$ is the noncentrality parameter (see Searle, 1971). Other authors (e.g., Stuart and Ord, 1987) define the noncentrality parameter as $\mathbf{m}'\mathbf{Q}\mathbf{m}$. If the non-centrality parameter is null, with a sufficient condition for this being $\mathbf{m} = \mathbf{0}$, then $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is said to have a central chi-square distribution.

The mean value of the distribution (1.101) is

$$E\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=\mathbf{m}'\mathbf{Q}\mathbf{m}+tr\left(\mathbf{Q}\mathbf{V}\right) \tag{1.102}$$

and the variance can be shown to be equal to

$$Var\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=4\mathbf{m}'\mathbf{Q}\mathbf{V}\mathbf{Q}\mathbf{m}+2tr\left(\mathbf{Q}\mathbf{V}\right)^{2}. \tag{1.103}$$

In the special case when $\mathbf{m}=\mathbf{0}$, then $E\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=tr\left(\mathbf{Q}\mathbf{V}\right)$ and

$$Var\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=2\,tr\left(\mathbf{Q}\mathbf{V}\right)^{2}=2\,tr\left(\mathbf{Q}\mathbf{V}\right).$$

Typically,

$$rank\left(\mathbf{Q}\right)<rank\left(\mathbf{V}\right),$$

so

$$tr\left(\mathbf{Q}\mathbf{V}\right)=rank\left(\mathbf{Q}\mathbf{V}\right)=rank\left(\mathbf{Q}\right),$$

this resulting from the idempotency of $\mathbf{QV}$. Then, for $\mathbf{m}=\mathbf{0}$,

$$E\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=rank\left(\mathbf{Q}\right)$$

and

$$Var\left(\mathbf{y}'\mathbf{Q}\mathbf{y}\right)=2\,rank\left(\mathbf{Q}\right),$$

so the mean and variance of the distribution are given by the number of degrees of freedom, and by twice the degrees of freedom, respectively.

Example 1.19 *Distribution of estimates of the variance of a normal distribution*

In Chapter 3, it will be shown that for the sampling model

$$\mathbf{y}\sim N\left(\mathbf{X}\boldsymbol{\beta},\mathbf{I}\sigma_{e}^{2}\right),$$

the maximum likelihood (ML) estimator of the variance σ_e^2 is given by

$$\begin{aligned}\widehat{\sigma_{e}^{2}} &= \frac{\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)}{n}\\ &= \frac{\mathbf{y}'\mathbf{Q}\mathbf{y}}{n},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ is the ordinary least-squares estimator of $\boldsymbol{\beta}$, $\mathbf{Q}=[\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}']$ is an idempotent $n\times n$ matrix, and n is the order of $\mathbf{y}$. Here it is assumed that the matrix $\mathbf{X}$ has full column rank equal to p, the order of $\boldsymbol{\beta}$. Now

$$\begin{aligned}\widehat{\sigma_{e}^{2}} &= \frac{\sigma_{e}^{2}}{n}\frac{\mathbf{y}'\mathbf{Q}\mathbf{y}}{\sigma_{e}^{2}}=\frac{\sigma_{e}^{2}}{n}\mathbf{y}'\mathbf{Q}\left(\frac{\mathbf{I}}{\sigma_{e}^{2}}\right)\mathbf{Q}\mathbf{y}\\ &= \frac{\sigma_{e}^{2}}{n}\mathbf{y}^{*\prime}\left(\frac{\mathbf{I}}{\sigma_{e}^{2}}\right)\mathbf{y}^{*}.\end{aligned}$$

Hence, $\widehat{\sigma_e^2}$ is a multiple of the quadratic form $\mathbf{y}^{*\prime}\left(\mathbf{I}\sigma_e^{-2}\right)\mathbf{y}^*$, where $\mathbf{y}^* = \mathbf{Q}\mathbf{y}$. It will be verified that this new quadratic form has a chi-square distribution. First note that $\mathbf{y}^*$ is normal by virtue of being a linear combination of $\mathbf{y}$, with

$$E\left(\mathbf{y}^*\right) = \mathbf{Q}\mathbf{E}\left(\mathbf{y}\right) = \mathbf{Q}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

and

$$Var\left(\mathbf{y}^*\right) = \mathbf{Q}\sigma_e^2.$$

Further

$$\left(\frac{\mathbf{I}}{\sigma_e^2}\right) Var\left(\mathbf{y}^*\right) = \left(\frac{\mathbf{I}}{\sigma_e^2}\right)\mathbf{Q}\sigma_e^2 = \mathbf{Q}$$

is idempotent. Hence,

$$\mathbf{y}^{*\prime}\left(\frac{\mathbf{I}}{\sigma_e^2}\right)\mathbf{y}^* = \frac{\mathbf{y}'\mathbf{Q}\mathbf{y}}{\sigma_e^2}$$

has a central chi-square distribution (since $E\left(\mathbf{y}^*\right) = \mathbf{0}$) with degrees of freedom equal to

$$\begin{aligned} rank\left(\frac{\mathbf{Q}}{\sigma_e^2}\right) &= rank\left(\mathbf{Q}\right) = tr\left(\mathbf{Q}\right) = tr\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \\ &= n - tr\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right] \\ &= \left[n - tr\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right] = n - p. \end{aligned}$$

It follows that

$$\begin{aligned} \widehat{\sigma_e^2} &= \frac{\sigma_e^2}{n}\mathbf{y}^{*\prime}\left(\frac{\mathbf{I}}{\sigma_e^2}\right)\mathbf{y}^* \\ &= \frac{\sigma_e^2}{n}\chi^2_{n-p}, \end{aligned}$$

so the ML estimator of σ_e^2 is distributed as a scaled chi-square random variable. ■

1.4.6 The Wishart and Inverse Wishart Distributions

Multivariate analysis is important in quantitative and evolutionary genetics. For example, in plant and animal breeding, selection for multiple attributes, e.g., yield and quality of wheat or growth rate and feed efficiency in beef cattle, is the rule rather than the exception. This requires knowledge of genetic variances and covariances between traits. Correlation and covariance matrices are also of interest in evolution because, for example, a genetic variance–covariance matrix contains information about possible

bottlenecks under natural selection (Lynch and Walsh, 1998), or about the evolutionary trajectory of several traits (Roff, 1997). The Wishart and inverse Wishart distributions appear in connection with inference about covariance matrices involving attributes that follow a multivariate normal distribution, and play an important role in multivariate analysis (Mardia et al., 1979; Anderson, 1984). Deriving these distributions is technically involved, so only a few features are sketched here.

Suppose there is data on p traits or attributes, each expressed as a deviation from their respective expectations, on each of n individuals. The p-dimensional vector of traits is assumed to follow the p-variate $N_p(\mathbf{0}, \mathbf{\Sigma})$ distribution, where $\mathbf{\Sigma}$ is a positive-definite variance–covariance matrix between attributes. Let $\mathbf{Y}_{(n\times p)}$ be a data matrix containing in row i, say, the p measurements taken on individual i. Often $\mathbf{Y}_{(n\times p)}$ can be assumed to represent a collection of n independent draws from this normal sampling model. Form now $\mathbf{M} = \mathbf{Y}'\mathbf{Y}$, a random $p \times p$ matrix of sums of squares and crossproducts. Given the normality assumption, the symmetric matrix $\mathbf{M}$ is said to have a Wishart distribution of order p with scale matrix $\mathbf{\Sigma}$ and degrees of freedom parameter equal to n. We write $\mathbf{M} \sim W_p(\mathbf{\Sigma}, n)$ to denote this random process. The Wishart distribution is a matrix generalization of the chi-squared distribution, as we shall see later, and it involves $p(p+1)/2$ random quantities: the p distinct sums of squares and the $p(p-1)/2$ sums of products.

The p.d.f. of the Wishart distribution is

$$p(\mathbf{M}|\mathbf{\Sigma}, n) = \frac{|\mathbf{M}|^{(n-p-1)/2} \exp\left[-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1}\mathbf{M}\right)\right]}{2^{np/2}\pi^{p(p-1)/4}|\mathbf{\Sigma}|^{n/2}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]}$$
$$\propto |\mathbf{M}|^{(n-p-1)/2} \exp\left[-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1}\mathbf{M}\right)\right] \tag{1.104}$$

with the sample space being such that $|\mathbf{M}| > 0$. A special situation is when $p = 1$, in which case $\mathbf{M}$ and $\mathbf{\Sigma}$ are scalars; here $M = \sum_{i=1}^{n} Y_i^2$ is a sum of squares, and $\Sigma = \sigma^2$ is the variance of the normal distribution. In this situation, the Wishart density reduces to the univariate p.d.f.

$$p\left(M|\sigma^2, n\right) \propto M^{(n-2)/2} \exp\left(-\frac{M}{2\sigma^2}\right). \tag{1.105}$$

This is the kernel of a gamma density with parameters $n/2$ and $(2\sigma^2)^{-1}$. It is also the density of the distribution of $\sigma^2\chi^2_{(n)}$, a scaled chi-square random variable, with scale parameter σ^2 and n degrees of freedom. Return to the general case, and put $\mathbf{M} = \{M_{ij}\}$ and $\mathbf{\Sigma} = \{\sigma_{ij}\}$. It can be shown that the expected (matrix) value of the random matrix $\mathbf{M}$ with p.d.f. given in (1.104) is $E(\mathbf{M}|\mathbf{\Sigma}, n) = n\mathbf{\Sigma}$ so, for any element of the matrix $E(M_{ij}) = n\sigma_{ij}$.

The Inverse Wishart Distribution

A related distribution is that of the inverse of a matrix of sums of squares and products of n randomly drawn vectors from $N_p(\mathbf{0}, \mathbf{\Sigma})$. The distribution of $\mathbf{T} = \mathbf{M}^{-1}$, for $|\mathbf{\Sigma}| > 0$ and $n \geq p$, is called the inverse Wishart distribution; it is symbolized as $\mathbf{T} \sim W_p^{-1}(\mathbf{\Sigma}, n)$ or as $\mathbf{T} \sim IW_p(\mathbf{\Sigma}, n)$. The density of an inverse Wishart matrix is

$$\begin{aligned} p(\mathbf{T}|\mathbf{\Sigma}, n) &= \frac{|\mathbf{T}|^{-(n+p+1)/2} \exp\left[-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1}\mathbf{T}^{-1}\right)\right]}{2^{np/2}\pi^{p(p-1)/4}|\mathbf{\Sigma}|^{n/2}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]} \\ &\propto |\mathbf{T}|^{-(n+p+1)/2} \exp\left[-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1}\mathbf{T}^{-1}\right)\right]. \end{aligned} \tag{1.106}$$

The expected value of $\mathbf{T}$ is

$$E(\mathbf{T}|\mathbf{\Sigma}, n) = \frac{\mathbf{\Sigma}^{-1}}{(n-p-1)}, \tag{1.107}$$

provided $n \geq p+2$. An important special case is when $\mathbf{T}$ is a scalar ($p = 1$). For example, in a Bayesian context one may consider using the scalar version of (1.106) to describe uncertainty about the variance σ^2. Here, $T = \sigma^2$, and $\mathbf{\Sigma}^{-1} = S$ is now the scale parameter. Then (1.106) reduces to:

$$p\left(\sigma^2|S, n\right) \propto \left(\sigma^2\right)^{-\left(\frac{n}{2}+1\right)} \exp\left(-\frac{S}{2\sigma^2}\right). \tag{1.108}$$

This is the kernel of the density of a scaled inverted chi-square distribution with parameters S and n (also, of an inverted gamma process with parameters $n/2$ and $S/2$). An alternative representation is that of, e.g., Gelman et al. (1995), who put $S = nS^*$. In a sense, S can be interpreted, Bayesianly speaking, as a prior sum of squares, whereas S^* would play the role of a value of σ^2 that, a priori, is viewed as very likely (actually, the most likely one when n is very large). In a Bayesian setting, n can be interpreted as a "degree of belief" in S^*.

More generally, if the random matrix is now $\mathbf{T} = \mathbf{\Sigma}$, the variance–covariance matrix of a normal distribution, then (1.106) will describe uncertainty about it. The scale parameter would be interpretable as a matrix of "prior sums of squares and products". If one has a prior opinion about the value of an unknown variance–covariance matrix $\mathbf{\Sigma}$, and this is $\mathbf{\Sigma}^*$, say, the value of the scale matrix (which we denote now as $\mathbf{S}^{-1}$ to avoid confusion with the covariance matrix that one wishes to infer, this being $\mathbf{\Sigma}$) can be assessed from (1.107) as $\mathbf{S}^{-1} = (n-p-1)\mathbf{\Sigma}^*$, given a value of n.

Properties of Wishart and Inverse Wishart Distributions

Some properties of Wishart and inverse Wishart distributions are summarized below. The material is taken from Korsgaard et al. (1999), where

the results presented here are given in a more general setting. Let $\mathbf{M}$ be a 2×2 symmetric, positive-definite random matrix having the Wishart distribution

$$\mathbf{M}|\mathbf{V}, n \sim W_2(\mathbf{V}, n),$$

where $\mathbf{V}$ is the scale matrix. Let $\mathbf{T} = \mathbf{M}^{-1}$ so that

$$\mathbf{T}|\mathbf{V}, n \sim IW_2(\mathbf{V}, n).$$

The symmetric matrices $\mathbf{M}, \mathbf{V}, \mathbf{T}$ have elements

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$

and

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix},$$

respectively. Using properties of partitioned matrices one can write

$$\begin{aligned} M_{11} &= \left(T_{11} - T_{12}^2 T_{22}^{-1}\right)^{-1}, \\ M_{12} &= -T_{12} / \left(T_{11} T_{22} - T_{12}^2\right), \\ M_{22} &= \left(T_{22} - T_{12}^2 T_{11}^{-1}\right)^{-1}. \end{aligned}$$

Define

$$\begin{aligned} X_1 &= M_{11}, \\ X_2 &= M_{11}^{-1} M_{12}, \\ X_3 &= M_{22} - M_{12}^2 M_{11}^{-1}. \end{aligned}$$

From these equalities it can be deduced that the following one-to-one relationships exist between the Xs and the Ts:

$$T_{11} = X_1^{-1} + X_2^2 X_3^{-1}, \tag{1.109}$$
$$T_{12} = -X_2 X_3^{-1}, \tag{1.110}$$
$$T_{22} = X_3^{-1}. \tag{1.111}$$

Then the following properties can be shown to hold:

$$X_1 \sim W_1(V_{11}, n), \tag{1.112}$$
$$(X_2|X_1 = x_1) \sim N\left(V_{11}^{-1} V_{12}, x_1^{-1} V_{22.1}\right), \tag{1.113}$$
$$X_3 \sim W_1(V_{22.1}, n-1), \tag{1.114}$$
$$p(x_1, x_2|x_3) = p(x_1, x_2), \tag{1.115}$$

where $V_{22.1} = V_{22} - V_{12}^2 V_{11}^{-1}$. This means that X_1 and X_3 have univariate Wishart (or gamma) distributions with appropriate parameters, that the conditional distribution $[X_2|X_1 = x_1]$ is normal, and the joint distribution $[X_1, X_2]$ is independent of X_3. All these distributions are easy to sample from. Thus, if one wishes to draw a random matrix from an $W_2(\mathbf{V}, n)$ distribution, the sampling procedure consists of:
(a) draw x from the three preceding distributions;
(b) compute $\mathbf{T}^*$ from x. This is a realized value from $\mathbf{T}|\mathbf{V}$, $n \sim IW_2(\mathbf{V}, n)$;
(c) finally, invert $\mathbf{T}^*$ to obtain a draw from $W_2(\mathbf{V}, n)$.

In a joint analysis of Gaussian and discrete data (the latter employing a probit model) with Bayesian methods, the following problem is often encountered. Draws of 2×2 covariance matrices are to be obtained from a certain posterior distribution, subject to the restriction that one of the variances is equal to 1 (this is the residual variance in the probit scale). In Chapter 14 we show how to exploit the properties described above, in order to perform a Markov chain Monte Carlo (MCMC) based Bayesian analysis of Gaussian and binary distributed traits.

Simulation of an Inverse Wishart Distribution

An efficient way of simulating a p-dimensional Wishart distribution with n degrees of freedom and scale matrix $\mathbf{S}$, $W_p(\mathbf{S}, n)$**,** is described by Odell and Feiveson (1966). The algorithm is as follows:

- Compute the Cholesky factorization of $\mathbf{S} = \mathbf{L}'\mathbf{L}$, such that $\mathbf{L}'$ is lower triangular.
- Construct a lower triangular matrix

$$\mathbf{T} = \{t_{ij}\},\ i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, p,$$

  with $t_{ii} = \sqrt{\chi^2_{n+1-i}}$, $t_{ij} \sim N(0, 1)$, if $i > j$, and $t_{ij} = 0$ if $i < j$.
- Compute the product $\mathbf{L}'\mathbf{T}\mathbf{T}'\mathbf{L}$. This matrix is distributed as $W_p(\mathbf{S}, n)$.
- The matrix $\left(\mathbf{L}'\mathbf{T}\mathbf{T}'\mathbf{L}\right)^{-1}$ is distributed as $IW_p(\mathbf{S}, n)$.

As discussed later in this book, a Bayesian analysis requires posing a prior distribution for all the parameters of the model. If a covariance matrix $\mathbf{C}$ of dimension $p \times p$, say, is assigned a priori the distribution $IW_p(\mathbf{V}, n)$, one way of choosing the scale matrix $\mathbf{V}$ is from consideration of the expected value of $\mathbf{C}$:

$$E(\mathbf{C}|\mathbf{V}, n) = \frac{\mathbf{V}^{-1}}{n - p - 1}.$$

Then set $\mathbf{V}^{-1} = (n - p - 1)\,\widetilde{E}(\mathbf{C}|\mathbf{V}, n)$, where $\widetilde{E}(\mathbf{C}|\mathbf{V}, n)$ is some "reasonable" value chosen on the basis of prior information.

1.4.7 The Multivariate-t Distribution

Density of the Distribution

Suppose a random vector has the conditional multivariate normal distribution with p.d.f.

$$\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w \sim N\left(\mathbf{y}|\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{w}\right), \tag{1.116}$$

where w, in turn, is a scalar random variable following a

$$Ga\left(\frac{\nu}{2},\frac{\nu}{2}\right)$$

process; here $\nu > 0$ is a parameter. The density of the joint distribution of $\mathbf{y}$ and w, using (1.40) and (1.77), is then

$$\begin{aligned} p\left(\mathbf{y},w|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right) &= p\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w\right)p\left(w|\nu\right) \\ &= \left|2\pi\left(\frac{\boldsymbol{\Sigma}}{w}\right)\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}\right)'\left(\frac{\boldsymbol{\Sigma}}{w}\right)^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)\right] \\ &\quad\times\frac{\left(\nu/2\right)^{\frac{\nu}{2}}}{\Gamma\left(\nu/2\right)}w^{\frac{\nu}{2}-1}\exp\left[-\frac{\nu w}{2}\right]. \end{aligned} \tag{1.117}$$

The marginal density of $\mathbf{y}$ is obtained by integrating the joint density above with respect to w, yielding

$$\begin{aligned} p\left(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right) &= \left|2\pi\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\frac{\left(\nu/2\right)^{\frac{\nu}{2}}}{\Gamma\left(\nu/2\right)} \\ \times\int_0^\infty w^{\frac{n+\nu}{2}-1}\exp\left[-w\frac{\left(\mathbf{y}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)+\nu}{2}\right]dw. \end{aligned} \tag{1.118}$$

Reference to (1.40) indicates that the integrand is the kernel of the density

$$Ga\left(w|\frac{n+\nu}{2},\frac{\left(\mathbf{y}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)+\nu}{2}\right).$$

Hence, the integral in (1.118) is equal to the reciprocal of the integration constant of the corresponding distribution, that is

$$\begin{aligned} &\int_0^\infty w^{\frac{n+\nu}{2}-1}\exp\left[-w\frac{\left(\mathbf{y}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)+\nu}{2}\right]dw \\ &\qquad=\frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\left[\frac{\left(\mathbf{y}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)+\nu}{2}\right]^{\frac{n+\nu}{2}}}. \end{aligned} \tag{1.119}$$

Employing (1.119) in (1.118), and rearranging

$$p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{(\nu)^{\frac{\nu}{2}}\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)|\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu\right]^{-\frac{n+\nu}{2}}$$
$$= \frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)|\nu\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[1+\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{n+\nu}{2}}. \quad (1.120)$$

This is the density of an n-dimensional multivariate-t distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degrees of freedom parameter ν. Note that ν can take any value in the positive part of the real line and does not need to be an integer. When $n = 1$, a univariate-t distribution results, and the density has already been given in (1.51).

It is interesting to observe that the t distribution results by averaging an infinite number of normal processes with a randomly varying covariance matrix $\boldsymbol{\Sigma}w^{-1}$ over a gamma (or inverse gamma, or scaled inverse chi-square) distribution. For this reason, it is often stated in the literature that the t distribution is a mixture of an infinite number of Gaussian processes (Gelman et al., 1995).

The mean vector and covariance matrix of the multivariate-t distribution can be deduced from (1.116) using iterated expectations (see below), as this leads to

$$E(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = E_w(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w) = E_w(\boldsymbol{\mu}) = \boldsymbol{\mu}$$

and to

$$Var(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = E_w[Var(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w)] + Var_w[E(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w)]$$
$$= E_w[Var(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w)] = E_w\left(\frac{\boldsymbol{\Sigma}}{w}\right) = \boldsymbol{\Sigma}E_w\left(\frac{1}{w}\right).$$

It is shown in Chapter 2 that the average value of the reciprocal of a gamma random variable is given by

$$E_w\left(\frac{1}{w}\right) = \frac{a}{b-1} = \frac{\frac{\nu}{2}}{\frac{\nu}{2}-1} = \frac{\nu}{\nu-2}$$

in this case. Thus

$$Var(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}.$$

Marginal and Conditional Distributions

A similar development can be adopted to show that all marginal and conditional distributions deriving from a multivariate-t distribution are univariate or multivariate-t as well. This is so since for any arbitrary partition of $\mathbf{y}$ in (1.116), say,

$$\left[\begin{array}{c}\mathbf{y}_1\\ \mathbf{y}_2\end{array}\right]\Bigg|\,\boldsymbol{\mu},\boldsymbol{\Sigma},w \sim N\left(\left[\begin{array}{c}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{array}\right],\left[\begin{array}{cc}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{array}\right]\frac{1}{w}\right),$$

all marginal and conditional distributions are normal, with appropriate mean vector and covariance matrix. Integration over the $Ga(\nu/2, \nu/2)$ distribution leads to the desired result directly. For example, the conditional distribution of $\mathbf{y}_1$ given $\mathbf{y}_2$ is an n_1-dimensional (the order of $\mathbf{y}_1$) multivariate-t with mean vector

$$E(\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22})^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2),$$

covariance matrix

$$Var(\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\nu}{\nu - 2}\left[\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22})^{-1}\boldsymbol{\Sigma}_{21}\right],$$

and degrees of freedom ν. A similar reasoning leads to the result that any linear combination of a vector that has a multivariate-t distribution must be multivariate- (or univariate-) t distributed also.

1.5 Distributions with Constrained Sample Space

In genetics, situations where the sampling space of random variables or where the space of unknowns is restricted, are not uncommon. For example, a Poisson sampling model might be sensible for describing litter size in pigs. However, litters of size 0 may not be reported in practice! In this situation, one may consider adopting a Poisson distribution with probabilities "normalized" such that the event $x = 0$ is not observable. Likewise, many "selection" models have been proposed in the context of the multivariate normal distribution and these often involve a restriction of the sampling space of the variables entering into the problem. Perhaps the best known one is selection by truncation (Pearson, 1903; Aitken, 1934). Henderson et al. (1959), Curnow (1961), Thompson (1973), Henderson (1975), Thompson (1976), Robertson (1977), Bulmer (1980), and Gianola et al. (1989), among others, have discussed the effects of selection of multivariate normal variates in a genetic context, from the point of view of either parameter estimation (fixed effects; variance components) or of prediction of breeding values.

Let X be a discrete random variable with p.m.f. $p(x)$ and let a and b be constants lying within the support of the domain of p. Then, the doubly truncated p.m.f. of X, given that $a < X \leq b$, is

$$\begin{aligned}\Pr(X = x|a < X \leq b) &= \frac{\Pr(X = x, a < X \leq b)}{\Pr(a < X \leq b)} \\ &= \frac{\Pr(X = x)}{\Pr(a < X \leq b)}, \quad \text{for } a < X \leq b.\end{aligned}$$

Therefore,

$$\Pr(X = x|a < X \leq b) = \begin{cases} 0, & \text{if } x \leq a \text{ or } x > b, \\ \frac{p(x)}{F(b) - F(a)}, & \text{if } a < x \leq b, \end{cases}$$

where $F(\cdot)$ is the distribution function. In particular, given a truncation point t, the p.m.f. of X, given that $X > t$ (i.e., truncated below), is

$$\Pr(X = x|X > t) = \begin{cases} 0, & \text{if } x \leq t, \\ \frac{p(x)}{1-F(t)}, & \text{if } x > t. \end{cases} \tag{1.121}$$

Example 1.20 *Sibship size*

As an example of a discrete truncated distribution, let s denote the number of children in a nuclear family. Assuming that a male or a female birth are equally likely (i.e., $\theta = 1/2$), the probability that there will be exactly x girls in a family of size s, assuming a binomial sampling model, is

$$\begin{aligned} \Pr(X = x|\theta, s) &= \binom{s}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{s-x} \\ &= \binom{s}{x} \left(\frac{1}{2}\right)^s, \quad x = 0, 1, \ldots, s. \end{aligned}$$

Suppose that there is interest in the distribution of the number of girls in nuclear families of size s that have at least one girl. The corresponding probability, using (1.121), is

$$\begin{aligned} \Pr(x|X > 0, \theta, s) &= \frac{\Pr(X = x|\theta, s)}{1 - \Pr(X = 0|\theta, s)} \\ &= \frac{\binom{s}{x} \left(\frac{1}{2}\right)^s}{1 - \left(\frac{1}{2}\right)^s}, \quad x = 1, 2, \ldots, s. \end{aligned}$$

For example, for $s = 4$, the unconstrained and truncated probability distributions are

x	0	1	2	3	4
$p(x\|\theta, s)$	0.0625	0.2500	0.3750	0.2500	0.0625
$p(x\|X > 0, \theta, s)$	0	0.2667	0.4000	0.2667	0.0667

The two distributions do not differ by much because $\Pr(X = 0|\theta, s)$ takes a low value in the untruncated binomial distribution. ■

Let a random vector $\mathbf{z}$ have a multivariate distribution with density $p(\mathbf{z}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. If the original sampling space of $\mathbf{z}$, R_z is constrained such that this vector is observed only in a more limited space, R_z^c, and this happens with probability

$$\int p(\mathbf{z}|\boldsymbol{\theta})\, I(\mathbf{z} \in R_z^c)\, d\mathbf{z} = P$$

where $I(\cdot)$ is an indicator function, then the density of the constrained distribution of $\mathbf{z}$ is $p(\mathbf{z}|\boldsymbol{\theta})/P$, which can be seen to integrate to 1 over

the space R_z^c. This setting has been used extensively by Pearson (1903) and Henderson (1975), among others, in genetic applications. A special case is the truncation selection model of quantitative genetics, which is described now for a bivariate situation. Let two random variables X and Y have a bivariate distribution with joint p.d.f. $p(x, y)$. Assume that selection operates on X such that this random variable is observed only if $X \geq t$, where t is a known truncation point, or threshold of selection, whereas the sampling space of Y remains unrestricted. The density of the joint distribution of X and Y (after selection), using the result given above, is

$$p(x, y|X \geq t) = \frac{p(x, y)}{\int_{-\infty}^{\infty} \int_{t}^{\infty} p(x, y)\, dx\, dy} = \frac{p(x, y)}{\int_{t}^{\infty} p(x)\, dx} = \frac{p(x, y)}{P}, \qquad (1.122)$$

where P is now the proportion selected. The conditional density of Y given X (after selection) is, by definition

$$p(y|x, X \geq t) = \frac{p(y, x|X \geq t)}{p(x|X \geq t)} = \frac{p(y, x)/P}{p(x)/P} = p(y|x). \qquad (1.123)$$

It follows that this conditional distribution is the same as in the absence of selection. This result is not unexpected because, intuitively, the conditional distribution $[Y|X]$ is defined for any X, so whatever happens with the sample space of X is irrelevant. An important corollary is that the form of $E(Y|X)$, the regression function, is unaffected by selection operating on X, this being true for any distribution. However, the marginal distribution of Y is altered by selection, unless X and Y are independent. This is illustrated below.

Example 1.21 *Parental selection*

Suppose there are two unrelated parents and one offspring; let their additive genetic values be a_1, a_2 and a_3 respectively. Under multivariate normality the conditional distribution of the additive genetic value of a_3 given a_1 and a_2 is normal with mean $E(a_3|a_1, a_2) = (a_1 + a_2)/2$ and variance $\sigma_a^2/2$; this variance can be found using formula (1.83), taking into account that the covariance between the additive genetic values of a parent and of an offspring is $\sigma_a^2/2$. This conditional distribution holds true for any pair of parents, selected or unselected, so the expected value of the additive genetic value of an offspring, conditionally on parental breeding values, is always equal to the average of the additive genetic values of parents. However, the unconditional mean and variance of a_3 are affected by the selection process. Letting E_s denote expected value under selection, it follows, by virtue of the theorem of double expectation (to be discussed below), that

$$E_s(a_3) = E_s[E(a_3|a_1, a_2)] = E_s\left[\frac{a_1 + a_2}{2}\right] \neq E\left[\frac{a_1 + a_2}{2}\right],$$

so the mean of the additive genetic values in the offspring, following selection, in general, will not be equal to the mean breeding value of parents selected at random, unless selection is of a stabilizing form (Bulmer, 1980). Likewise, the variance of breeding values in progeny of selected parents, Var_s, is

$$\begin{aligned} Var_s\left(a_3\right) &= E_s\left[Var_s\left(a_3|a_1,a_2\right)\right]+Var_s\left[E_s\left(a_3|a_1,a_2\right)\right] \\ &= \frac{\sigma_a^2}{2}+Var_s\left[\frac{a_1+a_2}{2}\right] \\ &= \frac{1}{2}\left[\sigma_a^2+\sigma_{a_s}^2\right], \end{aligned}$$

where $\sigma_{a_s}^2$ is the variance of the breeding values within selected parents. Most often, the latter will not be equal to the variance in the absence of selection. It can be lower or larger depending on the form of selection. ■

In a Bayesian analysis, inferences about a parameter vector $\boldsymbol{\theta}$ are made from the conditional distribution $[\boldsymbol{\theta}|\mathbf{y}]$, called posterior distribution in this specific context. Parameters are treated in the Bayesian approach as random variables (the randomness arises from the state of uncertainty about their values) and the analysis is made conditionally on the observed data, $\mathbf{y}$. Suppose $\mathbf{y}_0$ and $\mathbf{y}_1$ are vectors of data of a selection experiment collected at generations 0 and 1, respectively. Assume that individuals with data $\mathbf{y}_1$ are the offspring of selected parents, and that selection was based on the available phenotypic records, $\mathbf{y}_0$. Further, suppose that selection is by truncation, as before, and that the truncation point t is known. Had there been no selection, the density of the posterior distribution (using all data) is

$$p\left(\boldsymbol{\theta}|\mathbf{y}_0,\mathbf{y}_1\right)=\frac{p\left(\boldsymbol{\theta},\mathbf{y}_0,\mathbf{y}_1\right)}{p\left(\mathbf{y}_0,\mathbf{y}_1\right)}. \tag{1.124}$$

However, if the joint distribution $[\mathbf{y}_0,\mathbf{y}_1]$ is modified by selection such that only individuals whose phenotypic records exceed t (this is informally denoted as $\mathbf{y}_0>t$) are used as parents, the posterior distribution must be written as

$$\begin{aligned} p\left(\boldsymbol{\theta}|\mathbf{y}_0,\mathbf{y}_1,\mathbf{y}_0>t\right) &= \frac{p\left(\boldsymbol{\theta},\mathbf{y}_0,\mathbf{y}_1|\mathbf{y}_0>t\right)}{p\left(\mathbf{y}_0,\mathbf{y}_1|\mathbf{y}_0>t\right)} \\ &= \frac{p\left(\boldsymbol{\theta},\mathbf{y}_0,\mathbf{y}_1\right)}{P}\frac{P}{p\left(\mathbf{y}_0,\mathbf{y}_1\right)} \\ &= p\left(\boldsymbol{\theta}|\mathbf{y}_0,\mathbf{y}_1\right), \end{aligned} \tag{1.125}$$

where

$$P=\int_t^\infty\int_{-\infty}^\infty p\left(\mathbf{y}_0,\mathbf{y}_1\right)d\mathbf{y}_1\,d\mathbf{y}_0$$

is the probability of selection. The important point is that when all data on which selection operates is included in the analysis, inferences about

$\boldsymbol{\theta}$ can be made from the posterior density $p(\boldsymbol{\theta}|\mathbf{y}_0, \mathbf{y}_1)$, as if selection had not taken place. Including all data (records from parents, nonparents and offspring) in the analysis is necessary, to justify using the sampling model $[\mathbf{y}_0, \mathbf{y}_1|\boldsymbol{\theta}]$ as if selection had not occurred. This sampling model is needed for constructing the posterior density (1.125). Gianola and Fernando (1986), and more recently Sorensen et al. (2001), discuss this result in a more general setting.

Simulation of Univariate Truncated Distributions

An efficient algorithm for sampling from truncated distributions can be found in Devroye (1986) and is as follows. Let Y be a random variable from a normal distribution, truncated between a (lower bound) and b (upper bound). To sample from the truncated normal $TN_{(a,b)}\left(\mu, \sigma^2\right)$, where μ and σ^2 are the mean and variance before truncation:

- Simulate U from a uniform distribution $Un(0,1)$.
- The truncated normal is

$$Y = \mu + \sigma\Phi^{-1}\left[p_1 + U\left(p_2 - p_1\right)\right],$$

where $\Phi^{-1}(\cdot)$ is the inverse c.d.f. of the normal distribution,

$$p_1 = \Phi\left[\left(a - \mu\right)/\sigma\right]$$

and

$$p_2 = \Phi\left[\left(b - \mu\right)/\sigma\right].$$

The method can be generalized to any univariate distribution truncated in the interval $[a, b]$. If the c.d.f. of the untruncated variate is F, then a draw from the truncated distribution is

$$y = F^{-1}\left\{F(a) + U\left[F(b) - F(a)\right]\right\}. \tag{1.126}$$

The proof that (1.126) is a value from the desired truncated distribution in the interval $[a, b]$ is as follows

$$\begin{aligned}
\Pr\left(Y \leq y\right) &= \Pr\left[F^{-1}\left\{F(a) + U\left[F(b) - F(a)\right]\right\} \leq y\right] \\
&= \Pr\left[F(a) + U\left[F(b) - F(a)\right] \leq F(y)\right] \\
&= \Pr\left[U \leq \frac{F(y) - F(a)}{F(b) - F(a)}\right] \\
&= \int_0^{\frac{F(y)-F(a)}{F(b)-F(a)}} du \\
&= \frac{F(y) - F(a)}{F(b) - F(a)}.
\end{aligned}$$

1.6 Iterated Expectations

Let $\mathbf{x}$ and $\mathbf{y}$ be two random vectors with joint p.d.f. $p(\mathbf{x},\mathbf{y})$ and let $p(\mathbf{x}|\mathbf{y})$ denote the conditional p.d.f. of $\mathbf{x}$ given $\mathbf{y}$ (the distinction between a random variable and its realized value is dropped in this section). Then

$$E(\mathbf{x}) = E_y\left[E(\mathbf{x}|\mathbf{y})\right]. \tag{1.127}$$

This also holds for an arbitrary function of $\mathbf{x}$, $f(\mathbf{x})$, in which case $\mathbf{x}$ is replaced above by $f(\mathbf{x})$. The proof of (1.127) is as follows:

$$\begin{aligned} E(\mathbf{x}) &= \int \mathbf{x}p(\mathbf{x})\,d\mathbf{x} = \int\int \mathbf{x}p(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y} \\ &= \int\left[\int \mathbf{x}p(\mathbf{x}|\mathbf{y})\,d\mathbf{x}\right]p(\mathbf{y})\,d\mathbf{y} \\ &= \int\left[E(\mathbf{x}|\mathbf{y})\right]p(\mathbf{y})\,d\mathbf{y} = E_y\left[E(\mathbf{x}|\mathbf{y})\right], \end{aligned}$$

where $p(\mathbf{y})$ is the marginal density of $\mathbf{y}$. Thus the mean of $\mathbf{x}$ can be obtained by averaging conditional (given $\mathbf{y}$) means over the marginal distribution of $\mathbf{y}$. This result is at the root of Monte Carlo methods that use conditional distributions. For example, one can estimate $E(\mathbf{x})$ by either direct drawings from the distribution of $\mathbf{x}$ or, alternatively, by drawing samples from the distribution of $\mathbf{y}$ and computing $E(\mathbf{x}|\mathbf{y})$ for each sample. Then weighted averages of these conditional expectations are calculated, where the weights are assigned according to the density that the values of $\mathbf{y}$ take in their distribution. As shown below, since $Var(\mathbf{x}) \geq Var_y\left[E(\mathbf{x}|\mathbf{y})\right]$, inferences using the conditional mean rather than direct drawings from the distribution of $\mathbf{x}$ are usually more precise.

A similar result holds for the covariance

$$Cov(\mathbf{x},\mathbf{y}') = E_z\left[Cov(\mathbf{x},\mathbf{y}'|\mathbf{z})\right] + Cov_z\left[E(\mathbf{x}|\mathbf{z}),E(\mathbf{y}'|\mathbf{z})\right], \tag{1.128}$$

where $Cov(\cdot)$ now indicates a covariance matrix. Using the above result, observe that by definition of a covariance matrix (Searle, 1971),

$$\begin{aligned} E_z\left[Cov(\mathbf{x},\mathbf{y}'|\mathbf{z})\right] &= E_z\left\{E(\mathbf{x}\mathbf{y}'|\mathbf{z}) - \left[E(\mathbf{x}|\mathbf{z})\right]\left[E(\mathbf{y}'|\mathbf{z})\right]\right\} \\ &= E_z\left[E(\mathbf{x}\mathbf{y}'|\mathbf{z})\right] - E_z\left\{\left[E(\mathbf{x}|\mathbf{z})\right]\left[E(\mathbf{y}'|\mathbf{z})\right]\right\} \\ &= E(\mathbf{x}\mathbf{y}') - \left[E(\mathbf{x})E(\mathbf{y}')\right] - E_z\left\{\left[E(\mathbf{x}|\mathbf{z})\right]\left[E(\mathbf{y}'|\mathbf{z})\right]\right\} \\ &\quad + E_z\left[E(\mathbf{x}|\mathbf{z})\right]E_z\left[E(\mathbf{y}'|\mathbf{z})\right] \\ &= Cov(\mathbf{x},\mathbf{y}') - Cov_z\left[E(\mathbf{x}|\mathbf{y}),E(\mathbf{y}'|\mathbf{z})\right] \end{aligned}$$

and this leads to the desired result directly. The expression for the variance is obtained immediately, setting $\mathbf{x} = \mathbf{y}$,

$$Var(\mathbf{x}) = E_y\left[Var(\mathbf{x}|\mathbf{y})\right] + Var_y\left[E(\mathbf{x}|\mathbf{y})\right]. \tag{1.129}$$

Example 1.22 *Predicting a random variable using the conditional mean*
Let Y be a random variable to be predicted using some function $u(X)$ involving the known random variable $X = x$. For any function $u(X)$, consider $E[(Y - u(X))^2]$ equal to

$$E\{[Y - u(X)]^2\} = \int\int [y - u(x)]^2 p(x, y)\, dxdy. \tag{1.130}$$

Expression (1.130) is minimized when $u(x) = E(Y|X = x)$. The proof is as follows. Let $E(Y|X = x) = w(x)$ and write (1.130) as

$$\begin{aligned} E\{[Y - u(X)]^2\} &= E\left\{[(Y - w(X)) + (w(X) - u(X))]^2\right\} \\ &= E\left\{[Y - w(X)]^2\right\} + E\left\{[w(X) - u(X)]^2\right\} \\ &\quad +2E\left\{[Y - w(X)][w(X) - u(X)]\right\}. \end{aligned}$$

The expectation involving the cross-product term can be written as

$$\begin{aligned} &E\left\{[Y - w(X)][w(X) - u(X)]\right\} \\ &= E_X\left\{E_Y\left[(Y - w(X))(w(X) - u(X))|X\right]\right\} \\ &= E_X\left\{(w(X) - u(X))E_Y\left[(Y - w(X))|X\right]\right\}. \end{aligned}$$

However

$$E_Y\left[(Y - w(X))|X\right] = E(Y|X) - w(X) = 0.$$

Then it follows that

$$E\{[Y - u(X)]^2\} = E\left\{[Y - w(X)]^2\right\} + E\left\{[w(X) - u(X)]^2\right\}.$$

The first term in the right-hand side does not involve $u(X)$, and

$$E\left\{[w(X) - u(X)]^2\right\} \geq 0$$

with equality if $u(X) \equiv w(X) = E(Y|x)$. Defining the "best predictor" as that for which $E[(Y - u(X))^2]$ is a minimum, then the best choice for $u(x)$ is

$$u(x) = E(Y|X = x).$$

■

Example 1.23 *Variation in gene frequencies: the beta–binomial distribution*
Return to Example 1.8, where clusters of n alleles are drawn at random, and where $\theta = \Pr(\text{allele } A)$. Suppose now that θ varies between clusters according to a beta distribution, with density

$$p\left(\theta|\overline{\theta}, c\right) = \frac{\Gamma(c)}{\Gamma\left(c\overline{\theta}\right)\Gamma\left[c\left(1-\overline{\theta}\right)\right]}\theta^{c\overline{\theta}-1}(1-\theta)^{c\left(1-\overline{\theta}\right)-1}.$$

Recall that, in the parameterization of Wright (1968), $c = a + b$ and

$$\overline{\theta} = \frac{a}{a+b}.$$

The two alternative parameterizations will be used interchangeably here. Let X be the number of A alleles in a cluster of size n and suppose that, given θ, its distribution is binomial. Then, $X = \sum_{i=1}^{n} X_i$, where, given θ, X_i is a Bernoulli variable with success probability θ. Hence

$$E(X|\theta) = \sum_{i=1}^{n} E(X_i) = n\theta$$

and

$$Var(X|\theta) = \sum_{i=1}^{n} Var(X_i) = n\theta(1-\theta),$$

since the Bernoulli variables are assumed to be conditionally independent. Now, using (1.127), the mean value of X over clusters is

$$E(X) = E[E(X|\theta)] = E(n\theta) = n\overline{\theta} = \frac{na}{a+b}. \tag{1.131}$$

Employing (1.129),

$$\begin{aligned} Var(X) &= E[Var(X|\theta)] + Var[E(X|\theta)] \\ &= E[n\theta(1-\theta)] + Var(n\theta) \\ &= n\left[E(\theta) - E^2(\theta) + n\,Var(\theta)\right]. \end{aligned}$$

Using the mean and variance of the beta distribution, given in (1.37) and (1.38), respectively, one obtains after algebra,

$$\begin{aligned} Var(X) &= \frac{nab(a+b+n)}{(a+b)^2(a+b+1)} \\ &= n\overline{\theta}\left(1-\overline{\theta}\right)\left(\frac{c+n}{c+1}\right). \end{aligned} \tag{1.132}$$

If the cluster has size $n = 1$, the mean of the marginal distribution of X is $\overline{\theta}$, and the variance is $\overline{\theta}\left(1-\overline{\theta}\right)$.
Next, we derive the correlation between alleles within a cluster. Note that for a cluster of size n

$$\begin{aligned} Var(X) &= Var\left(\sum_{i=1}^{n} X_i\right) = n\,Var(X_i) + n(n-1)\,Cov(X_i, X_j) \\ &= n\,Var(X_i)[1 + (n-1)\rho] \\ &= n\overline{\theta}\left(1-\overline{\theta}\right)[1 + (n-1)\rho], \end{aligned}$$

where ρ is the correlation between alleles; if $\rho = 0$, there is no within-cluster aggregation and sampling is purely binomial. The correlation between alleles is

$$\rho = \frac{1}{n-1}\left[\frac{Var\,(X)}{n\overline{\theta}\left(1-\overline{\theta}\right)} - 1\right].$$

The marginal distribution of X is called beta–binomial, which is discrete. Its form can be derived by noting that if $X|\theta$ is binomial, and θ follows a beta process, then the marginal distribution of X is given by

$$\begin{aligned}
\Pr\left(X = x|n, c, \overline{\theta}\right) &= \int_0^1 \left(\begin{array}{c} n \\ x \end{array}\right) \theta^x \left(1-\theta\right)^{n-x} \\
&\times \frac{\Gamma\left(c\right)}{\Gamma\left(c\overline{\theta}\right)\Gamma\left[c\left(1-\overline{\theta}\right)\right]} \theta^{c\overline{\theta}-1}\left(1-\theta\right)^{c\left(1-\overline{\theta}\right)-1} d\theta \\
&= \left(\begin{array}{c} n \\ x \end{array}\right) \frac{\Gamma\left(c\right)}{\Gamma\left(c\overline{\theta}\right)\Gamma\left[c\left(1-\overline{\theta}\right)\right]} \\
&\times \int_0^1 \theta^{x+c\overline{\theta}-1}\left(1-\theta\right)^{n-x+c\left(1-\overline{\theta}\right)-1} d\theta \\
&= \left(\begin{array}{c} n \\ x \end{array}\right) \frac{\Gamma\left(c\right)}{\Gamma\left(c\overline{\theta}\right)\Gamma\left[c\left(1-\overline{\theta}\right)\right]} \\
&\times \frac{\Gamma\left(x+c\overline{\theta}\right)\Gamma\left[n-x+c\left(1-\overline{\theta}\right)\right]}{\Gamma\left(n+c\right)}.
\end{aligned} \tag{1.133}$$

The last expression results from use of the integral in (1.36). Below is another example of the beta–binomial process. ■

Example 1.24 *Deconditioning a binomial distribution having a random parameter*

As discussed before, the sum of n independent random variables X_i, each following a Bernoulli probability distribution $Br\,(\theta)$, is a random variable that has a binomial distribution. Assume now that the probability of success, θ, is unknown, and that this random variable is assigned a beta distribution $Be\,(a, b)$. Thus, conditionally on θ

$$\Pr\left(X_i = x_i|\theta\right) = \begin{cases} Br\left(x_i|\theta\right) = \theta^{x_i}\left(1-\theta\right)^{1-x_i}, & \text{for } x_i = 0, 1, \\ 0, \quad \text{for other values of } x_i. \end{cases} \tag{1.134}$$

The density of the beta distribution for θ is

$$p\left(\theta|a, b\right) = Be\left(\theta|a, b\right) = \begin{cases} C\theta^{a-1}\left(1-\theta\right)^{b-1}, & \text{for } 0 \le \theta_i \le 1, \\ 0, \quad \text{for other values of } \theta, \end{cases} \tag{1.135}$$

where a and b are the parameters of the beta distribution. Recall that the mean and variance of the beta distribution are given, respectively, by

$$\begin{aligned} E(\theta|a,b) &= \frac{a}{a+b}, \\ Var(\theta|a,b) &= \frac{ab}{\left[(a+b)^2(a+b+1)\right]}. \end{aligned}$$

Consider the random variable $Y = \sum_{i=1}^{n} X_i$, the total number of "successes" in n trials. Given θ, the random variable Y is $Bi(\theta, n)$, and its marginal distribution is in the form of a beta–binomial, which is generated by the mixture

$$Bb(y|a,b,n) = \int_0^1 Bi(y|n,\theta)\, Be(\theta|a,b)\, d\theta, \tag{1.136}$$

as in the previous example. Any of the moments of $[Y|a,b,n]$ can be obtained from (1.136). In animal breeding, the beta–binomial model could arise in the following context. Suppose that in a given cluster (a cow in a given herd, say), a cow is artificially inseminated and pregnancy is registered. Pregnancy can be modeled as a Bernoulli random variable, with unknown probability equal to θ associated with the particular cluster. Here we obtain the mean and variance of Y, the rate of calving of the cow in n trials, unconditionally on θ, using iterated expectations (in a more realistic set up, there would be several clusters, each with its own probability of pregnancy). The mean of the marginal distribution of Y is obtained directly as follows:

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_i E_\theta\left[E(X_i|\theta)\right] \\ &= \sum_i E_\theta(\theta) \\ &= \frac{na}{a+b}. \end{aligned}$$

Because, given θ, the X_i are independent, the conditional variance is

$$\begin{aligned} Var(Y|\theta) &= Var\left(\sum_{i=1}^{n}(X_i|\theta)\right) \\ &= \sum_{i=1}^{n} Var(X_i|\theta) = n\theta(1-\theta). \end{aligned}$$

Recalling (1.129), the marginal variance of Y is

$$Var\left(Y\right) = E_\theta\left[Var\left(Y|\theta\right)\right] + Var_\theta\left[E\left(Y|\theta\right)\right].$$

Since $E\left(Y|\theta\right) = n\theta$, and θ has a beta distribution,

$$Var_\theta\left[E\left(Y|\theta\right)\right] = Var\left(n\theta\right) = \frac{n^2 ab}{\left(a+b\right)^2\left(a+b+1\right)}.$$

Also, $Var\left(Y|\theta\right) = n\theta\left(1-\theta\right)$. Then

$$\begin{aligned} E_\theta\left[Var\left(Y|\theta\right)\right] &= E_\theta\left[n\theta\left(1-\theta\right)\right] \\ &= n\int_0^1 \theta\left(1-\theta\right) p\left(\theta|a,b\right) d\theta \\ &= Cn\int_0^1 \theta\left(1-\theta\right)\theta^{a-1}\left(1-\theta\right)^{b-1} d\theta \\ &= \frac{nab}{\left(a+b\right)\left(a+b+1\right)}. \end{aligned}$$

Thus

$$\begin{aligned} Var\left(Y\right) &= \frac{nab}{\left(a+b\right)^2}\frac{a+b+n}{a+b+1} \\ &= nE\left(\theta|a,b\right)\left[1-E\left(\theta|a,b\right)\right]\frac{a+b+n}{a+b+1}. \end{aligned}$$

The variance of the beta–binomial with mean probability $a/\left(a+b\right)$ is greater by a factor $\left(a+b+n\right)/\left(a+b+1\right)$ than the binomial with the same probability. When $n = 1$, there is no information available to distinguish between the beta and binomial variation, and both models have equal variances. ■

Example 1.25 *Genetic markers and the covariance between half-sibs*
Suppose a male is drawn at random from a population in equilibrium (i.e., where mating is at random, so gene and genotypic frequencies are constant from generation to generation, and there is no inbreeding or assortative mating). Let the additive genetic variance of a trait of interest at an autosomal additive locus be V_g. Since allelic effects would be independently and identically distributed in this population, V_g is equal to twice the variance between either paternal or maternal allelic effects. Assume that this male is randomly mated to an unknown female, also sampled at random, and that two half-sibs are born from this mating. Let x and y designate the values of the alleles (haplotypes) that the two half-sibs received from their

father and assume that $E(x) = E(y) = 0$. The allelic variance is

$$Var(x) = Var(y) = \frac{1}{2}V_g.$$

The additive genetic covariance between half-sibs, $Cov(x, y)$, can be derived as follows. Let z be a Bernoulli random variable taking the value 1 if the paternally derived alleles from both individuals are identical by descent (IBD), and 0 otherwise. Using (1.128),

$$\begin{aligned} Cov(x, y) &= E_z[Cov(x, y|z)] + Cov_z[E(x|z), E(y|z)] \\ &= E_z[Cov(x, y|z)]. \end{aligned}$$

This is so, because $E(x|z) = E(x) = 0$, a constant, with the same holding for y. In other words, the mean value of an allelic effect is not altered by knowledge of identity by descent. Taking expectations with respect to the distribution of z yields

$$\begin{aligned} &Cov(x, y) = E_z[Cov(x, y|z)] \\ &= Cov(x, y|z = 0)\Pr(z = 0) + Cov(x, y|z = 1)\Pr(z = 1). \end{aligned} \tag{1.137}$$

The term $Cov(x, y|z = 0)$ is null. This is because, if the two alleles are not IBD, the allelic effects are independently distributed; thus

$$Cov(x, y|z = 0) = E(xy|z = 0) = E(x)E(y) = 0.$$

Further

$$Cov(x, y|z = 1) = Var(x) = \frac{1}{2}V_g, \tag{1.138}$$

because if the two alleles are identical by descent, $x \equiv y$. In order to calculate the probability of IBD, suppose the father has genotype A_1A_2, where subscripts 1 and 2 are labels for chromosomes 1 and 2, respectively. Thus, in principle, alleles at position A_1A_2 can be identical in state. We refer to the event "drawing allele A_i $(i = 1, 2)$ from the first half-sib with value x" as "A_i in x". Similarly, the event, "drawing allele A_i from the other half-sib with value y" is written as "A_i in y". The required probability is

$$\begin{aligned} \Pr(z = 1) &= \omega_{11}\Pr(A_1 \text{ in } x \cap A_1 \text{ in } y) \\ &\quad + \omega_{22}\Pr(A_2 \text{ in } x \cap A_2 \text{ in } y) \\ &\quad + \omega_{12}[\Pr(A_1 \text{ in } x \cap A_2 \text{ in } y) + \Pr(A_2 \text{ in } x \cap A_1 \text{ in } y)] \\ &= \omega_{11}\frac{1}{2}\frac{1}{2} + \omega_{22}\frac{1}{2}\frac{1}{2} + 2\omega_{12}\frac{1}{2}\frac{1}{2}, \end{aligned} \tag{1.139}$$

where $\omega_{ij} = \Pr(A_i \equiv A_j | A_i \text{ in } x \cap A_j \text{ in } y)$, $i, j = 1, 2$, which we write $\Pr(A_i \equiv A_j)$ for short. Above, the four terms of the form

$$\Pr(A_i \text{ in } x \cap A_j \text{ in } y), \quad i = 1, 2,\ j = 1, 2,$$

are equal to

$$\begin{aligned}\Pr\left(A_i \text{ in } x \cap A_j \text{ in } y\right) &= \Pr\left(A_i \text{ in } x\right)\Pr\left(A_j \text{ in } y\right) \\ &= \frac{1}{2}\frac{1}{2} = \frac{1}{4},\end{aligned}$$

since there is a probability of 1/2 of drawing one of the two alleles from each individual and the two drawings are independent. Expression (1.139) results from the fact that the probability involves four mutually exclusive and exhaustive events. Two such events pertain to the situation where the same allele is picked on each of the individuals, and here

$$\Pr\left(A_1 \equiv A_1\right) = \Pr\left(A_2 \equiv A_2\right) = 1.$$

The other two events involve different alleles, but it must be noted that $\Pr\left(A_1 \equiv A_2\right)$ may not always be zero, as these alleles might be copies of the same allele from a common ancestor. Since we assume there is no inbreeding, $\Pr\left(A_1 \equiv A_2\right) = 0$, so

$$\Pr\left(z = 1\right) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}. \tag{1.140}$$

Using (1.138) and (1.140) in (1.137), the half-sib covariance is

$$Cov\left(x, y\right) = \frac{1}{4}V_g. \tag{1.141}$$

Imagine now that the sire is known to be heterozygote for a genetic marker linked to the locus affecting the trait in question. Let the recombination fraction between the marker and the locus be r, and define a new random variable M, taking the value 1 if the marker alleles are the same in both half-sibs, and 0 otherwise. The additive genetic covariance between the two half-sibs, conditionally on their marker information, is expressible as

$$\begin{aligned}Cov\left(x, y|M\right) &= E_z\left[Cov\left(x, y|z, M\right)\right] \\ &+ Cov_z\left[E\left(x|z, M\right), E\left(y|z, M\right)\right] \\ &= E_z\left[Cov\left(x, y|z, M\right)\right],\end{aligned}$$

since the covariance between the conditional means is zero, following the same argument as before. Now take expectations with respect to the distribution of z. When the half-sibs receive the same marker allele from the sire ($M = 1$), one has

$$\begin{aligned}&E_z\left[Cov\left(x, y|z, M = 1\right)\right] \\ &= Cov\left(x, y|z = 0, M = 1\right)\Pr\left(z = 0|M = 1\right) \\ &+ Cov\left(x, y|z = 1, M = 1\right)\Pr\left(z = 1|M = 1\right) \\ &= Cov\left(x, y|z = 1, M = 1\right)\Pr\left(z = 1|M = 1\right).\end{aligned} \tag{1.142}$$

The conditional probability in the bottom line of (1.142) is

$$\Pr(z=1|M=1)=\frac{\Pr(z=1,M=1)}{\Pr(M=1)}. \tag{1.143}$$

The numerator in (1.143) is the probability of drawing independently two gametes from the sire that have the same marker allele and the same allele at the locus in question. These gametes are either nonrecombinant, with probability $\frac{1}{2}(1-r)^2$, or recombinant, with the corresponding probability being $\frac{1}{2}r^2$. Therefore the numerator of (1.143) is equal to

$$\frac{1}{2}(1-r)^2+\frac{1}{2}r^2.$$

Since marker alleles in the two half-sibs are equal to each other one-half of the time, $\Pr(M=1)=\frac{1}{2}$. Therefore

$$\Pr(z=1|M=1)=(1-r)^2+r^2.$$

The conditional covariance between the half-sibs, given that they inherited the same marker allele from their sire, is

$$\begin{aligned} Cov(x,y|M=1) &= Cov(x,y|z=1,M=1)\left[(1-r)^2+r^2\right] \\ &= \frac{(1-r)^2+r^2}{2}V_g. \end{aligned} \tag{1.144}$$

This is because, if the two alleles are identical by descent,

$$Cov(x,y|z=1,M=1)=\frac{V_g}{2}$$

in (1.142). Similar arguments lead to the following expression for the conditional covariance between the half-sibs, given that they inherited different marker alleles from their sire

$$Cov(x,y|M=0)=\frac{r(1-r)}{2}V_g. \tag{1.145}$$

As $r\to 0$, (1.144) and (1.145) tend to $Vg/2$ and to 0, respectively. On the other hand, when the marker provides less and less information about the locus in question (i.e., when $r\to 1/2$), both (1.144) and (1.145) tend to $Vg/4$, as in (1.141), as it should. ■

In this chapter, we have discussed and illustrated the most important univariate and multivariate distributions encountered in statistical genetics. This is extended in Chapter 2, which discusses random processes arising from functions of random variables.

This page intentionally left blank

2

Uncertainty about Functions of Random Variables

2.1 Introduction

It is seldom the case that the initial parameterization of a statistical model for genetic analysis will lead directly to inferences about all parameters of interest. One may also wish to learn about functions of the parameters of the model. An illustration is the use of a linear mixed model parameterized in terms of variance components. The investigator may wish to make inferences about variance ratios or functions thereof, such as intraclass correlations or heritabilities. As another example, consider data from a trial in which experimental mice are subjected to varying doses of a carcinogenic agent. The response variable is whether a tumor has developed or not at the end of the trial. Because of the binary nature of the response, a Bernoulli sampling model may be adopted, and a linear structure (with dose as an explanatory variable) may be imposed to the log of the ratio between the probability of developing a tumor at a given dose and that of the complementary event. This is a "generalized linear model" with a logit link function (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Then, for mouse j at dose k, one could have

$$\log \frac{p_{jk}}{1-p_{jk}} = \beta_0 + \beta_1 x_k, \tag{2.1}$$

where p_{jk} is the probability of response, x_k is the dose or a function thereof (such as the logarithm of the dose), and β_0 and β_1 are parameters of the dose-response process. It may be that the latter are not the parameters of primary interest, for example, one may wish to find the dose at which

the probability of developing a tumor is $\frac{1}{2}$ (this is usually called the LD-50 dose, with LD standing for "lethal dose"). At this probability, it is seen that

$$x_{\text{LD-50}} = -\frac{\beta_0}{\beta_1}. \tag{2.2}$$

In a model where the parameters are regarded as random and, therefore, having a joint distribution, it would follow that the LD-50 is also a random variable, by virtue of being a function of randomly varying quantities. Even if the parameters are not "random", in the usual sense of being drawn randomly from a conceptual population of values (as in a random effects model), one could perhaps argue from a Bayesian perspective, as follows. Since the LD-50 is unknown, it is an uncertain quantity and, therefore, there is randomness which is naturally measured by probability. Here the LD-50 would be viewed as random irrespective of whether or not the β's are drawn randomly from a population! Disregarding the issue of the origin of the randomness, model (2.1) would require invoking a bivariate distribution for β_0 and β_1, whereas the LD-50 in (2.2) involves a scalar distribution resulting from a nonlinear function of β_0 and β_1.

In this chapter, we review some concepts of distribution theory needed for functions of random variables, and illustrate the theory with examples. Single random variables are considered in the first section. Subsequently, functions of sets of random variables are discussed.

2.2 Functions of a Single Random Variable

2.2.1 Discrete Random Variables

If X is a discrete random variable having some p.m.f. $p(x)$, then any function of X, say $Y = f(X)$, is also a random variable. If the inverse transformation from Y to X is denoted by $f^{-1}(Y) = X$, then the p.m.f. of the random variable Y is

$$p_Y(y) = p_X\left(f^{-1}(y)\right). \tag{2.3}$$

Example 2.1 *A situation involving the binomial distribution*
Suppose that X has a binomial distribution, $Bi(p, n)$. Consider the random variable $Y = f(X) = n - X$. The inverse transformation is $f^{-1}(Y) = X = n - Y$. The probability function of Y is then

$$\begin{aligned} p_Y(y) &= p_X\left(f^{-1}(y)\right) = p_X(n-y) \\ &= \begin{pmatrix} n \\ n-y \end{pmatrix} p^{n-y}(1-p)^{n-(n-y)} \\ &= \begin{pmatrix} n \\ y \end{pmatrix} (1-p)^y p^{n-y}. \end{aligned}$$

Therefore, Y has also a binomial distribution, but now with parameters $(1-p)$ and n. The sample space of Y is the same as that of X, that is, $Y = 0, 1, ..., n$. ■

2.2.2 Continuous Random Variables

Let X be a continuous random variable having sample space A (denoted as $X \in A$) and p.d.f. $p(x)$ (so $p(x) = 0$ for $x \notin A$). Let Y be a function of X, $Y = f(X)$. Then, if $f(\cdot)$ is a monotone function and the inverse transformation is $X = f^{-1}(Y)$, the p.d.f. of Y is given by

$$\begin{aligned} p_Y(y) &= p_X\left(f^{-1}(y)\right)\left|\frac{d}{dy}f^{-1}(y)\right| \\ &= p_X\left(f^{-1}(y)\right)|J(y)|, \quad y \in f(A), \end{aligned} \tag{2.4}$$

and $p_Y(y) = 0$ for $y \notin f(A)$. In (2.4), $|J(y)|$ is the absolute value of the Jacobian of the transformation as a function of y. This ensures that the density is positive throughout (the derivative of the inverse function with respect to y may be negative at some values).

Interpretation of expression (2.4) can be facilitated recalling (1.21) which discloses that a p.d.f. has units: probability by unit of measurement of the random variable. A change in these units leads naturally to a change in the density function.

An equivalent, perhaps more suggestive way of writing (2.4) is

$$p_Y(y) = p_X(x)\left|\frac{dx}{dy}\right|, \quad y \in f(A), \; x = f^{-1}(y). \tag{2.5}$$

Result (2.4) is not intuitively obvious, and its proof is as follows (e.g. Hoel et al., 1971). Let F and G denote the respective c.d.f.s of X and Y. Suppose first that $Y = f(X)$ is strictly increasing, i.e., $f(x_1) < f(x_2)$ if $x_1 < x_2$, with $x_1 \in A$ and $x_2 \in A$. Then f^{-1} is strictly increasing on $f(A)$, for $y \in f(A)$. One can write

$$\begin{aligned} G(y) &= \Pr(Y \leq y) \\ &= \Pr(f(X) \leq y) \\ &= \Pr\left(X \leq f^{-1}(y)\right) \\ &= F\left(f^{-1}(y)\right). \end{aligned}$$

Using the chain rule of differentiation yields

$$\begin{aligned} \frac{d}{dy}G(y) &= \frac{d}{dy}F\left(f^{-1}(y)\right) \\ &= \frac{d}{df^{-1}(y)}F\left(f^{-1}(y)\right)\frac{df^{-1}(y)}{dy} \\ &= p\left(f^{-1}(y)\right)\frac{df^{-1}(y)}{dy}. \end{aligned}$$

Now

$$\frac{df^{-1}(y)}{dy} = \left|\frac{df^{-1}(y)}{dy}\right|$$

since f^{-1} is strictly increasing, so this yields (2.4). Suppose next that f is strictly decreasing on A. Then f^{-1} is strictly decreasing on $f(A)$, and for $y \in f(A)$ the following holds:

$$\begin{aligned} G(y) &= \Pr(Y \leq y) \\ &= \Pr(f(X) \leq y) \\ &= \Pr\left(X \geq f^{-1}(y)\right) \\ &= 1 - F\left(f^{-1}(y)\right). \end{aligned}$$

Thus

$$\begin{aligned} \frac{d}{dy}G(y) &= -\frac{d}{dy}F\left(f^{-1}(y)\right) \\ &= \frac{d}{df^{-1}(y)}F\left(f^{-1}(y)\right)\left(-\frac{df^{-1}(y)}{dy}\right) \\ &= p\left(f^{-1}(y)\right)\left(-\frac{df^{-1}(y)}{dy}\right). \end{aligned}$$

Here

$$-\frac{df^{-1}(y)}{dy} = \left|\frac{df^{-1}(y)}{dy}\right|$$

because f^{-1} is strictly decreasing, thus yielding (2.4) again. Therefore, in either case, we see that the density of Y is given by (2.4).

Example 2.2 *The lognormal distribution*
A variable whose logarithm is normally distributed is said to have a lognormal distribution. Let $X \sim N\left(m, \sigma^2\right)$, for $-\infty < x < \infty$, and suppose one seeks the distribution of the transformed random variable $Y = f(X) = \exp(X)$. The inverse transformation is $X = f^{-1}(Y) = \ln(Y)$. The p.d.f. of Y is then, using (2.4),

$$\begin{aligned} p(y) &= p\left(f^{-1}(y)\right)\left|\frac{d}{dy}f^{-1}(y)\right| \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(\ln y - m)^2\right], \end{aligned}$$

where $1/y$ enters from the Jacobian of the transformation. Then, $Y = \exp(X)$ is said to have a lognormal distribution, with density as given above. This distribution arises, for example, in quantitative genetic analysis of the productive life of breeding animals (Ducrocq et al., 1988) and in survival analysis (Kleinbaum, 1996). ■

Example 2.3 *Distribution of the inverse of a lognormal random variable*
Consider now the transformation $Z = 1/Y$ where Y is lognormal, with density as given above. The absolute value of the Jacobian of the transformation is z^{-2} and, employing this in conjunction with the density above gives

$$\begin{aligned} p(z) &= \frac{1}{z^{-1}\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(-\ln z - m)^2\right] z^{-2} \\ &= \frac{1}{z\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\ln z + m)^2\right]. \end{aligned}$$

This implies that the reciprocal of a lognormal random variable is lognormal as well. ■

Example 2.4 *Transforming a uniform random variable*
Let X be a random variable having a uniform distribution in the interval $[0, 1]$, so its p.d.f. is

$$p(x|0,1) = \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

The uniform distribution was used by Bayes (1763) in an attempt to represent prior ignorance about the value of a parameter within a certain range. This "principle of insufficient reason" has been used extensively in quantitative genetics and in other fields. The uniform distribution assigns equal probabilities to all possible ranges of equal length within which the random variable can fall. Thus, it is intuitively appealing to use the uniform process to represent (in terms of probability) lack of prior knowledge. Often though, the uniform distribution is not a good choice for conveying vague prior knowledge. For example, one may think that a uniform distribution between -1 and 1 can be used to represent prior ignorance about a coefficient of correlation. However, Bayarri (1981) showed that the prior distribution that should be used for such purpose is not the uniform.

Suppose that one is interested in the variable $Y = f(X) = -\log X$. The inverse transformation is $X = f^{-1}(Y) = \exp(-Y)$. The interval $[0, 1]$, constituting the sample space of X, maps onto the interval $[0, \infty)$ as the sample space for Y. The absolute value of the Jacobian of the transformation is then

$$\left|\frac{d}{dy} f^{-1}(-y)\right| = \left|\frac{d}{dy}\exp(-y)\right| = \exp(-y).$$

The p.d.f. of Y is, therefore,

$$p(y) = p\left(f^{-1}(y)\right)\exp(-y) = \exp(-y), \quad \text{for } y > 0. \tag{2.6}$$

This is the density of an exponentially distributed random variable with mean and variance equal to 1. More generally, as noted in Chapter 1, the

density of an exponential distribution with parameter η is

$$p(y|\eta) = \eta \exp(-\eta y), \quad \text{for } y > 0$$

and the mean and variance can be shown to be η^{-1} and η^{-2}, respectively. From the developments leading to (2.6), it follows that if one wishes to simulate random variables from an exponential distribution with parameter η the draws can be computed as

$$y = -\frac{\ln(x)}{\eta},$$

where x is a draw from $Un(0,1)$. ■

Example 2.5 *From the beta to the logistic distribution*
Assume that X follows the beta distribution, $Be(a,b)$, and recall that its support is the set of values of X in the closed interval $[0,1]$. Applying the logistic transformation yields

$$Y = f(X) = \ln\left(\frac{X}{1-X}\right),$$

where Y is often referred to as a logit. Note that Y is defined in the space $(-\infty, \infty)$. The inverse transformation is

$$f^{-1}(Y) = \frac{\exp(Y)}{1+\exp(Y)}.$$

The Jacobian of the transformation, in this case being positive for all values of y, is

$$\frac{d}{dy}\left(f^{-1}(y)\right) = \frac{\exp(y)}{[1+\exp(y)]^2}$$

so the p.d.f. of Y is

$$\begin{aligned} p(y) &= C\left[\frac{\exp(y)}{1+\exp(y)}\right]^{a-1}\left[\frac{1}{1+\exp(y)}\right]^{b-1}\frac{\exp(y)}{[1+\exp(y)]^2} \\ &= C\left\{\frac{[\exp(y)]^a}{[1+\exp(y)]^{a+b}}\right\}, \quad -\infty < y < \infty, \end{aligned} \tag{2.7}$$

where the constant is:

$$C = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

as given in (1.36) of the previous chapter. For appropriate values of the parameters a and b, the density (2.7) can be shown to approximate well that of a normally distributed random variable. This is why linear models are often employed for describing the variation of logistically transformed

variables taking values between 0 and 1 (such as probabilities) before transformation. For example, Gianola and Foulley (1983) described methods for genetic analysis of logits of probabilities, using Bayesian ideas. ■

Example 2.6 *The standard logistic distribution*
Consider the random variable Y from a logistic distribution with p.d.f.

$$p(y) = \frac{\exp(y)}{[1+\exp(y)]^2}.$$

This distribution has mean 0 and variance $\pi^2/3$. There may be interest in finding the distribution of the linear combination $f(Y) = Z = \alpha + \beta Y$, where α and β are constants. Here the inverse transformation is

$$f^{-1}(Z) = \beta^{-1}(Z-\alpha),$$

and the Jacobian of the transformation is β^{-1}. The density of interest is then

$$p(z) = \frac{\exp\left(\frac{z-\alpha}{\beta}\right)}{\beta\left[1+\exp\left(\frac{z-\alpha}{\beta}\right)\right]^2}.$$

The density characterizes the sech-squared distribution (Johnson and Kotz, 1970b), that has mean and variance

$$E(Z) = \alpha$$

and

$$Var(Z) = \frac{(\beta\pi)^2}{3},$$

respectively. If, in the transformation, one takes $\alpha = 0$ and $\beta = \sqrt{3}/\pi$, the density becomes

$$p(z) = \frac{\pi\exp\left(\frac{\pi z}{\sqrt{3}}\right)}{\sqrt{3}\left[1+\exp\left(\frac{\pi z}{\sqrt{3}}\right)\right]^2} \tag{2.8}$$

this being the density of a standard logistic distribution, with mean 0 and variance 1. ■

Example 2.7 *Moment generating function of the logistic distribution*
The moment generating function (see Chapter 1) of a random variable following a logistic distribution is

$$E[\exp(tX)] = \int_{-\infty}^{\infty} \exp(tx)\frac{\exp(x)}{[1+\exp(x)]^2}dx. \tag{2.9}$$

Change variables to

$$Y = \frac{\exp(x)}{[1+\exp(x)]},$$

so

$$1 - Y = \frac{1}{[1+\exp(x)]},$$

and note that the sampling space of Y goes from 0 to 1. The inverse transformation is

$$X = \ln\left(\frac{Y}{1-Y}\right)$$

with Jacobian

$$\frac{dX}{dY} = \frac{1}{Y(1-Y)}.$$

Observe now that

$$\exp(tX) = \exp\left[t\ln\left(\frac{Y}{1-Y}\right)\right] = Y^t(1-Y)^{-t}.$$

Using the preceding in (2.9) yields the moment generating function

$$\begin{aligned} E[\exp(tX)] &= \int_0^1 y^{1+t-1}(1-y)^{1-t-1}\,dy = B(1+t,1-t) \\ &= \frac{\Gamma(1+t)\Gamma(1-t)}{\Gamma(2)} = \Gamma(1+t)\Gamma(1-t), \end{aligned} \tag{2.10}$$

where $B(\cdot)$ is called the beta function and $\Gamma(\cdot)$ is the gamma function, seen in Chapter 1. ■

Example 2.8 *The inverse chi-square distribution from a gamma process* As seen in Chapter 1, a gamma random variable has as p.d.f.

$$p(x|a,b) = Ga(x|a,b) = \left[\Gamma(a)b^{-a}\right]^{-1} x^{a-1}\exp(-bx), \quad x > 0, \tag{2.11}$$

where a and b are strictly positive parameters. If one sets $a = \nu/2$ and $b = 1/2$, where ν is an integer, the above p.d.f. becomes that of a chi-square random variable. The parameter ν is known as the degrees of freedom of the distribution. The p.d.f. of a chi-square random variable is

$$p(x|\nu) = \left[\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}\right]^{-1} x^{\frac{\nu}{2}-1}\exp\left[-\frac{1}{2}x\right], \quad x > 0. \tag{2.12}$$

Let $Y = 1/X$ be an "inverse chi-square" random variable. Noting that

$$\frac{d}{dy}f^{-1}(y) = -y^{-2},$$

then the p.d.f. of the inverse chi-square distribution is

$$p(y|\nu) = \left[\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}\right]^{-1} y^{-\left(\frac{\nu}{2}-1\right)} \exp\left(-\frac{1}{2y}\right) y^{-2}$$
$$= \left[\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}\right]^{-1} y^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{1}{2y}\right), \quad y > 0. \tag{2.13}$$

■

Example 2.9 *The scaled inverse chi-square distribution*
A by-product of the inverse chi-square distribution plays an important role in variance component problems. Assume that X is a random variable following an inverse chi-square distribution. Let S be a positive nonrandom quantity called the scale parameter, and consider the transformation $Y = f(X) = SX$, with inverse $X = f^{-1}(Y) = Y/S$. Noting that the Jacobian is

$$\frac{d}{dy} f^{-1}(y) = \frac{1}{S},$$

then the p.d.f. of Y, using (2.4) and (2.13), is

$$p(y|\nu, S) = \left[\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}\right]^{-1} S^{\left(\frac{\nu}{2}\right)} y^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{S}{2y}\right)$$
$$\propto y^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{S}{2y}\right), \quad y > 0,\ \nu > 0,\ S > 0. \tag{2.14}$$

This is the density of the scaled inverse chi-square distribution. An alternative parameterization (employed often) is obtained by defining the scale parameter as $S = \nu S^*$. For example, in certain Bayesian variance component problems (Lindley and Smith, 1972; Box and Tiao, 1973; Gelfand and Smith, 1990), a scaled inverse chi-square distribution is used to represent prior uncertainty about a variance component. Here the value of S^* may be interpreted as a statement about the mean or mode of this prior distribution of the variance component, and ν as a degree of belief in such value. The term "degree of belief" is appealing because ν is a strictly positive, continuous parameter, whereas "degrees of freedom" is normally employed in connection with linear models to refer to integer quantities pertaining to the rank of certain matrices (Searle, 1971). It will be shown later that when ν tends to infinity, the mean of the scaled inverse chi-square distribution tends toward S^*.
The scaled inverse chi-square is a special case of the inverse gamma distribution, whose density is

$$p(x|a, b) = Cx^{-(a+1)} \exp(-b/x), \quad x > 0,\ a, b > 0, \tag{2.15}$$

where $C = b^a/\Gamma(a)$. Note that (2.14) can be retrieved from (2.15), by setting $a = \nu/2$ and $b = S/2$.

Suppose one seeks the mean of the distribution with density kernel (2.14). First, the propriety (integrability to a finite value) of the distribution will be assessed. Consider

$$\int_0^\infty Cx^{-\left(\frac{\nu}{2}+1\right)} \exp\left[-\frac{S}{2x}\right] dx. \tag{2.16}$$

To evaluate this integral, make the change of variable $y = 1/x$, so $dx = -y^{-2}\, dy$. Then (2.16) is expressible as

$$C\int_0^\infty y^{\left(\frac{\nu}{2}+1\right)} \exp\left[-\frac{S}{2}y\right] y^{-2}\, dy = C\int_0^\infty y^{\left(\frac{\nu}{2}-1\right)} \exp\left[-\frac{S}{2}y\right] dy.$$

Recalling the result from gamma integrals employed earlier in the book (for details, see Abramowitz and Stegun, 1972), for $\alpha > 0$ and $\lambda > 0$,

$$\int_0^\infty z^{\alpha-1} \exp\left[-\lambda z\right] dz = \frac{\Gamma(\alpha)}{\lambda^\alpha}.$$

Making use of this in the expression above yields

$$C\int_0^\infty y^{\left(\frac{\nu}{2}-1\right)} \exp\left[-\frac{S}{2}y\right] dy = C\frac{\Gamma\left(\frac{\nu}{2}\right)}{\left(\frac{S}{2}\right)^{\frac{\nu}{2}}}$$

since

$$C = \left[\frac{\Gamma\left(\frac{\nu}{2}\right)}{\left(\frac{S}{2}\right)^{\frac{\nu}{2}}}\right]^{-1}.$$

Hence, the distribution is proper provided S and ν are both positive. Computation of the expected value of the distribution requires evaluation of

$$\begin{aligned} E(X|\nu, S) &= C\int_0^\infty x\; x^{-\left(\frac{\nu}{2}+1\right)} \exp\left[-\frac{S}{2}x^{-1}\right] dx \\ &= C\int_0^\infty y^{\left(\frac{\nu}{2}\right)} \exp\left[-\frac{S}{2}y\right] y^{-2} dy \\ &= C\int_0^\infty y^{\left(\frac{\nu}{2}-1-1\right)} \exp\left[-\frac{S}{2}y\right] dy. \end{aligned}$$

Making use of the gamma integral again:

$$E(X|\nu, S) = C\frac{\Gamma\left(\frac{\nu}{2}-1\right)}{\left(\frac{S}{2}\right)^{\frac{\nu}{2}-1}}, \quad \frac{\nu}{2} - 1 > 0,$$

with the preceding condition required for the gamma integral to exist. Finally, recalling that $\Gamma(\nu) = (\nu-1)\,\Gamma(\nu-1)$, this expression reduces to

$$E(X|\nu, S) = \frac{S}{\nu-2}, \quad \nu > 2. \tag{2.17}$$

If $\nu \leq 2$, the expected value is not finite. A similar development leads to the result that if $\nu \leq 4$, the variance of the distribution is not defined. However, the scaled inverse chi-square distribution is still proper, provided that the degree of belief parameter is positive. Note that for $S = \nu S^*$, the expectation in (2.17) becomes

$$E\left(X|\nu, S^*\right) = \frac{\nu S^*}{\nu - 2}, \quad \nu > 2.$$

Hence, $E\left(X|\nu, S^*\right) \to S^*$ as $\nu \to \infty$. It can also be verified that the mode of the scaled inverted chi-square distribution is

$$Mode\left(X|\nu, S^*\right) = \frac{\nu S^*}{\nu + 2}.$$

■

Many-to-one Transformations

There are situations in which the transformation is not one-to-one. For example consider $Y = X^2$, where X is a random variable with known distribution. Here the transformation from X to Y is clearly not one-to-one, because both X and $-X$ produce the same value of Y. This section discusses how to cope with these many-to-one transformations.

To formalize the argument, let X be a continuous random variable with p.d.f. $p_X(x)$ and let Y define the many-to-one transformation $Y = f(X)$. If $\mathcal{A}$ denotes the space where $p(x) > 0$, and $\mathcal{B}$ is the space where $g(y) > 0$, then there exist points in $\mathcal{B}$ that correspond to more than one point in $\mathcal{A}$. However, if $\mathcal{A}$ can be partitioned into k sets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$, such that f_i defines a one-to-one transformation of each $\mathcal{A}_i$ onto $\mathcal{B}_i$ (the $\mathcal{B}_i$ can be overlapping), then the p.d.f. of Y is (i.e., Rao, 1973)

$$g(y) = \sum_{i=1}^{k} I_i\left(y \in \mathcal{B}_i\right) p_X\left[f_i^{-1}(y)\right] \left|J_i(y)\right|, \tag{2.18}$$

where the indicator function I_i is 1 if $y \in \mathcal{B}_i$ and 0 otherwise, and $J_i(y) = df_i^{-1}(y)/dy$ is the Jacobian of the transformation in partition i. That is, within each region i we work with (2.4), and then add all parts $i = 1, 2, \ldots, k$.

Example 2.10 *The square of a normal random variable: the chi-squared distribution*

Let $X \sim N(0, 1)$, with density

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty.$$

Let $Y = f(X) = X^2$. The transformation f is not one-to-one over the given domain. However, one can partition the domain into disjoint regions within which the transformation is one-to-one. These two regions in the space of X are $\mathcal{A}_1 = (-\infty, 0)$ and $\mathcal{A}_2 = (0, \infty)$. The corresponding regions in the space of Y are $\mathcal{B}_1 = (0, \infty)$ and $\mathcal{B}_2 = (0, \infty)$, which in this case are completely overlapping. In $(\mathcal{A}_1, \mathcal{B}_1)$, $f_1^{-1}(Y) = -Y^{1/2}$, and in $(\mathcal{A}_2, \mathcal{B}_2)$, $f_2^{-1}(Y) = Y^{1/2}$. The absolute value of the Jacobian of the transformation in both partitions is $\frac{1}{2}Y^{-1/2}$. The density of Y, applying (2.18), is

$$\begin{aligned} g(y) &= \frac{1}{2}y^{-1/2}\left[p_X\left(-y^{1/2}\right) + p_X\left(y^{1/2}\right)\right] \\ &= \frac{(1/2)^{1/2}}{\sqrt{\pi}}y^{-1/2}\exp\left(-\frac{y^2}{2}\right) \\ &= \frac{(1/2)^{1/2}}{\Gamma(1/2)}y^{-1/2}\exp\left(-\frac{y^2}{2}\right), \quad 0 < y < \infty, \end{aligned}$$

where the last equality arises because $\Gamma(1/2) = \sqrt{\pi}$. Here the indicator function I_i is not needed because the regions in the space of Y are completely overlapping. From (1.42), $g(y)$ is the p.d.f. of a random variable from a chi-squared distribution with one degree of freedom. ■

Example 2.11 *Many-to-one transformation with partially overlapping regions*
Let X have p.d.f.

$$p_X(x) = C\exp(x), \quad -1 \le x \le 2,$$

where C is the integration constant. Consider again the many-to-one transformation $Y = f(X) = X^2$. Define the regions in the space of X, $\mathcal{A}_1 = (-1, 0)$ and $\mathcal{A}_2 = (0, 2)$, with corresponding regions in the space of Y, $\mathcal{B}_1 = (0, 1)$ and $\mathcal{B}_2 = (0, 4)$, which are partially overlapping. In $(\mathcal{A}_1, \mathcal{B}_1)$, $f_1^{-1}(Y) = -Y^{1/2}$, and in $(\mathcal{A}_2, \mathcal{B}_2)$, $f_2^{-1}(Y) = Y^{1/2}$, as in the previous example. The absolute value of the Jacobian of the transformation in both partitions is again equal to $Y^{-1/2}/2$. The density of Y, applying (2.18), is

$$\begin{aligned} g(y) &= \frac{1}{2}y^{-1/2}\left[p_X\left(-y^{1/2}\right)I_1(y \in \mathcal{B}_1) + p_X\left(y^{1/2}\right)I_2(y \in \mathcal{B}_2)\right] \\ &= \frac{1}{2}y^{-1/2}\left[C\exp\left(-\sqrt{y}\right)I_1(y \in \mathcal{B}_1)\right] + \left[C\exp\left(\sqrt{y}\right)I_2(y \in \mathcal{B}_2)\right]. \end{aligned}$$

This can also be expressed as

$$g(y) = \begin{cases} \frac{1}{2}y^{-1/2}\left[C\exp\left(-\sqrt{y}\right)\right] + \left[C\exp\left(\sqrt{y}\right)\right], & 0 < y \le 1, \\ \frac{1}{2}y^{-1/2}C\exp\left(\sqrt{y}\right), & 1 < y \le 4, \end{cases}$$

which shows the discontinuity at $y = 1$ more transparently. ■

2.2.3 Approximating the Mean and Variance

Sometimes it is extremely difficult to arrive at the distribution of functions of random variables by analytical means. Often, one may just wish to have a rough idea of a distribution by using an approximation to its mean and variance. In this subsection and the following one, two widely employed and useful methods for approximating the mean and variance of the distribution of a function of a random variable are presented, and examples are given to illustrate. The approximation in this subsection is arguably based on a mathematical, rather than statistical argument, and has been used extensively in quantitative genetics, specially for obtaining standard errors of estimates of heritabilities and genetic correlations (e.g., Becker, 1984; Dempster and Lerner, 1950), as these involve nonlinear functions of estimates of variance and covariance components. As discussed later in this book, more powerful computer-based weaponry is presently available.

Let X be a random variable and let $Y = f(X)$ be a function of X that is differentiable at least twice. Expanding $f(X)$ in a Taylor series about $X = E[X] = \mu$ gives

$$Y = f(X) \cong f[\mu] + \left.\frac{d}{dX} f(X)\right|_{X=\mu} (X-\mu) + \frac{1}{2}\left.\frac{d^2}{(dX)^2} f(X)\right|_{X=\mu} (X-\mu)^2 . \tag{2.19}$$

Taking expectations, one obtains, as a second-order approximation to $E(Y)$

$$E[Y] \cong f[\mu] + \frac{1}{2}\left.\frac{d^2}{(dX)^2} f(X)\right|_{X=\mu} Var(X) . \tag{2.20}$$

Now, taking variances over approximation (2.19) and retaining only second order terms

$$Var(Y) \cong \left[\left.\frac{d}{dX} f(X)\right|_{X=\mu}\right]^2 Var(X) .$$

Sometimes, only the linear term is retained in (2.20), and one uses as a rough guide

$$Y \sim \left\{ f(\mu), \left[\left.\frac{d}{dX} f(X)\right|_{X=\mu}\right]^2 Var(X) \right\} . \tag{2.21}$$

Example 2.12 *Mean and variance of a transformed, beta-distributed random variable*

In Example 2.5, it was noted that a logistic transformation of a beta random variable gives an approximately normally distributed process. Thus,

if $X \sim Be(a,b)$, then the variable

$$Y = f(X) = \ln[X/(1-X)]$$

is approximately distributed as $N[E(Y), Var(Y)]$. The mean and variance of the beta distribution are

$$E(X|a,b) = \frac{a}{a+b}, \quad Var(X|a,b) = \frac{ab}{\left[(a+b)^2(a+b+1)\right]}. \tag{2.22}$$

The mean of Y, using (2.21), is

$$E(Y|a,b) \cong \ln\left[\frac{E(X)}{1-E(X)}\right] = \ln\left(\frac{a}{b}\right),$$

and using (2.20), is

$$E(Y|a,b) \cong \ln\left(\frac{a}{b}\right) + \frac{(a-b)(a+b)}{2ab(a+b+1)}.$$

From (2.21) the resulting variance is

$$Var(Y|a,b) \cong \frac{(a+b)^2}{ab(a+b+1)}.$$

■

Example 2.13 *Genotypes in Gaussian and discrete scales: the threshold model*

Dempster and Lerner (1950) discussed the quantitative genetic analysis of a binary character (Y_o) following the ideas of Wright (1934) and of Robertson and Lerner (1949). These authors assumed that the expression of the trait ($Y_o = 0$ = attribute absent, $Y_o = 1$ = attribute present) is related to an underlying, unobservable normal process, and that gene substitutions take place at this level. Let

$$Y = \mu + G + E$$

be the Gaussian variable, where μ is the mean of Y and G and E are random terms representing the genetic and environmental effects on Y, respectively. Suppose that $G \sim N(0, V_G)$, $E \sim N(0, V_E)$ have independent distributions. Hence, the marginal distribution of the latent variable is $Y \sim N(\mu, V_G + V_E)$. Dempster and Lerner (1950) defined "genotype in the observable scale" as the conditional probability of observing the attribute, given the genotype in the latent scale, that is

$$\Pr(Y_o = 1|G) = \Pr(Y > t|G),$$

where t is a threshold (assume, subsequently, that $t = 0$). In other words, the attribute is observed if the value of the latent variable exceeds the threshold. The "outward" genotype can be expressed as

$$\begin{aligned} G_0 &= \Pr\left(Y_o = 1|G\right) = \Pr(Y - \mu - G > t - \mu - G|G) \\ &= 1 - \Pr\left(Z \leq \frac{-\mu - G}{\sqrt{V_E}}\right) \\ &= \Phi\left(\frac{\mu + G}{\sqrt{V_E}}\right), \end{aligned}$$

where $Z \sim N(0,1)$. A first-order approximation of the outward genotype about 0, the mean of the distribution of G, is

$$G_0 \cong \Phi\left(\frac{\mu}{\sqrt{V_E}}\right) + \phi\left(\frac{\mu}{\sqrt{V_E}}\right)\frac{G}{\sqrt{V_E}}.$$

The genetic variance in the outward scale is, approximately,

$$Var\left(G_0\right) = Var\left[\Phi\left(\frac{\mu + G}{\sqrt{V_E}}\right)\right] \cong \phi^2\left(\frac{\mu}{\sqrt{V_E}}\right)\frac{V_G}{V_E}.$$

The heritability (ratio between genetic and total variance) in the underlying scale is

$$h^2 = \frac{V_G}{V_G + V_E}$$

so "heritability in the outward scale" is approximately:

$$h_o^2 = \frac{Var\left(G_0\right)}{Var\left(Y_o\right)} \cong \frac{\phi^2\left(\frac{\mu}{\sqrt{V_E}}\right)V_G/V_E}{Var\left(Y_o\right)}.$$

Now

$$\begin{aligned} Var\left(Y_o\right) &= E\left(Y_o^2\right) - E^2\left(Y_o\right) \\ &= 0^2 \times \Pr(Y < t) + 1^2 \Pr\left(Y \geq t\right) - \left[\Pr\left(Y \geq t\right)\right]^2 \\ &= \Pr\left(Y \geq t\right)\left[1 - \Pr\left(Y \geq t\right)\right] \\ &= \Phi\left(\frac{-\mu}{\sqrt{V_G + V_E}}\right)\left[1 - \Phi\left(\frac{-\mu}{\sqrt{V_G + V_E}}\right)\right] \\ &= \left[1 - \Phi\left(\frac{\mu}{\sqrt{V_G + V_E}}\right)\right]\Phi\left(\frac{\mu}{\sqrt{V_G + V_E}}\right), \end{aligned}$$

so

$$h_o^2 = \frac{\phi^2\left(\frac{\mu}{\sqrt{V_E}}\right)V_G/V_E}{\left[1 - \Phi\left(\frac{\mu}{\sqrt{V_G + V_E}}\right)\right]\Phi\left(\frac{\mu}{\sqrt{V_G + V_E}}\right)}.$$

Since the latent variable cannot be observed, one can take the residual standard deviation in the underlying scale to be equal to 1, so all terms are expressed in units of $\sqrt{V_E}$. In this scale

$$h_o^2 = \frac{\phi^2(\mu)\, h^2/\left(1-h^2\right)}{\left[1-\Phi\left(\frac{\mu}{\sqrt{V_G+1}}\right)\right]\Phi\left(\frac{\mu}{\sqrt{V_G+1}}\right)}.$$

Alternatively, one could take $V_G + V_E = 1$, so $V_G = h^2$, heritability in the latent scale. Here

$$h_o^2 = \frac{\phi^2\left(\frac{\mu}{\sqrt{1-h^2}}\right) h^2/\left(1-h^2\right)}{\left[1-\Phi(\mu)\right]\Phi(\mu)}.$$

The two expressions for h_o^2 do not coincide with what was given by Dempster and Lerner (1950). The reason is that these authors used a different linear approximation to the genotype in the discrete scale.
Consider now a second-order approximation for the genotype in the outward scale. Here

$$G_0 = \Phi\left(\frac{\mu+G}{\sqrt{V_E}}\right)$$
$$\cong \Phi\left(\frac{\mu}{\sqrt{V_E}}\right) + \phi\left(\frac{\mu}{\sqrt{V_E}}\right)\frac{G}{\sqrt{V_E}} - \frac{\mu}{V_E}\phi\left(\frac{\mu}{\sqrt{V_E}}\right)G^2.$$

The genetic variance in the outward scale in this case is

$$Var(G_0) \cong \phi^2\left(\frac{\mu}{\sqrt{V_E}}\right)\frac{V_G}{V_E} + \left[\frac{\mu}{V_E}\phi\left(\frac{\mu}{\sqrt{V_E}}\right)\right]^2 Var(G^2)$$
$$+ 2\phi^2\left(\frac{\mu}{\sqrt{V_E}}\right)\frac{\mu}{(V_E)^{\frac{3}{2}}}Cov(G, G^2).$$

Using the moment generating function of the normal distribution or, directly, results for the variance of quadratic forms on normal variates (and for the covariance between a linear and a quadratic form, (Searle, 1971)), it can be established that

$$Var\left(G^2\right) = 2\left(V_G\right)^2$$

and

$$Cov\left(G, G^2\right) = 0.$$

Hence, the heritability in the outward scale resulting from the second-order approximation is

$$h_o^2 = \frac{\left\{\phi^2\left(\frac{\mu}{\sqrt{V_E}}\right) + 2\left[\mu\phi\left(\frac{\mu}{\sqrt{V_E}}\right)\right]^2(V_G/V_E)\right\}(V_G/V_E)}{\left[1-\Phi\left(\frac{\mu}{\sqrt{V_G+V_E}}\right)\right]\Phi\left(\frac{\mu}{\sqrt{V_G+V_E}}\right)}.$$

The threshold model is revisited in Chapters 4 and 14. ■

2.2.4 Delta Method

The approach based on the Taylor series described above yields approximate formulas for means and variances of functions of random variables. However, nothing is said here about the distributional properties of the derived statistics. A related large-sample based technique, that is more formally anchored statistically, known as the delta method, does this. Borrowing from Lehmann (1999), let T_n be a random variable where the subscript expresses its dependence on sample size n. As n increases towards infinity, suppose that the sequence of c.d.f.s of $\sqrt{n}\,(T_n - \mu)$ converges to the c.d.f. of a normal variable with mean 0 and variance σ^2. This limiting behavior is known as convergence in distribution, denoted here by

$$\sqrt{n}\,(T_n - \mu) \xrightarrow{D} N\left(0, \sigma^2\right). \tag{2.23}$$

The delta method provides the following limiting distribution for a function of T_n, $f\,(T_n)$:

$$\sqrt{n}\,[f\,(T_n) - f\,(\mu)] \xrightarrow{D} N\left(0, \sigma^2\left[f'\,(\mu)\right]^2\right), \tag{2.24}$$

where $f'\,(\mu)$ denotes the first derivative of $f\,(T_n)$ evaluated at μ. The proof of this result is based on a Taylor expansion of $f\,(T_n)$ around $f\,(\mu)$:

$$f\,(T_n) = f\,(\mu) + (T_n - \mu)\,f'\,(\mu) + o_p\,(T_n - \mu). \tag{2.25}$$

The notation $o_p\,(T_n - \mu)$ denotes a random variable of smaller order than $T_n - \mu$ for large n, in the sense that, for fixed $\epsilon > 0$,

$$\Pr\left(o_p\,(T_n - \mu)/\,(T_n - \mu) \leq \epsilon\right) \to 1$$

as $n \to \infty$. Therefore this last term converges in probability to 0 as n increases toward infinity. (If X_n converges in probability to X, then for $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \epsilon\right) = 0$$

and hence, for large n, $X_n \approx X$). Subtracting $f\,(\mu)$ from both sides and multiplying by $\sqrt{n}$ yields

$$\sqrt{n}\,[f\,(T_n) - f\,(\mu)] = \sqrt{n}\,(T_n - \mu)\,f'\,(\mu) + \sqrt{n}o_p\,(T_n - \mu).$$

As $n \to \infty$, the second term in the right-hand side vanishes; therefore, the left-hand side has the same limiting distribution as

$$\sqrt{n}\,(T_n - \mu)\,f'\,(\mu).$$

Since $T_n - \mu$ is approximately normal with variance σ^2/n, then $f(T_n) - f(\mu)$ is approximately normal with variance $\sigma^2 [f'(\mu)]^2 / n$ and result (2.24) follows.

When the term $f'(\mu) = 0$, in (2.25), it is natural to carry the expansion to a higher-order term, provided that the second derivative exists and that it is not 0:

$$f(T_n) = f(\mu) + \frac{1}{2}(T_n - \mu)^2 f''(\mu) + o_p(T_n - \mu)^2,$$

where $f''(\mu)$ is the second derivative of $f(T_n)$ evaluated at μ. The last term tends to 0 for large n; therefore we can write, loosely,

$$n[f(T_n) - f(\mu)] = \frac{1}{2} f''(\mu)\, n(T_n - \mu)^2. \tag{2.26}$$

Now from (2.23) it follows that

$$\frac{n(T_n - \mu)^2}{\sigma^2} \to \chi_1^2.$$

Therefore, using (2.26),

$$n[f(T_n) - f(\mu)] \to \frac{1}{2}\sigma^2 f''(\mu)\, \chi_1^2. \tag{2.27}$$

Example 2.14 *Bernoulli random variables*
Let X_i $(i = 1, 2, \ldots)$ be independent Bernoulli random variables with parameter θ and let $T_n = n^{-1} \sum_{i=1}^n X_i$. By the central limit theorem,

$$\sqrt{n}(T_n - \theta) \to N[0, \theta(1-\theta)]$$

because $E(T_n) = \theta$ and $Var(T_n) = \theta(1-\theta)/n$. Imagine that one is interested in the large sample behavior of the statistic $f(T_n) = T_n(1 - T_n)$ as an estimate of $f(\theta) = \theta(1-\theta)$. From (2.24), since $f'(\theta) = 1 - 2\theta$,

$$\sqrt{n}[T_n(1-T_n) - \theta(1-\theta)] \to N\left[0, \theta(1-\theta)(1-2\theta)^2\right]$$

for $\theta \neq 1/2$. When $\theta = 1/2$, $f'(1/2) = 0$. Then, using (2.27), for $\theta = 1/2$, since $f(1/2) = 1/4$ and $f''(1/2) = -2$,

$$n\left[T_n(1-T_n) - \frac{1}{4}\right] \to \frac{1}{2}\frac{1}{4}(-2)\chi_1^2 = -\frac{1}{4}\chi_1^2$$

or, equivalently,

$$4n\left[\frac{1}{4} - T_n(1-T_n)\right] \to \chi_1^2.$$

■

The delta method generalizes straightforwardly to functions of random vectors. Let $\mathbf{T}_n = (T_{n1}, T_{n2}, \ldots, T_{np})'$ be asymptotically multivariate normal, with mean $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ and covariance matrix $\boldsymbol{\Sigma}/n$. Suppose the function $f(t_1, t_2, \ldots, t_p)$ has nonzero differential $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_p)'$ at $\boldsymbol{\theta}$, where

$$\Delta_i = \left.\frac{\partial f}{\partial t_i}\right|_{\mathbf{t}=\boldsymbol{\theta}}.$$

Then,

$$\sqrt{n}\left[f(\mathbf{T}_n) - f(\boldsymbol{\theta})\right] \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Delta}'\boldsymbol{\Sigma}\boldsymbol{\Delta}\right). \tag{2.28}$$

2.3 Transformations Involving Several Discrete or Continuous Random Variables

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ be a random vector with p.m.f. or p.d.f. equal to $p_X(\mathbf{x})$, and let $\mathbf{Y} = f(\mathbf{X})$ be a one-to-one transformation. Let the sample space of $\mathbf{X}$, denoted as $S \subseteq \mathbb{R}^n$, be such that

$$\Pr\left[(X_1, X_2, \ldots, X_n) \in S\right] = 1,$$

or, equivalently, that the integral of $p_X(\mathbf{x})$ over S is equal to 1. Define $T \subseteq \mathbb{R}^n$ to be the image of S under the transformation, that is, as the values of $X_1, X_2, \ldots, X_n$ vary over S, the values of $Y_1, Y_2, \ldots, Y_n$ vary over T. Corresponding to each value of $Y_1, Y_2, \ldots, Y_n$ in the set T there is a unique value of $X_1, X_2, \ldots, X_n$ in the set S, and vice-versa. This ensures that the inverse transformation exists and this is denoted by $\mathbf{X} = f^{-1}(\mathbf{Y})$.

The elements of $\mathbf{Y} = \mathbf{f}(\mathbf{X})$ are

$$\begin{aligned}
Y_1 &= f_1(X_1, X_2, \ldots, X_n), \\
Y_2 &= f_2(X_1, X_2, \ldots, X_n), \\
&\vdots \\
Y_n &= f_n(X_1, X_2, \ldots, X_n),
\end{aligned}$$

whereas $\mathbf{X} = f^{-1}(\mathbf{Y})$ has elements

$$\begin{aligned}
X_1 &= f_1^{-1}(Y_1, Y_2, \ldots, Y_n), \\
X_2 &= f_2^{-1}(Y_1, Y_2, \ldots, Y_n), \\
&\vdots \\
X_n &= f_n^{-1}(Y_1, Y_2, \ldots, Y_n).
\end{aligned}$$

Now let S' be a subset of S and let T' denote the mapping of S' under the transformation. The events $(X_1, X_2, \ldots, X_n) \in S'$ and $(Y_1, Y_2, \ldots, Y_n) \in T'$

are said to be equivalent. Hence,

$$\Pr\left[(Y_1, Y_2, \ldots, Y_n) \in T'\right] = \Pr\left[(X_1, X_2, \ldots, X_n) \in S'\right] \tag{2.29}$$

which, in the continuous case, is equal to

$$\int p_X\left(x_1, x_2, \ldots, x_n\right) dx_1\, dx_2 \ldots dx_n,$$

where the integral is multidimensional and taken over S'. In order to find the p.d.f. of $\mathbf{Y}$, a change of variable of integration must be effected in (2.29), such that $x_1 = f_1^{-1}(\mathbf{y}), x_2 = f_2^{-1}(\mathbf{y}), \ldots, x_n = f_n^{-1}(\mathbf{y})$. In calculus books (e.g., Kaplan, 1993) it is shown that this change of variables results in the following expression:

$$\begin{aligned}&\Pr\left[(Y_1, Y_2, \ldots, Y_n) \in T'\right] \\ &= \int p_X\left(f_1^{-1}(\mathbf{y}), f_2^{-1}(\mathbf{y}), \ldots, f_n^{-1}(\mathbf{y})\right) |\mathbf{J}(\mathbf{y})|\, dy_1 dy_2 \ldots dy_n,\end{aligned} \tag{2.30}$$

where $|\mathbf{J}(\mathbf{y})|$ is the absolute value of the Jacobian of the transformation. This implies that the p.d.f. of the vector $\mathbf{Y}$ is (writing from now onwards $\mathbf{J}(\mathbf{y}) = \mathbf{J}$)

$$p_Y(\mathbf{y}) = \begin{cases} p_X\left(f_1^{-1}(\mathbf{y}), f_2^{-1}(\mathbf{y}), \ldots, f_n^{-1}(\mathbf{y})\right) |\mathbf{J}|, & \mathbf{y} \in T, \\ 0, \quad \text{otherwise.} \end{cases} \tag{2.31}$$

In the multivariate situation, the Jacobian is the determinant of a matrix of first derivatives and is given by

$$\mathbf{J} = \det \begin{bmatrix} \dfrac{\partial f_1^{-1}(\mathbf{y})}{\partial y_1} & \cdots & \dfrac{\partial f_1^{-1}(\mathbf{y})}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n^{-1}(\mathbf{y})}{\partial y_1} & \cdots & \dfrac{\partial f_n^{-1}(\mathbf{y})}{\partial y_n} \end{bmatrix}, \tag{2.32}$$

where $\partial f_j^{-1}(\mathbf{y}) / \partial y_k$ is the first partial derivative of the jth element of $f^{-1}(\mathbf{Y})$ with respect to the kth element of $\mathbf{Y}$. The Jacobian is also denoted

$$\mathbf{J} = \det \left[\frac{\partial\left(x_1, x_2, \ldots, x_n\right)}{\partial\left(y_1, y_2, \ldots, y_n\right)}\right]. \tag{2.33}$$

It may be that there is only one function of interest, for example, $Y_1 = f_1(X_1, \ldots, X_n)$. By defining appropriate additional arbitrary functions f_2, $\ldots, f_n$, such that the transformation is one-to-one, then the joint p.d.f. of $\mathbf{Y}$ can be obtained by the method above. The p.d.f. of Y_1 can be derived subsequently by integrating $p_Y(\mathbf{y})$ over the space spanned by $Y_2, Y_3 \ldots, Y_n$.

For discrete random variables, the joint probability of $\mathbf{Y} = \mathbf{y}$, defined within the sample space T, is given directly by

$$\begin{aligned}&\Pr\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\right)\\&= p\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right),\end{aligned} \tag{2.34}$$

where $x_1 = f_1^{-1}\left(y_1, y_2, \ldots, y_n\right), \ldots, x_n = f_n^{-1}\left(y_1, y_2 \ldots, y_n\right)$.

If $\mathbf{Y} = f\left(\mathbf{X}\right)$ is a many-to-one differentiable transformation, the density of $\mathbf{Y}$ can be obtained by applying (2.31) to each solution of the inverse transformation separately, and then summing the transformed densities for each solution. This is exactly in the same spirit as in (2.18).

The Derivative and the Jacobian as a Local Magnification of a Projection

In this section a heuristic motivation of expression (2.31) is provided. A more rigorous treatment can be found in standard books on multivariate calculus (e.g., Kaplan, 1993; Williamson et al., 1972).

Consider the projection of a point x of a line onto the point $g\left(x\right)$ of another line. The magnification at x, when x is projected onto $g\left(x\right)$, is the absolute value of the derivative of g at x. For example, if $g\left(x\right) = 2x$, then the magnification at $x = u$ is $\left|g'\left(u\right)\right| = 2$. There is a magnification by a factor of 2 at all points, irrespective of the value of x. This means that a distance $R = x_1 - x$ on the original line is projected onto a distance $S = g\left(x_1\right) - g\left(x\right)$, and that the magnification of the distance at x is given by the absolute value of the derivative of g evaluated at x:

$$\left|\lim_{x_1 \to x} \frac{g\left(x_1\right) - g\left(x\right)}{x_1 - x}\right| = \left|g'\left(x\right)\right|. \tag{2.35}$$

As $x_1 \to x$, the length S is, approximately,

$$\text{length of } S = \left|g'\left(x\right)\right| \text{length of } R$$

or, equivalently,

$$dS = \left|g'\left(x\right)\right| dR. \tag{2.36}$$

This argument extends to the multivariate case as follows, and two variables are used to illustrate. Let R be a region in the uv plane and let S be a region in the xy plane. Consider a mapping given by the functions

$$\begin{aligned}x &= f_1\left(u, v\right),\\y &= f_2\left(u, v\right),\end{aligned}$$

and let S be the image of R under this mapping. As R approaches zero,

$$\frac{\text{area of } S}{\text{area of } R} \to \left|\mathbf{J}\right|$$

or, equivalently,

$$\text{area of } S = |\mathbf{J}| \text{ area of } R, \tag{2.37}$$

where $|\mathbf{J}|$ is the absolute value of the Jacobian of the transformation. The Jacobian is equal to the determinant

$$\mathbf{J} = \det\left[\frac{\partial(f_1, f_2)}{\partial(u, v)}\right]$$

which is a function of (u, v). The absolute value of the Jacobian is a measure of the local magnification at the point (u, v). In the bivariate case, if area of S is approximated by $dx\,dy$ and area of R is approximated by $du\,dv$, (2.37) can be written as

$$dx\,dy = |\mathbf{J}|\,du\,dv. \tag{2.38}$$

Then,

$$\int_S h(x, y)\,dx\,dy = \int_R h(f_1(u, v), f_2(u, v))\,|\mathbf{J}|\,du\,dv. \tag{2.39}$$

In a trivariate case, the absolute value of the Jacobian would represent the magnification of a volume, and so on.

Example 2.15 *Transforming a multinomial distribution*

This example illustrates a multivariate transformation in the discrete case and, also, how a conditional distribution can be derived from the joint and marginal probability distributions of the transformed variables. As seen in Chapter 1, the multinomial distribution applies to a sampling model where n independent draws are made from the same population. The outcome of each draw is a realization into one of C mutually exclusive and exhaustive classes, or categories of response. The categories can be ordered or unordered. For example, in beef cattle breeding the degree of ease of calving is often scored into four classes ($C = 4$), for example, "no difficulty", "some assistance is needed", "mechanical pull", or "caesarean section required", so the categories are ordered. On the other hand, unordered categories appear, for example, in genetic analyses of leg deformities in chickens. Here the possible classes cannot be ordered in a meaningful way.

Let n_i be the number of observations falling into the ith class, and let p_i be the probability that an individual observation falls in the ith class, for $i = 1, 2, 3$. Then $n = n_1 + n_2 + n_3$, and the joint probability function of (n_1, n_2) is (the dependence on parameters p_1, p_2, and n is omitted)

$$p(n_1, n_2) = \frac{n!}{n_1!\,n_2!\,(n - n_1 - n_2)!} p_1^{n_1} p_2^{n_2} (1 - p_1 - p_2)^{n - n_1 - n_2}. \tag{2.40}$$

Suppose that one needs to find the conditional probability distribution of n_1, given $n_1 + n_2$. That is, the conditional probability function is

$$p(n_1 | n_1 + n_2) = \frac{p(n_1, n_1 + n_2)}{p(n_1 + n_2)}. \tag{2.41}$$

To simplify the notation, let $(n_1, n_2) = (X, Y)$ and $(n_1, n_1 + n_2) = (U, V)$, and omit the parameters (p_1, p_2) as arguments of $p(\cdot)$. The probability function of $(n_1, n_2) = (X, Y)$ is then expressible as

$$p(X,Y) = \frac{n!}{X!\,Y!\,(n-X-Y)!} p_1^X p_2^Y (1-p_1-p_2)^{n-X-Y}. \tag{2.42}$$

To derive the numerator in (2.41), that is, $p(n_1, n_{1+}n_2) = p(U, V)$, note that the transformation $(X, Y) \to (U, V)$ can be written as

$$\begin{bmatrix} U \\ V \end{bmatrix} = f(X,Y) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ X+Y \end{bmatrix} \tag{2.43}$$

with inverse transformation

$$\begin{bmatrix} X \\ Y \end{bmatrix} = f^{-1}(U,V) = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V-U \end{bmatrix}.$$

Therefore, from (2.34),

$$p_{U,V}(U,V) = \frac{n!}{U!\,(V-U)!\,(n-V)!} p_1^U p_2^{V-U} (1-p_1-p_2)^{n-V}. \tag{2.44}$$

To obtain $p(n_1|n_1+n_2)$, (2.44) must be divided by $p(V)$. Now the random variable V follows the binomial distribution

$$V \sim Bi(p_1+p_2, n).$$

This is so because the three classes can be regrouped into two "wider" categories, one where the counts $(n_1 + n_2)$ are observed to fall, and the other involving the third original category with counts n_3. In view of the independence of the draws, it follows that $(n_1 + n_2)$ is binomially distributed. Dividing $p_{U,V}(U, V)$ in (2.44) by the marginal probability function $p_V(V)$ we obtain

$$\begin{aligned} p(n_1|n_1+n_2) &= p(U|V) \\ &= \frac{p(U,V)}{p(V)} \\ &= \frac{V!}{U!\,(V-U)!} \frac{p_1^U p_2^{V-U}}{(p_1+p_2)^V} \\ &= \frac{(n_1+n_2)!}{n_1!\,n_2!} \frac{p_1^{n_1} p_2^{n_2}}{(p_1+p_2)^{n_1+n_2}} \\ &= \frac{(n_1+n_2)!}{n_1!\,n_2!} \left(\frac{p_1}{p_1+p_2}\right)^{n_1} \left(\frac{p_2}{p_1+p_2}\right)^{n_2}. \end{aligned}$$

This implies that

$$[n_1|n_1+n_2] \sim Bi\left(\frac{p_1}{p_1+p_2}, n_1+n_2\right).$$

Hence, the conditional distribution $[n_1|n_1+n_2]$ has mean,

$$E\left(n_1|n_1+n_2\right)=\left(n_1+n_2\right)\frac{p_1}{p_1+p_2}$$

and variance

$$Var\left(n_1|n_1+n_2\right)=\left(n_1+n_2\right)\frac{p_1}{p_1+p_2}\frac{p_2}{p_1+p_2}.$$

■

Example 2.16 *Distribution of the ratio between two independent random variables*

Suppose that two random variables are i.i.d., with densities equal to $p\left(x_i\right)=2x_i$, for $i=1,2$, with the sample space being the set of all points contained in the interval $(0,1)$. Their joint p.d.f. is then

$$p\left(x_1,x_2\right)=\begin{cases}4x_1x_2, & \text{for } 0<x_1<1 \text{ and } 0<x_2<1,\\ 0, & \text{otherwise.}\end{cases} \tag{2.45}$$

We wish to find the p.d.f. of the ratio $Y_1=X_1/X_2$. This is a situation where an auxiliary variable is needed in order to make the transformation one-to-one. To find the p.d.f. of the ratio, the auxiliary variable is integrated out from the joint density of all transformed variables. Let the auxiliary variable be $Y_2=X_2$. Then

$$\begin{aligned}Y_1&=f_1\left(X_1,X_2\right)=X_1/X_2,\\ Y_2&=f_2\left(X_1,X_2\right)=f_2\left(X_2\right)=X_2.\end{aligned}$$

The inverse transformation is

$$\begin{aligned}X_1&=f_1^{-1}\left(Y_1,Y_2\right)=Y_1Y_2,\\ X_2&=f_2^{-1}\left(Y_1,Y_2\right)=f_2^{-1}\left(Y_2\right)=Y_2.\end{aligned}$$

In view of the support of $p\left(x_1,x_2\right)$, to find the sample space of the joint distribution $[Y_1,Y_2]$, observe that

$$\begin{aligned}0<y_1<\infty,\\ 0<y_2<1.\end{aligned}$$

Now, from

$$0<y_1y_2<1,$$

the following relationship also holds

$$0<y_2<\frac{1}{y_1}.$$

The Jacobian of the transformation is

$$\begin{aligned}\mathbf{J} &= \det\begin{bmatrix} \dfrac{\partial f_1^{-1}(y_1,y_2)}{\partial y_1} & \dfrac{\partial f_1^{-1}(y_1,y_2)}{\partial y_2} \\ \dfrac{\partial f_2^{-1}(y_1,y_2)}{\partial y_1} & \dfrac{\partial f_2^{-1}(y_1,y_2)}{\partial y_2} \end{bmatrix} \\ &= \det\begin{bmatrix} \dfrac{\partial (y_1 y_2)}{\partial y_1} & \dfrac{\partial (y_1 y_2)}{\partial y_2} \\ \dfrac{\partial y_2}{\partial y_1} & \dfrac{\partial y_2}{\partial y_2} \end{bmatrix} \\ &= \det\begin{bmatrix} y_2 & y_1 \\ 0 & 1 \end{bmatrix} = y_2.\end{aligned}$$

The p.d.f. of the vector $\mathbf{Y} = [Y_1, Y_2]'$ is obtained as follows: in the density $p(x_1, x_2)$, replace x_1 by $y_1 y_2$, x_2 by y_2, and then multiply the result by the absolute value of $\mathbf{J}$. This leads to

$$p(y_1, y_2) = \begin{cases} 4y_1 y_2^3, & 0 < y_1 < \infty, 0 < y_2 < 1, 0 < y_1 y_2 < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.46}$$

We now check whether or not this is a proper p.d.f.: the integral of $p(y_1, y_2)$ over the sampling space induced by the transformation must be finite. We then have

$$\begin{aligned}&\int\int p(y_1, y_2)\, dy_1\, dy_2 \\ &= \int_{y_2=0}^{1}\int_{y_1=0}^{1} p(y_1, y_2)\, dy_1 dy_2 + \int_{y_2=0}^{1/y_1}\int_{y_1=1}^{\infty} p(y_1, y_2)\, dy_1\, dy_2 \\ &= \int_{y_1=0}^{1} 4y_1 \left(\frac{y_2^4}{4}\bigg|_0^1 \right) dy_1 + \int_{y_1=1}^{\infty} 4y_1 \left[\frac{y_2^4}{4}\bigg|_0^{\frac{1}{y_1}} \right] dy_1 \\ &= \frac{1}{2} + \frac{1}{2} = 1.\end{aligned}$$

Thus, propriety is established. The marginal p.d.f. of $Y_1 = X_1/X_2$ is obtained by integrating the joint density with respect to Y_2, yielding

$$p(y_1) = \begin{cases} \displaystyle\int_{y_2=0}^{1} p(y_1, y_2)\, dy_2 = 4y_1 \left(\frac{y_2^4}{4}\bigg|_0^1 \right) = y_1,\ 0 < y_1 < 1, \\ \displaystyle\int_{y_2=0}^{1} p(y_1, y_2)\, dy_2 = 4y_1 \left(\frac{y_2^4}{4}\bigg|_0^{\frac{1}{y_1}} \right) = \frac{1}{y_1^3},\ 1 < y_1 < \infty. \end{cases}$$

Consider now the calculation of

$$\begin{aligned}\Pr\left(X_1 < \frac{1}{2}, X_2 < 0.7 \right) &= \int_{x_2=0}^{0.7}\int_{x_1=0}^{\frac{1}{2}} 4x_1 x_2\, dx_1\, dx_2 \\ &= 0.1225.\end{aligned}$$

In view of the relationship between (X_1, X_2) and (Y_1, Y_2) this joint probability can be written in terms of (Y_1, Y_2) as

$$\begin{aligned}\Pr\left(Y_1Y_2 < \frac{1}{2}, Y_2 < 0.7\right) &= \Pr\left(Y_1 < \frac{1}{2Y_2}, Y_2 < 0.7\right) \\ &= \int_{y_2=0}^{0.7} \int_{y_1=0}^{\frac{1}{2y_2}} 4y_1y_2^3 \, dy_1 \, dy_2 \\ &= 0.1225,\end{aligned}$$

corroborating (2.29). ■

Example 2.17 *Parameterization of a variance components model*
Consider a Gaussian linear mixed effects model for animal breeding data, and let this model have two sets of random effects with variances σ_a^2 and σ_e^2, respectively, both unknown. For example, σ_a^2 could be the additive genetic variance, and σ_e^2 the environmental variance for a certain trait. Suppose that in a Bayesian setting (where unknown parameters are treated as random variables), the two variance components (both strictly positive) are assigned a proper joint prior density equal to

$$p(\sigma_a^2, \sigma_e^2) = p(\sigma_a^2)p(\sigma_e^2), \quad \sigma_a^2 \geq 0, \, \sigma_e^2 > 0,$$

so that there is independence between the two components, a priori. Suppose, further, that there is a family structure, so that observations can be clustered into families of half-sibs, such that the observations within a cluster are equicorrelated, whereas those in different clusters are independent. Let the variance between half-sib clusters be σ_s^2 and the variance within clusters be σ_w^2. Conceivably, one may wish to parameterize the model in terms of random effects having variances σ_s^2 and σ_w^2. From a classical perspective, the two models are said to be equivalent (Henderson, 1984) if the same likelihood (see Chapter 3 for a formal definition of the concept) is conferred by the data to values of the same parameter under each of the models. For this equivalence to hold, a relationship is needed to establish a link between the two sets of parameters, and this comes from genetic theory. Under additive inheritance (Fisher, 1918) in a randomly mated population in linkage equilibrium, one has:
1) $\sigma_s^2 = \frac{1}{4}\sigma_a^2$, and
2) $\sigma_w^2 = \frac{3}{4}\sigma_a^2 + \sigma_e^2$.
This is a one-to-one linear transformation expressible in matrix notation as

$$\begin{bmatrix} \sigma_s^2 \\ \sigma_w^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & 0 \\ \frac{3}{4} & 1 \end{bmatrix} \begin{bmatrix} \sigma_a^2 \\ \sigma_e^2 \end{bmatrix}.$$

Since heritability $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ necessarily takes values between 0 and 1, it follows that σ_s^2/σ_w^2 must take a value between 0 and 1/3. In order to

arrive at the probabilistic beliefs in the parameterization in terms of half-sib clusters, one must effect a change of variables in the prior distribution. It turns out that the joint prior density of σ_s^2 and σ_w^2 is:

$$p\left(\sigma_s^2, \sigma_w^2\right) = 4p\left(4\sigma_s^2\right) p\left(\sigma_w^2 - 3\sigma_s^2\right), \quad \sigma_w^2 > 0,\ 0 \leq \sigma_s^2 \leq \frac{\sigma_w^2}{3}. \tag{2.47}$$

This shows that σ_s^2 and σ_w^2 are not independent a priori. It follows that a Bayesian that takes σ_a^2 and σ_e^2 as independent a priori, has different prior beliefs than one that assigns independent prior distributions to σ_s^2 and σ_w^2. Even if the likelihood in the two models confers the same strength to values of the same parameter, the two models would not be equivalent, at least in the Bayesian sense, unless the joint prior density in the second parameterization has the form given in (2.47). This illustrates that constructing models that are probabilistically consistent (or, in a Bayesian context, that reflect beliefs in a coherent manner) require careful consideration not only of the statistics, but of the subject matter of the problem as well. ■

Example 2.18 *Conditioning on a function of random variables*
Suppose there are two continuously distributed random vectors $\mathbf{x}$ and $\mathbf{y}$ having joint p.d.f. $p(\mathbf{x}, \mathbf{y})$. Here the distinction between a random variable and its realized value is omitted. Let $\mathbf{z} = \mathbf{f}(\mathbf{y})$ be a vector valued function of $\mathbf{y}$, and let $\mathbf{y} = \mathbf{f}^{-1}(\mathbf{z})$ be the inverse transformation. We wish to find the p.d.f. of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$. Assume that each of the elements of any of the vectors can take any value in the real line. The joint density of $\mathbf{x}$ and $\mathbf{z}$ is

$$p_{XZ}(\mathbf{x}, \mathbf{z}) = p_{XY}\left(\mathbf{x}, \mathbf{f}^{-1}(\mathbf{z})\right) |\mathbf{J}|$$

where the Jacobian is

$$\begin{aligned}\mathbf{J} &= \det \begin{bmatrix} \dfrac{\partial \mathbf{x}}{\partial \mathbf{x}'} & \dfrac{\partial \mathbf{x}}{\partial \mathbf{z}'} \\ \dfrac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{x}'} & \dfrac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'} \end{bmatrix} = \det \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \dfrac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'} \end{bmatrix} \\ &= \det \left[\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'} \right].\end{aligned}$$

Therefore

$$p_{XZ}(\mathbf{x}, \mathbf{z}) = p\left(\mathbf{x}, \mathbf{f}^{-1}(\mathbf{z})\right) \left| \det \left[\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'} \right] \right|.$$

Recall that

$$p_Z(\mathbf{z}) = p_Y\left(\mathbf{f}^{-1}(\mathbf{z})\right) \left| \det \left[\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'} \right] \right|.$$

To obtain $p(\mathbf{x}|\mathbf{z})$, use is made of the fact that

$$\begin{aligned} p_{X|Z}(\mathbf{x}|\mathbf{z}) &= \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z})} \\ &= \frac{p_{XY}\left(\mathbf{x},\mathbf{f}^{-1}(\mathbf{z})\right)\left|\det\left[\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'}\right]\right|}{p_Y\left(\mathbf{f}^{-1}(\mathbf{z})\right)\left|\det\left[\frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}'}\right]\right|} \\ &= p_{X|Y}\left(\mathbf{x}|\mathbf{f}^{-1}(\mathbf{z})\right). \end{aligned} \tag{2.48}$$

This is an important result: it shows that if $\mathbf{z}$ is a one-to-one transformation of $\mathbf{y}$, the conditional p.d.f. (or distribution) of $\mathbf{x}$, given $\mathbf{z}$, is the same as the conditional p.d.f. of $\mathbf{x}$, given $\mathbf{y}$. Arguing intuitively, this implies that the same inferences about parameters of this conditional distribution are drawn irrespective of whether one bases inferences on $\mathbf{x}|\mathbf{y}$ or on $\mathbf{x}|\mathbf{z}$. Suppose, in a Bayesian context, that $\mathbf{x}$ is unknown and that $\mathbf{y}$ is an observed data vector. Often, all information about $\mathbf{x}$ will be contained in a vector $\mathbf{z}$ having a lower dimension than $\mathbf{y}$ (loosely speaking, one can refer to this as a principle of sufficiency, but see Chapter 3); in this case, the posterior distribution of $\mathbf{x}$ based on $\mathbf{z}$ will lead to the same inferences about $\mathbf{x}$ than the corresponding posterior distribution based on $\mathbf{y}$, but with a considerable reduction in dimensionality. For example, in Bayesian inferences about the coefficient of correlation of a bivariate normal distribution from which n pairs have been sampled at random (so the data $\mathbf{y}$ are in a vector of order $2n \times 1$), the posterior distribution of the correlation parameter can be shown to depend on the data only through its ML estimator, which is a scalar variable (Box and Tiao, 1973; Bayarri, 1981; Bernardo and Smith, 1994).

To illustrate (2.48), suppose that phenotypic values for a certain trait y are recorded on members of nuclear families consisting of an offspring (y_0), a father (y_f), and a mother (y_m). Assume that their joint p.d.f. is the trivariate normal process

$$\begin{pmatrix} y_0 \\ y_f \\ y_m \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma_a^2/2 & \sigma_a^2/2 \\ \sigma_a^2/2 & \sigma^2 & 0 \\ \sigma_a^2/2 & 0 & \sigma^2 \end{bmatrix}\right),$$

where σ^2 is the variance of the phenotypic values, and σ_a^2 is the additive genetic variance for this trait in the population from which individuals are sampled. The independence between parental records reflects the often-made assumption that the two parents are not genetically related or mated assortatively. The distribution $[y_0|y_f, y_m]$ is also normal, with expected value,

$$E(y_0|y_f, y_m) = \mu + \frac{1}{2}h^2(y_f - \mu) + \frac{1}{2}h^2(y_m - \mu)$$

and variance

$$Var\left(y_0|y_f, y_m\right) = \sigma^2\left(1 - \frac{h^4}{2}\right),$$

where $h^2 = \sigma_a^2/\sigma^2$. Consider now the random variables

$$\hat{a}_f = h^2\left(y_f - \mu\right) \quad \text{and} \quad \hat{a}_m = h^2\left(y_m - \mu\right).$$

These are the means of the conditional distribution of the father's and mother's additive genetic values, respectively, given their phenotypic values. The inverse transformations are $y_f = h^{-2}\hat{a}_f + \mu$ and $y_m = h^{-2}\hat{a}_m + \mu$, respectively. Suppose that one wishes to derive the distribution $[y_0|\hat{a}_f, \hat{a}_m]$. According to (2.48), this distribution must be the same as $[y_0|y_f, y_m]$, where we replace, in the latter, y_f by $h^{-2}\hat{a}_f + \mu$ and y_m by $h^{-2}\hat{a}_m + \mu$. That is, $[y_0|\hat{a}_f, \hat{a}_m]$ is normal, with expected value:

$$E\left(y_0|\hat{a}_f, \hat{a}_m\right) = \mu + \frac{1}{2}\hat{a}_f + \frac{1}{2}\hat{a}_m$$

and variance as before. ■

Example 2.19 *The Box–Muller transformation*

The following technique was devised for generating standard normal random variables (Box and Muller, 1958). Let U_1 and U_2 be independent random variables having the same uniform distribution $Un\left(0, 1\right)$ and construct the transformation

$$X = f_1\left(U_1, U_2\right) = \left(-2\ln U_1\right)^{\frac{1}{2}}\cos\left(2\pi U_2\right),$$
$$Y = f_2\left(U_1, U_2\right) = \left(-2\ln U_1\right)^{\frac{1}{2}}\sin\left(2\pi U_2\right).$$

It will be shown that X and Y are i.i.d. $N\left(0, 1\right)$. The proof below makes use of the following trigonometric relationships:

$$\sin^2\left(x\right) + \cos^2\left(x\right) = 1,$$
$$\tan\left(x\right) = \frac{\sin\left(x\right)}{\cos\left(x\right)},$$
$$\tan^{-1}\left(\tan x\right) = x,$$
$$\frac{d}{dx}\tan^{-1}\left(u\right) = \frac{1}{1+u^2}\frac{du}{dx}.$$

Using these relationships, the inverse transformations are obtained as

$$U_1 = f_1^{-1}\left(X, Y\right) = \exp\left[-\frac{1}{2}\left(X^2 + Y^2\right)\right],$$
$$U_2 = f_2^{-1}\left(X, Y\right) = \left(2\pi\right)^{-1}\tan^{-1}\left(\frac{Y}{X}\right).$$

Then, the joint p.d.f. of (X, Y) is

$$p_{X,Y}(x,y) = p_{U_1U_2}\left(f_1^{-1}(X,Y), f_2^{-1}(X,Y)\right)|\mathbf{J}| = |\mathbf{J}|. \tag{2.49}$$

The Jacobian is given by

$$\begin{aligned}\mathbf{J} &= \det\begin{bmatrix} \dfrac{\partial f_1^{-1}(X,Y)}{\partial X} & \dfrac{\partial f_1^{-1}(X,Y)}{\partial Y} \\ \dfrac{\partial f_2^{-1}(X,Y)}{\partial X} & \dfrac{\partial f_2^{-1}(X,Y)}{\partial Y} \end{bmatrix} \\ &= \det\begin{bmatrix} -\exp\left[-\frac{1}{2}\left(X^2+Y^2\right)\right]X & -\exp\left[-\frac{1}{2}\left(X^2+Y^2\right)\right]Y \\ -Y\left[2\pi\left(1+\frac{Y^2}{X^2}\right)X^2\right]^{-1} & \left[2\pi\left(1+\frac{Y^2}{X^2}\right)X\right]^{-1} \end{bmatrix} \\ &= -(2\pi)^{-1}\exp\left[-\frac{1}{2}\left(X^2+Y^2\right)\right]. \end{aligned} \tag{2.50}$$

Using the absolute value of this in (2.49) gives

$$p_{X,Y}(x,y) = \left(\sqrt{2\pi}\right)^{-1}\exp\left(-\frac{X^2}{2}\right)\left(\sqrt{2\pi}\right)^{-1}\exp\left(-\frac{Y^2}{2}\right)$$

which can be recognized as the p.d.f. of two independent standard normal variables. ■

Example 2.20 *Implied distribution of beliefs about heritability*
Suppose a quantitative geneticist wishes to undertake a Bayesian analysis of the heritability $\left(h^2\right)$ of a certain trait. Let there be two sources of variance in the population, genetic and environmental, and let the values of the corresponding variance components be unknown. As pointed out earlier (and see Chapter 5), the Bayesian approach requires the specification of an uncertainty distribution, the prior distribution, which is supposed to describe beliefs about heritability before the data are observed. Let the positive random variables Y and X denote the genetic and environmental components of variance, respectively. Suppose that this geneticist uses proper uniform distributions to represent vague prior knowledge about components of variance. In addition, the geneticist assumes that prior beliefs about X are independent of those about Y, and takes as prior densities

$$\begin{aligned} p(x) &= \frac{1}{a}, \quad \text{for } 0 < x < a, \\ p(y) &= \frac{1}{b}, \quad \text{for } 0 < y < b, \end{aligned}$$

where a and b are the maximum values that X and Y, respectively, are allowed to take. These upper bounds are perhaps established on the basis

of some knowledge of the population in question, or on mechanistic considerations. Since X and Y are assumed to be independent, the joint prior density of X and Y is simply the product of the two densities above. The problem is to derive the induced prior p.d.f. of the heritability, denoted here W and defined as the ratio

$$W = \frac{Y}{Y+X}.$$

It will be shown that uniform prior distributions for each of the two variance components lead to discontinuous and perhaps sharp prior distributions of W, depending on the values of a and b adopted. In order to make a one-to-one transformation, define the auxiliary random variable $U = Y$. Thus, the transformation can be written as

$$\begin{aligned} W &= f_1(X,Y) = \frac{Y}{X+Y}, \\ U &= f_2(X,Y) = f_2(Y) = Y, \end{aligned}$$

and the inverse transformation is

$$\begin{aligned} X &= f_1^{-1}(U,W) = \frac{U(1-W)}{W}, \\ Y &= f_2^{-1}(U,W) = f_2^{-1}(U) = U. \end{aligned}$$

Then the joint p.d.f. of U and W is

$$p(u,w) = p\left(f_1^{-1}(u,w), f_2^{-1}(u)\right)|\mathbf{J}|,$$

where

$$\begin{aligned} \mathbf{J} &= \det \begin{bmatrix} \dfrac{\partial f_1^{-1}(u,w)}{\partial u} & \dfrac{\partial f_1^{-1}(u,w)}{\partial w} \\ \dfrac{\partial f_2^{-1}(u)}{\partial u} & \dfrac{\partial f_2^{-1}(u)}{\partial w} \end{bmatrix} \\ &= \det \begin{bmatrix} \frac{1-w}{w} & \frac{-u}{w^2} \\ 1 & 0 \end{bmatrix} = \frac{u}{w^2}. \end{aligned}$$

Therefore,

$$p(u,w) = \frac{u}{abw^2}. \tag{2.51}$$

The next step involves finding the support of this joint p.d.f. The relationship $U = Y$ implies

$$\text{(i) } 0 < u < b$$

and the relationship $W = Y/(Y+X)$ implies

$$\text{(ii) } 0 < w < 1.$$

Further, the fact that $X = U(1-W)/W$ implies

$$\text{(iiia)}\ 0 < \frac{u(1-w)}{w} < a,$$
$$\text{(iiib)}\ 0 < u(1-w) < wa,$$
$$\text{(iiic)}\ 0 < u < \frac{wa}{1-w}.$$

From (i) and (iiic) the following inequalities must be satisfied:

$$(u < b) \quad \text{and} \quad \left(u < \frac{wa}{1-w}\right).$$

Since a and b are given, and u must be smaller than b, one must determine the values of w such that $wa/(1-w)$ is smaller than b. We have

$$\frac{wa}{1-w} < b \Longrightarrow wa < b(1-w) \Longrightarrow w < \frac{b}{b+a}.$$

From all these relationships, the sample space of the joint distribution $[U, W]$, with density in (2.51), is

$$0 < w < 1,$$
$$0 < u < \frac{wa}{1-w}, \quad \text{for } 0 < w < \frac{b}{b+a},$$
$$0 < u < b, \quad \text{for } \frac{b}{b+a} < w < 1.$$

Hence, $p(u,w)$ is discontinuous at $b/(b+a)$. Verifying that $p(u,w)$ in (2.51) is a proper p.d.f. requires integrating the density over the range of all possible values of U and W:

$$\int\int p(u,w)\,du\,dw$$
$$= \frac{1}{ab}\int_{w=0}^{\frac{b}{a+b}}\int_{u=0}^{\frac{wa}{1-w}} \frac{u}{w^2}du\,dw + \frac{1}{ab}\int_{w=\frac{b}{a+b}}^{1}\int_{u=0}^{b} \frac{u}{w^2}du\,dw$$
$$= \frac{1}{ab}\left[\frac{ab}{2} + \frac{ab}{2}\right] = 1,$$

so propriety is established. The marginal density of heritability (W) is

$$p(w) = \begin{cases} \displaystyle\int_{u=0}^{\frac{wa}{1-w}} \frac{u}{abw^2}du, & \text{for } 0 < w < \frac{b}{a+b}, \\ \displaystyle\int_{u=0}^{b} \frac{u}{abw^2}du, & \text{for } \frac{b}{a+b} \le w < 1. \end{cases}$$

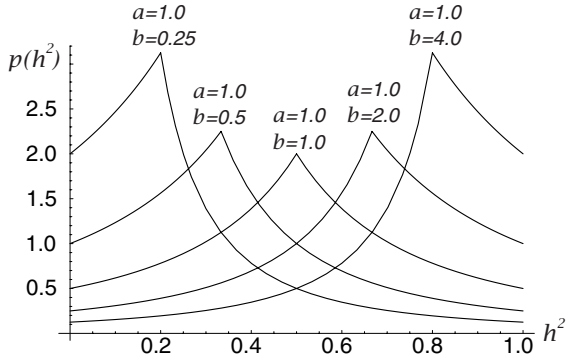

FIGURE 2.1. Prior distributions of heritability $\left(h^2 = W\right)$ implied by assuming uniform prior distributions of the variance components, for different bounds a and b.

This yields

$$p(w) = \begin{cases} \frac{a}{2b(1-w)^2}, & \text{for } 0 < w < \frac{b}{a+b}, \\ \frac{b}{2aw^2}, & \text{for } \frac{b}{a+b} \leq w < 1. \end{cases} \tag{2.52}$$

Figure 2.1 shows the implied prior distribution of heritability (W), given the assumed prior distributions for the variance components, for different values of a and b. Five different prior densities are depicted. It can be seen that independent, proper, uniform distributions, for each of the two variance components, induce a spiked prior density for heritability, with a degree of sharpness or asymmetry that depends on the bounds a and b. ■

Example 2.21 *Revisited implied distribution of beliefs about heritability*
Suppose now that the additive genetic (Y) and the environmental (X) components of variance follow independent inverse gamma (or scaled inverse chi-square) distributions, a priori. It follows immediately that their reciprocals $G = 1/Y$ and $T = 1/X$ possess independent gamma distributions. Note that heritability can be written as

$$W = \frac{Y}{X+Y} = \frac{T}{T+G}.$$

Once again, the geneticist wishes to derive the implied prior distribution of W. The prior gamma densities of T and G are

$$p(t) = \frac{b_t^{a_t}}{\Gamma(a_t)} t^{a_t - 1} \exp(-b_t t), \quad t, b_t, a_t > 0,$$

$$p(g) = \frac{b_g^{a_g}}{\Gamma(a_g)} g^{a_g - 1} \exp(-b_g g), \quad g, b_g, a_g > 0,$$

where b_t, a_t, b_g, and a_g are parameters of the appropriate distributions. In view of the independence between T and G, their joint density is

$$p(t,g) = \frac{b_t^{a_t} b_g^{a_g} t^{a_t-1} g^{a_g-1}}{\Gamma(a_t)\Gamma(a_g)} \exp(-b_t t - b_g g).$$

In order to arrive at the marginal distribution of W, make the one-to-one transformation

$$\begin{aligned} W &= f_1(T,G) = \frac{T}{T+G}, \quad 0 < W < 1, \\ U &= f_2(T,G) = T+G, \quad U > 0, \end{aligned}$$

where U is an auxiliary random variable. The inverse transformation is

$$\begin{aligned} T &= f_1^{-1}(U,W) = UW, \\ G &= f_2^{-1}(U,W) = U(1-W). \end{aligned}$$

The joint p.d.f. of U and W is

$$p_{UW}(u,w) = p_{TG}\left(f_1^{-1}(U,W), f_2^{-1}(U,W)\right)|\mathbf{J}|,$$

where

$$\begin{aligned} \mathbf{J} &= \det \begin{bmatrix} \dfrac{\partial f_1^{-1}(u,w)}{\partial u} & \dfrac{\partial f_1^{-1}(u,w)}{\partial w} \\ \dfrac{\partial f_2^{-1}(u,w)}{\partial u} & \dfrac{\partial f_2^{-1}(u,w)}{\partial w} \end{bmatrix} \\ &= \det \begin{bmatrix} w & u \\ 1-w & -u \end{bmatrix} = -u. \end{aligned}$$

Therefore

$$\begin{aligned} p_{UW}(u,w) = {} & \frac{b_t^{a_t} b_g^{a_g}}{\Gamma(a_e)\Gamma(a_g)} u (uw)^{a_t-1} [u(1-w)]^{a_g-1} \\ & \times \exp[-b_t uw - b_g u(1-w)]. \end{aligned}$$

To arrive at the desired marginal distribution, one must integrate the preceding joint density with respect to U:

$$\begin{aligned} p(w) &= \int_0^\infty p_{UW}(u,w)\,du \\ &= \frac{b_t^{a_t} b_g^{a_g}}{\Gamma(a_t)\Gamma(a_g)} w^{a_t-1} (1-w)^{a_g-1} \\ &\times \int_0^\infty u^{a_t+a_g-1} \exp[-(b_g + b_t w - b_g w) u]\,du. \end{aligned}$$

Now the integrand is the kernel of the density of a gamma distribution, so

$$\int_0^\infty u^{a_t+a_g-1}\exp\left[-\left(b_g+b_tw-b_gw\right)u\right]du$$
$$=\left[\frac{\left(b_g+b_tw-b_gw\right)^{a_t+a_g}}{\Gamma\left(a_t+a_g\right)}\right]^{-1}.$$

Hence, the prior density of heritability is

$$p\left(w\right)=\frac{\Gamma\left(a_t+a_g\right)b_t^{a_t}b_g^{a_g}}{\Gamma\left(a_t\right)\Gamma\left(a_g\right)}w^{a_t-1}\left(1-w\right)^{a_g-1}\left(b_g+b_tw-b_gw\right)^{-a_t-a_g}. \tag{2.53}$$

The corresponding distribution does not have an easily recognizable form. Consider now the special case where $b_g=b_t$; then (2.53) reduces to

$$p\left(w\right)=\frac{\Gamma\left(a_t+a_g\right)w^{a_t-1}\left(1-w\right)^{a_g-1}}{\Gamma\left(a_t\right)\Gamma\left(a_g\right)}. \tag{2.54}$$

This is the density of a beta distribution with parameters a_t, a_g. In the hypothetical Bayesian analysis of variance components, the parameters of the gamma distribution assigned to the reciprocal of the variance components, would be equal to those of the inverse gamma (or scaled inverse chi-square) process assigned to the variance components. A typical specification would be $a_t=\nu_t/2$, $a_g=\nu_g/2$, $b_t=\nu_tS_t/2$, and $b_g=\nu_gS_g/2$. Here, ν and S are parameters of the scaled inverse chi-square distributions associated with the variance components. The implied prior density of heritability would be the beta form given by (2.54), provided that $\nu_tS_t=\nu_gS_g$; otherwise, the implied prior density of heritability would be as in (2.53). ■

2.3.1 Linear Transformations

A special case is when the transformation is linear, as in (2.43). Let $\mathbf{x}$ be a random vector possessing p.d.f. $p\left(\mathbf{x}\right)$, and let $\mathbf{y}=f\left(\mathbf{x}\right)=\mathbf{A}\mathbf{x}$ be a one-to-one linear transformation, so that the matrix of constants $\mathbf{A}$ is nonsingular. The inverse transformation is, therefore

$$\mathbf{x}=f^{-1}\left(\mathbf{y}\right)=\mathbf{A}^{-1}\mathbf{y}.$$

From (2.31), the p.d.f. of $\mathbf{Y}$ is given by

$$p_Y\left(\mathbf{y}\right)=p_X\left(f^{-1}\left(\mathbf{y}\right)\right)\left|\mathbf{J}\right|$$
$$=p_X\left(f^{-1}\left(\mathbf{y}\right)\right)\left|\det\left(\frac{\partial f^{-1}\left(\mathbf{y}\right)}{\partial\mathbf{y}'}\right)\right|.$$

Now, the matrix of derivatives is

$$\frac{\partial f^{-1}\left(\mathbf{y}\right)}{\partial\mathbf{y}'}=\frac{\partial\mathbf{A}^{-1}\mathbf{y}}{\partial\mathbf{y}'}=\mathbf{A}^{-1}.$$

Hence

$$\begin{aligned} p_Y(\mathbf{y}) &= p_X\left(f^{-1}(\mathbf{y})\right)\left|\det\left(\mathbf{A}^{-1}\right)\right| \\ &= p_X\left(f^{-1}(\mathbf{y})\right)\left|\frac{1}{\det(\mathbf{A})}\right|. \end{aligned} \tag{2.55}$$

Example 2.22 *Samples from the multivariate normal distribution*
Imagine one wishes to obtain realizations from $\mathbf{y} \sim N(\mathbf{m}, \mathbf{V})$. The starting point consists of drawing a vector of independent standard normal deviates, $\mathbf{x}$, from a $N(\mathbf{0}, \mathbf{I})$ distribution having the same order (n, say) as $\mathbf{y}$. Then write $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$, so

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2}\right).$$

Because $\mathbf{V}$ is a nonsingular variance–covariance matrix and, therefore, positive definite, it can be expressed as $\mathbf{V} = \mathbf{L}'\mathbf{L}$, where $\mathbf{L}'$ is a nonsingular, lower triangular matrix; this is called the Cholesky decomposition. Then $\mathbf{V}^{-1} = \mathbf{L}^{-1}(\mathbf{L}')^{-1}$ and, further, the following relationships can be established

$$\left|\mathbf{V}^{-1}\right| = \left|\mathbf{L}^{-1}\right|^2 \Rightarrow \left|\mathbf{L}^{-1}\right| = |\mathbf{V}|^{-\frac{1}{2}}.$$

Now, the linear transformation

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{m} + \mathbf{L}'\mathbf{x}$$

has the desired distribution, by being a linear combination of the normal vector $\mathbf{x}$. To verify this, note that the inverse transformation is

$$\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y}) = (\mathbf{L}')^{-1}(\mathbf{y} - \mathbf{m})$$

with the absolute value of the Jacobian of the transformation being

$$\left|\det\left(\frac{\partial \mathbf{f}^{-1}(\mathbf{y})}{\partial \mathbf{y}'}\right)\right| = \left|\det(\mathbf{L}')^{-1}\right| = \left|\det(\mathbf{V})^{-\frac{1}{2}}\right|.$$

In the usual notation employed for the Gaussian model, we write

$$\det(\mathbf{V})^{-\frac{1}{2}} = |\mathbf{V}|^{-\frac{1}{2}}.$$

Then, applying (2.55), and recalling that the initial distribution is that of n independent standard normal variables, gives, as p.d.f. of $\mathbf{Y}$,

$$p(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{m})' \mathbf{L}^{-1}(\mathbf{L}')^{-1}(\mathbf{y} - \mathbf{m})\right]$$

which proves the result. ■

Example 2.23 *Bivariate normal distribution with null means*
Consider the centered (null means) bivariate normal density function:

$$p(x,y) = C \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - \frac{2\rho xy}{\sigma_X \sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)\right], \tag{2.56}$$

where $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$, ρ is the coefficient of correlation between X and Y, and C is the integration constant. Hereinafter, we will retain only the kernel of this density. If $\rho = 0$, the joint density can be factorized as

$$p(x,y) = p(x)\, p(y),$$

where

$$p(x) \propto \exp\left(-\frac{x^2}{2\sigma_X^2}\right), \quad p(y) \propto \exp\left(-\frac{y^2}{2\sigma_Y^2}\right).$$

This verifies that a null value of the correlation is a sufficient condition for independence in the bivariate normal distribution (as discussed in Chapter 1, a sufficient condition for mutual independence in the multivariate distribution is that the variance–covariance matrix has a diagonal form).
Returning to the general case of nonnull correlation, now let U and V be random variables arising through the linear transformation of X and Y

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \frac{X}{\sigma_x} - \frac{Y}{\sigma_y} \\ \frac{X}{\sigma_x} + \frac{Y}{\sigma_y} \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_x^{-1} & -\sigma_y^{-1} \\ \sigma_x^{-1} & \sigma_y^{-1} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \mathbf{A} \begin{bmatrix} X \\ Y \end{bmatrix},$$

where $\mathbf{A}$ is the coefficient matrix preceding the $[X,Y]'$ vector. Hence, U is a contrast between the standardized X and Y variables, whereas V is their sum. Since U and V are linear combinations of normal random variables, they must also follow a joint bivariate normal distribution, and are also normal at the margin. It can be verified that the covariance between U and V is equal to 0, so these two variates are independent. We proceed to verify this formally. The inverse transformation is

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \sigma_x^{-1} & -\sigma_y^{-1} \\ \sigma_x^{-1} & \sigma_y^{-1} \end{bmatrix}^{-1} \begin{bmatrix} U \\ V \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sigma_x & \sigma_x \\ -\sigma_y & \sigma_y \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix}.$$

Therefore,

$$X = f_1^{-1}(U,V) = \frac{\sigma_x (U+V)}{2},$$
$$Y = f_2^{-1}(U,V) = \frac{\sigma_y (V-U)}{2}.$$

The absolute value of the Jacobian of the transformation is

$$|\mathbf{J}| = \left| \frac{1}{\det \mathbf{A}} \right| = \frac{\sigma_x \sigma_y}{2}.$$

Using (2.55) the joint p.d.f. of U and V is

$$\begin{aligned} p(u,v) &= p\left(f_1^{-1}(u,v), f_2^{-1}(u,v)\right) \frac{\sigma_x \sigma_y}{2} \\ &\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{\left(u^2+v^2+2uv\right)}{4} - \frac{2\rho\left(v^2-u^2\right)}{4} \right.\right. \\ &\qquad \left.\left. + \frac{\left(u^2+v^2-2uv\right)}{4} \right] \right\} \\ &\propto \exp\left[-\frac{1}{4(1-\rho^2)} \left(u^2+v^2-\rho\left(v^2-u^2\right)\right) \right] \end{aligned}$$

which can be factorized as

$$\begin{aligned} p(u,v) &= C \exp\left[-\frac{1}{4(1-\rho^2)} u^2 (1+\rho) \right] \\ &\quad \times \exp\left[-\frac{1}{4(1-\rho^2)} v^2 (1-\rho) \right]. \end{aligned}$$

Hence,

$$p(u,v) \propto p(u)\, p(v)$$

which shows that U and V are independent. ■

2.3.2 Approximating the Mean and Covariance Matrix

Let $\mathbf{x}$ be a random vector having mean $\mathbf{m}$ and covariance matrix $\mathbf{V}$, and suppose one is interested in a scalar valued function, $Y = f(\mathbf{x})$, of the vector $\mathbf{x}$. Assume that this function admits at least up to second-order partial differentiability with respect to $\mathbf{x}$. Put

$$\left[\frac{\partial}{\partial \mathbf{x}'}, f(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{m}} \right] = \mathbf{b}'$$

and

$$\left[\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'} f(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{m}} \right] = \mathbf{B}.$$

Expanding $f(\mathbf{x})$ in a second-order Taylor series about $\mathbf{m}$ gives

$$\begin{aligned} f(\mathbf{x}) &\cong f(\mathbf{m}) + \mathbf{b}'(\mathbf{x}-\mathbf{m}) \\ &+ \frac{1}{2}(\mathbf{x}-\mathbf{m})' \mathbf{B} (\mathbf{x}-\mathbf{m}). \end{aligned} \tag{2.57}$$

Taking expectations and variances, one obtains the following approximations to $E(Y)$ and $Var(Y)$:

(1) First order:

$$E(Y) = E[f(\mathbf{x})] \cong f[\mathbf{m}], \tag{2.58}$$

$$Var(Y) = Var[f(\mathbf{x})] \cong \mathbf{b}'\mathbf{V}\mathbf{b}. \tag{2.59}$$

(2) Second order:

$$E(Y) = E[f(\mathbf{x})] \cong f(\mathbf{m}) + \frac{1}{2} tr[\mathbf{B}\mathbf{V}], \tag{2.60}$$

$$Var(Y) = Var[f(\mathbf{x})] \cong \mathbf{b}'\mathbf{V}\mathbf{b}$$
$$+\frac{1}{4} Var[tr(\mathbf{S}_{\mathbf{x}}\mathbf{B})] + Cov\left[\mathbf{b}'(\mathbf{x}-\mathbf{m}), (\mathbf{x}-\mathbf{m})'\mathbf{B}(\mathbf{x}-\mathbf{m})\right], \tag{2.61}$$

where

$$\mathbf{S}_{\mathbf{x}} = (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'$$

is a matrix. The feasibility of evaluating the variance of the quadratic form in (2.61) depends on the distribution of the vector $\mathbf{x}$ (Searle, 1971; Rao, 1973). For example, if its distribution is normal, the covariance term vanishes because the third moments from the mean are null.

Example 2.24 *Approximate mean and variance of a ratio*

Let $\mathbf{x}' = [x_1, x_2]$, $E(\mathbf{x})' = \mathbf{m}' = [m_1, m_2]$, and

$$Var(\mathbf{x}) = \mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Consider the function

$$Y = f(\mathbf{x}) = \frac{X_1}{X_2}.$$

Then, from (2.58), the linear approximation gives as mean

$$E(Y) = E\left(\frac{X_1}{X_2}\right) \cong \frac{m_1}{m_2}.$$

The vector of first derivatives is

$$\mathbf{b} = \left.\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{m}} = \begin{bmatrix} \frac{\partial Y}{\partial X_1} \\ \frac{\partial Y}{\partial X_2} \end{bmatrix}_{\substack{X_1=m_1 \\ X_2=m_2}} = \begin{bmatrix} \frac{1}{m_2} \\ -\frac{m_1}{m_2^2} \end{bmatrix}.$$

Then, from (2.59)

$$\left[\frac{\partial}{\partial \mathbf{X}'} f(\mathbf{X})\Big|_{x=\mathbf{m}}\right] \mathbf{V} \left[\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})\Big|_{\mathbf{X}=\mathbf{m}}\right]$$
$$= \begin{bmatrix} \frac{1}{m_2} & -\frac{m_1}{m_2^2} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \frac{1}{m_2} \\ -\frac{m_1}{m_2^2} \end{bmatrix}$$

and, finally,

$$Var\left(\frac{X_1}{X_2}\right) \cong \left(\frac{m_1}{m_2}\right)^2 \left(\frac{\sigma_1^2}{m_1^2} + \frac{\sigma_2^2}{m_2^2} - 2\frac{\sigma_{12}}{m_1 m_2}\right)$$
$$= \left(\frac{m_1}{m_2}\right)^2 \left(C_1^2 + C_2^2 - 2\rho C_1 C_2\right),$$

where ρ is the coefficient of correlation and C_1 and C_2 are the corresponding coefficients of variation. ■

At this point, we have completed a review of the basic elements of distribution theory needed by geneticists to begin the study of parametric methods of inference. In the next chapters, a discussion of two important methods for drawing inferences is presented. Chapters 3 and 4 give a description of the principles of likelihood-based inference, whereas Chapters 5, 6, 7, and 8 present an overview of the Bayesian approach to statistical analysis.

Part II

Methods of Inference

This page intentionally left blank

3

An Introduction to Likelihood Inference

3.1 Introduction

A well-established problem in statistics, especially in what today is called the classical approach, is the estimation of a single parameter or an ensemble of parameters from a set of observations. Data are used to make statements about a statistical model proposed to describe aspects of the state of nature. This model is characterized in terms of parameters that have a "true" value. The description of uncertainty, typically, is in terms of a distribution that assigns probabilities or densities to the values that all random variables in the model, including the observations, can take. In general, the values of the parameters of this distribution are unknown and must be inferred from observable quantities or data. In genetics, the observables can consist of phenotypic measurements for quantitative or discrete traits, and/or information on molecular polymorphisms.

In this chapter, an important method of statistical inference called estimation by maximum likelihood (ML) is discussed. There is an extensive literature on this method, originally due to Fisher (1922) and now firmly entrenched in statistics. A historical account can be found in Edwards (1974). A readable introduction is given by King (1989), and a discussion of related concepts, from the point of view of animal breeding, is in Blasco (2001). Edwards (1992), one of the strong adherents to the use of likelihood inference in statistics, starts his book as follows:

> "Likelihood is the central concept in statistical inference. Not only does it lead to inferential techniques in its own right,

> but it is as fundamental to the repeated-sampling theories of estimation advanced by the "classical" statisticians as it is to the probabilistic reasoning advanced by the Bayesian. It is the key to an understanding of fiducial probability, and the concept without which no serious discussion of the philosophy of statistical inference can begin."

Although likelihood is indeed a central concept in statistical inference, it is not free from controversy. In fact, all methods of inference are the subject of some form of controversy. Malécot (1947) gives an interesting critique of approaches based on likelihood, and Bernardo and Smith (1994) discuss it from a Bayesian perspective, another approach to inference in which parameters are viewed as random variables, with the randomness stemming from subjective uncertainty. It will be shown in Chapter 5 how likelihood enters into the Bayesian paradigm.

This chapter is organized as follows. The likelihood function and the ML estimator are presented, including a measure of the information about the parameters contained in the data. First-order asymptotic theory of likelihood inference is covered subsequently in some detail. This gives the basis for the appeal of the method, at least from a frequentist point of view. The chapter ends with a discussion of the functional invariance of the ML estimator.

3.2 The Likelihood Function

Assume that the observed data $\mathbf{y}$ (scalar, vector, or matrix) is the outcome of a stochastic model (i.e., a random process), that can be characterized in terms of a p.d.f. $p(\mathbf{y}|\boldsymbol{\theta})$ indexed by a parameter(s) $\boldsymbol{\theta}$ taking values in the interior of a parameter space $\boldsymbol{\Omega}$. (Hereinafter, unless it is clear from the context, we use boldface for the parameters, to allow for the possibility that θ may be a vector). For example, in a bivariate normal distribution, the parameter space of the correlation coefficient includes all real numbers from -1 to 1. This gives an illustration of a bounded parameter space. On the other hand, in a regression model with p coefficients, each of which can take any value in the real line, the parameter space is the p-dimensional hypervolume $\mathbb{R}^p$. There are situations where the parameter space may be constrained, due to restrictions imposed by the model, or due to mechanistic considerations. It was seen previously that, in a random effects model where the total variation is partitioned into between and within half-sib family components, purely additive genetic action imposes the constraint that the variance within families must be at least three times as large as the variance between families. Without this constraint, the statistical model would not be consistent with the biological system it is supposed to describe. Hence, taking these constraints into account is important in

likelihood-based inference. This is because one is interested in finding the maximum values the parameters can take inside their allowable space, given the data, so constrained maximization techniques must be employed.

Given the probability model and the parameter $\boldsymbol{\theta}$, the joint probability density (or distribution, if the random variable under study is discrete) of the observations, $p(\mathbf{y}|\boldsymbol{\theta})$, is a function of $\mathbf{y}$. This describes the plausibility of the different values $\mathbf{y}$ can take in its sampling space, at a given value of $\boldsymbol{\theta}$. The likelihood function or, just likelihood, is based on an "inversion" of the preceding concept. By definition, the likelihood is any function of $\boldsymbol{\theta}$ that is proportional to $p(\mathbf{y}|\boldsymbol{\theta})$; it is denoted as $L(\boldsymbol{\theta}|\mathbf{y})$ or $L(\boldsymbol{\theta})$. Thus, the likelihood is a mathematical function of the parameter for fixed data, whereas the p.d.f. is viewed as varying with $\mathbf{y}$ at fixed values of $\boldsymbol{\theta}$. Therefore, the likelihood is not a probability density or a probability function, so the different values $\boldsymbol{\theta}$ takes in the likelihood cannot be interpreted in the usual probabilistic sense. Further, because the true value of a parameter is fixed, one cannot apply the probability calculus to a likelihood function, at least in principle. Blasco (2001) pointed out that Fisher proposed the likelihood function as a rational measure of degree of belief but without sharing the properties of probability. The adherents of the likelihood approach to inference view the entire likelihood as a complete description of the information about $\boldsymbol{\theta}$ contained in the data, given the model.

Now, by definition

$$L(\boldsymbol{\theta}|\mathbf{y}) = k(\mathbf{y})p(\mathbf{y}|\boldsymbol{\theta}) \propto p(\mathbf{y}|\boldsymbol{\theta}), \tag{3.1}$$

where $k(\mathbf{y})$ is a function that does not depend on $\boldsymbol{\theta}$, but may depend on the data. For $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the value $L(\boldsymbol{\theta}^*|\mathbf{y})$ is called the likelihood of $\boldsymbol{\theta}^*$. It is apparent that a likelihood, by construction, must be positive, because any density (or probability) is positive for any $\boldsymbol{\theta}$ in the allowable space. While a probability takes values between 0 and 1, a likelihood evaluated at a given point has no specific meaning. On the other hand, it is meaningful to compare the ratio of likelihoods from the same data. To be more specific, consider a one-to-one transformation $f(\mathbf{y}) = \mathbf{z}$. For example, the transformation could represent different scales of measurement associated with $\mathbf{y}$ and $\mathbf{z}$, centimeters and meters respectively, say. Then the likelihood function based on $\mathbf{z}$ is

$$L(\boldsymbol{\theta}|\mathbf{z}) = L(\boldsymbol{\theta}|\mathbf{y}) \left| \frac{\partial \mathbf{y}}{\partial \mathbf{z}} \right|.$$

Then for two possible values of the parameter, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{**}$, the likelihood ratio is the relevant quantity to study

$$\frac{L(\boldsymbol{\theta}^*|\mathbf{y})}{L(\boldsymbol{\theta}^{**}|\mathbf{y})},$$

as opposed to the difference

$$[L(\boldsymbol{\theta}^*|\mathbf{y}) - L(\boldsymbol{\theta}^{**}|\mathbf{y})] \left|\frac{\partial \mathbf{y}}{\partial \mathbf{z}}\right|.$$

The latter is arbitrarily affected by the particular transformation used, in the above example, by the arbitrary scale of measurement.

The ratio of the likelihoods can be interpreted as a measure of support brought up by the data set for one value of $\boldsymbol{\theta}$ relative to the other. It must be emphasized that likelihood values obtained from different data sets cannot be compared.

3.3 The Maximum Likelihood Estimator

Suppose for the moment that the random variable Y is discrete taking values y with probabilities depending on a parameter $\boldsymbol{\theta}$:

$$\Pr(Y = y|\boldsymbol{\theta}) = p(y|\boldsymbol{\theta}). \tag{3.2}$$

We said that the likelihood of $\boldsymbol{\theta}$, $L(\boldsymbol{\theta}|y)$, is proportional to (3.2), and the value of $\boldsymbol{\theta}$ that maximizes the likelihood $L(\boldsymbol{\theta}|y)$ is the ML estimate of $\boldsymbol{\theta}$; it is denoted by $\widehat{\boldsymbol{\theta}}$. The ML estimate $\widehat{\boldsymbol{\theta}}$ can be viewed as the most likely value of $\boldsymbol{\theta}$ given the data, in the following sense. If $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two possible values for $\boldsymbol{\theta}$, and if $\Pr(Y = y|\boldsymbol{\theta}_1) > \Pr(Y = y|\boldsymbol{\theta}_2)$, then the probability of observing what was actually observed, i.e., $Y = y$, is greater when the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_1$. Therefore it is more likely that $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ than $\boldsymbol{\theta} = \boldsymbol{\theta}_2$. The ML estimate provides the best explanation for observing the data point y, under the probability model (3.2). This does not mean that $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximizes the probability of observing the data y. This role is played by the true value of the parameter $\boldsymbol{\theta}$, in the sense discussed in Subsection 3.7.1.

The above interpretation can be adapted to the case where Y is continuously distributed, provided we view $p(y|\boldsymbol{\theta})$ as the probability that Y takes values in a small set containing y.

Often, rather than working with the likelihood it is more convenient to work with the logarithm of the likelihood function. This log-likelihood is denoted $l(\boldsymbol{\theta}|\mathbf{y})$. The maximizer of the likelihood is also the maximizer of the log-likelihood. The value of the ML estimate of $\boldsymbol{\theta}$ must be inside the parameter space and must be a global maximum, in the sense that any other value of the parameter would produce a smaller likelihood.

Fisher (1922) suggested that it would be meaningful to rescale all likelihood values relative to the maximum value it can take for a specific data set. For example, if the maximizer of $L(\boldsymbol{\theta}|\mathbf{y})$ is $\widehat{\boldsymbol{\theta}}$, one could rescale values as

$$r(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{y})}, \tag{3.3}$$

so that rescaled values are between 0 and 1.

Beyond providing a means of viewing uncertainty in terms of a relative fall in likelihood, it turns out that the maximizer of the likelihood function plays an important role in statistical inference. If $\widehat{\boldsymbol{\theta}}$ is used as a point estimator of $\boldsymbol{\theta}$, conceptual repeated sampling over the distribution of $\mathbf{y}$ generates a distribution of $\widehat{\boldsymbol{\theta}}$ values, one corresponding to each realization of $\mathbf{y}$. In fact, the sampling distribution of $\widehat{\boldsymbol{\theta}}$ has interesting coverage probabilities in relation to the true value of $\boldsymbol{\theta}$. This distribution will be discussed in a later section.

If the p.d.f. is everywhere continuous and there are no corner solutions, the ML estimator, if it exists, can be found by differentiating the log-likelihood with respect to $\boldsymbol{\theta}$, setting all partial derivatives equal to zero, and solving the resulting equations for $\boldsymbol{\theta}$. It is a joint maximization with respect to all elements in $\boldsymbol{\theta}$ that must be achieved. If this parameter has p elements, then there are p simultaneous equations to be solved, one for each of its components. The vector of first-order partial derivatives of the log-likelihood with respect to each of the elements of $\boldsymbol{\theta}$ is called the gradient or score, and is often denoted as $S\left(\boldsymbol{\theta}|\mathbf{y}\right)$ or as $\mathbf{l}'\left(\boldsymbol{\theta}|\mathbf{y}\right)$.

There are some potential difficulties inherent to inferences based on likelihood. First, there is the issue of constructing the likelihood. For example, consider generalized mixed effects linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). These are models that can include several sets of random effects ($\mathbf{u}$) and where the response variable ($\mathbf{y}$) may not be normal. The starting point in building such models is the specification of the marginal distribution of $\mathbf{u}$, followed by an assumption about the form of the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$. This produces the joint distribution $\left[\mathbf{y},\mathbf{u}|\boldsymbol{\theta}\right]$. In order to form the likelihood as in (3.1), one must obtain the marginal p.d.f. of the data. This requires evaluating the integral:

$$L(\boldsymbol{\theta}|\mathbf{y}) \propto p\left(\mathbf{y}|\theta\right) = \int p\left(\mathbf{y}|\mathbf{u},\boldsymbol{\theta}\right) p\left(\mathbf{u}|\boldsymbol{\theta}\right) d\mathbf{u}, \tag{3.4}$$

where $p\left(\mathbf{y}|\mathbf{u},\boldsymbol{\theta}\right)$ is the density of the conditional distribution of the data given the random effects, and $p\left(\mathbf{u}|\boldsymbol{\theta}\right)$ is the marginal density of the random effects. The integration in (3.4) seldom leads to a likelihood function that is expressible in closed form. This creates difficulties in evaluating likelihoods and in finding the ML estimator of $\boldsymbol{\theta}$.

A second issue is that it is often not straightforward to locate the global maximum of a likelihood, especially in high-dimensional models (those having many parameters). This is particularly troublesome in animal breeding because of the potentially large number of parameters a model can have. For example, consider a 17-variate analysis of type traits in dairy cattle, such as scores on body condition and feet and legs, and suppose that the same Gaussian linear model with two random factors is entertained for each trait. These factors could be the breeding value of a cow plus a random residual.

Here there are two covariance matrices, each having 153 potentially distinct elements to be estimated, unless these matrices are structured as a function of a smaller number of parameters. If, additionally, the number of location parameters or fixed effects is f, the order of $\boldsymbol{\theta}$ would be $f + 306$. Typically, the order of f is in the thousands when field records are employed in the analysis. In this situation, encountered often in practice, it would be extremely difficult to verify that a global maximum has been found. The zeros of the first derivatives only locate extreme points in the interior of the domain of a function. If extrema occur on the boundaries or corners, the first derivative may not be zero at that point. First derivatives can also be null at local minima or maxima, or at inflection points. To ensure that a maximum (local or global) has been found, the second derivative of the log-likelihood with respect to the parameter must be negative, and this is relatively easy to verify in a single parameter model. If $\boldsymbol{\theta}$ is a vector, the conditions for a maximum are:
(i) The vector of partial derivatives of $l(\boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$ must be null.
(ii) The symmetric matrix of mixed second-order partial derivatives of $l(\boldsymbol{\theta}|\mathbf{y})$ with respect to all parameters:

$$\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \tag{3.5}$$

must be negative definite. A symmetric matrix is negative definite if all its eigenvalues are negative. In the dairy cattle example given above, $f + 306$ eigenvalues would need to be computed to check this. Further, even if all eigenroots of (3.5) are negative, there is no assurance that the stationary point would be a global maximum.

The form of a likelihood function depends on the sampling model hypothesized for the observations. Maximum likelihood is a parametric method, so the distributional assumptions made are central. However, there is a certain automatism in the calculations required: find first and second derivatives; if iterative methods are employed, iterate until convergence; verify that a stationary point has been reached and attempt to ascertain that it is a supremum. However, the chores of finding the ML estimator of $\boldsymbol{\theta}$, in a model where normality is assumed, differ from those in a model where, say, Student-t distributions with unknown degrees of freedom are used.

Another important issue is the potential susceptibility of the ML estimator of $\boldsymbol{\theta}$ to small changes in the data. A slightly different sample of observations can produce very different estimates when the likelihood is flat in the neighborhood of the maximum. In theory, it is advisable to explore the likelihood as much as possible. However this can be complicated, if not impossible, in multidimensional models. In the dairy cattle breeding example, how would one represent a likelihood in $f + 306$ dimensions?

One of the main difficulties of ML is encountered in multiparameter situations. The problem is caused by the existence of nuisance parame-

ters. For example, suppose that the multivariate normal sampling model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ is proposed, where $\boldsymbol{\beta}$ is an unknown location vector to be estimated, and $\mathbf{V}$ is a variance–covariance matrix, also unknown. This matrix is needed for a correct specification of the model but it may not be of primary interest. The ML procedure, at least in its original form, treats all parameters identically, and does not account well for the possible loss of information incurred in estimating $\mathbf{V}$. An approach for dealing with this shortcoming, at least in some models, will be discussed later.

A related problem is that of inferences about random effects in linear models (Henderson, 1963, 1973; Searle et al., 1992). In the notation employed in connection with (3.4), suppose that $\mathbf{u}$ is a vector of genetic effects, and that one wishes to make statements about the conditional distribution $[\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}]$, with unknown $\boldsymbol{\theta}$ (e.g., $\boldsymbol{\theta}$ may include $\boldsymbol{\beta}$ and $\mathbf{V}$). Intuitively, one may wish to use the approximation $[\mathbf{u}|\mathbf{y}, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}]$, where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$. However, this does not take into account the fact that there is an error of estimation associated with $\widehat{\boldsymbol{\theta}}$, and it is not obvious why one should use $\widehat{\boldsymbol{\theta}}$ as opposed to any other point estimate of $\boldsymbol{\theta}$. Arguably, likelihood inference was developed for assessing the plausibility of values of $\boldsymbol{\theta}$, and not for making statements about unobservable random variables from conditional distributions.

3.4 Likelihood Inference in a Gaussian Model

Suppose a single data point $y = 10$ is drawn from a normal distribution with unknown mean m and known variance $\sigma^2 = 3$. The likelihood resulting from this single observation follows from the corresponding normal p.d.f.

$$L\left(m|\sigma^2 = 3, y = 10\right) = \frac{1}{\sqrt{6\pi}} \exp\left[-\frac{(10-m)^2}{6}\right]. \tag{3.6}$$

The function $k(y)$ in (3.1) is $1/\sqrt{6\pi}$, and the only part that matters (from a likelihood inference viewpoint) is the exponential function in (3.6). This is interpreted as a function of m for a given value of σ^2 and of the observed data point $y = 10$. A plot of the likelihood function using $k(y) = 1/\sqrt{6\pi}$ is in Figure 3.1.

The likelihood is symmetric about $m = 10$, its maximizer, as shown in Figure 3.1. Likelihoods are not probability functions or density functions and do not necessarily yield a finite value (e.g., 1) when integrated with respect to $\boldsymbol{\theta}$. In this example, however, if (3.6) is integrated with respect to m (which would be meaningless in a likelihood setting because m is not stochastic), the value of the integral is precisely equal to one. At any rate, the main issues in likelihood inference are the shape of the function and the relative heights.

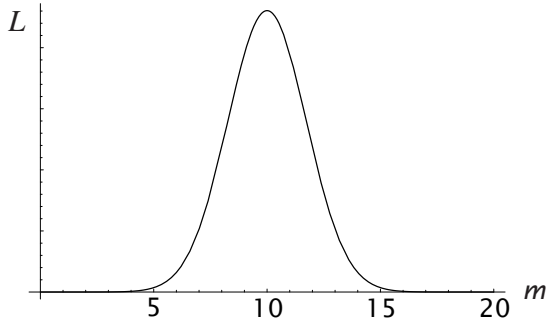

FIGURE 3.1. Likelihood of m based on a sample of size 1.

Suppose now that four additional independent random samples are drawn from the same distribution and that the new data set consists of the 5×1 vector

$$\mathbf{y}' = [y_1 = 10, y_2 = 8, y_3 = 12, y_4 = 9, y_5 = 11].$$

The likelihood is now built from the joint p.d.f. of the observations. This, by virtue of the independence assumptions, is the product of five normal densities each with mean m and variance $\sigma^2 = 3$. For this example, the likelihood of m is

$$L\left(m|\sigma^2 = 3, \mathbf{y}\right) = \frac{1}{\left(\sqrt{6\pi}\right)^5} \exp\left[-\frac{(10-m)^2 + \cdots + (11-m)^2}{6}\right]. \quad (3.7)$$

This can be written as

$$\begin{aligned} L\left(m|\sigma^2 = 3, \mathbf{y}\right) &= \left(\sqrt{6\pi}\right)^{-5} \exp\left[-\frac{(10-10)^2 + \cdots + (11-10)^2}{6}\right] \\ &\quad \times \exp\left[-\frac{5\,(10-m)^2}{6}\right] \\ &\propto k\,(\mathbf{y}) \exp\left[-\frac{5\,(10-m)^2}{6}\right], \end{aligned}$$

where $k(\mathbf{y}) = \left(\sqrt{6\pi}\right)^{-5} \exp\left\{-[(10-10)^2 + \cdots + (11-10)^2]/6\right\}$. The product of the two $\exp(\cdot)$ functions arises because the term in the exponent in (3.7) can be written as

$$\sum_{i=1}^{5} (y_i - m)^2 = \sum_{i=1}^{5} (y_i - \overline{y})^2 + 5\,(\overline{y} - m)^2, \quad (3.8)$$

where

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{5}.$$

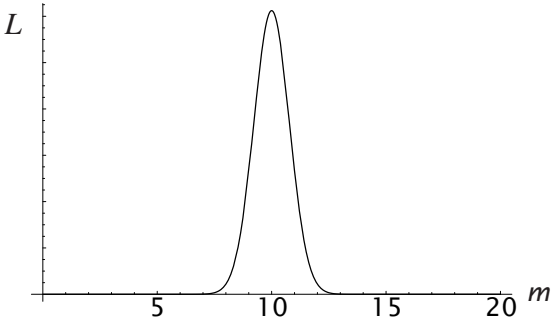

FIGURE 3.2. Likelihood of m based on a sample of size 5.

The relevant part is that containing m. It follows that

$$L\left(m|\sigma^2=3,\mathbf{y}\right)\propto\exp\left[-\frac{5\left(\overline{y}-m\right)^2}{6}\right]. \tag{3.9}$$

The relative values of the likelihood are the same irrespective of whether (3.7) or (3.9) are used. The plot of the likelihood is shown in Figure 3.2.

Note that the likelihood is much sharper than that presented in Figure 3.1. The curvature of the likelihood at its maximum is larger with the larger sample. This is because the likelihood of m based on a sample of size 5 has more information about m than likelihood (3.6), a concept to be defined formally later in this chapter. In both data sets, the maximum value of the likelihood is obtained when $m=10$, so the ML estimate of this parameter is $\widehat{m}=10$ in the two situations. In practice, this would rarely happen, but we have chosen the situation deliberately to compare the sharpness of two likelihood functions having the same maximum. For example, if the observed value in the first data set had been $y=7.6$, this would have been the corresponding ML estimate of m.

With a sample of n independent observations, it follows from (3.8) that the log-likelihood can be expressed as

$$l\left(m|\sigma^2,\mathbf{y}\right)=\text{constant}-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2+n\left(\overline{y}-m\right)^2\right]. \tag{3.10}$$

Maximizing (3.10) with respect to m is equivalent to minimizing the sum of squared deviations $\sum_{i=1}^{n}\left(y_i-m\right)^2$, so there is a relationship with the least-squares criterion in this case. Taking the derivative of (3.10) with respect to m gives the linear form in m :

$$\frac{dl\left(m|\sigma^2,\mathbf{y}\right)}{dm}=\frac{n\left(\overline{y}-m\right)}{\sigma^2}. \tag{3.11}$$

Setting (3.11) to zero and solving for m gives, as ML of m,

$$\widehat{m} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}. \tag{3.12}$$

Consequently, the ML estimator of the expected value of a normal distribution with known variance is the arithmetic mean of the observations (the estimator would have been the same if the variance had been unknown, as shown later). It is also the least-squares estimator, but this coincides with the ML estimator of m only when normality is assumed, as in this example. Differentiating (3.11) with respect to m gives $-n/\sigma^2$. Because this derivative is negative, it follows that $\widehat{m}$ is a maximum.

Suppose, temporarily, that m is a random variable in the real line, and that σ^2 is an unknown parameter. Now integrate (3.7) with respect to m to obtain, for a sample of size n,

$$\int L\left(m|\sigma^2, \mathbf{y}\right) dm$$
$$= \frac{1}{\left(2\pi\sigma^2\right)^{(n-1)/2}\sqrt{n}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right]. \tag{3.13}$$

which, given the data, is a function of σ^2 only. (The limits of integration are $\pm\infty$, but this is generally omitted unless the context dictates otherwise). Expression (3.13) is called an integrated or marginal likelihood, and has the same form as the so-called restricted likelihood, proposed by Patterson and Thompson (1971). In fact, anticipating a subject to be discussed later in this book, when σ^2 is an unknown parameter, maximization of (3.13) with respect to σ^2 yields

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{n-1},$$

which is the restricted maximum likelihood estimator (REML) of σ^2.

3.5 Fisher's Information Measure

3.5.1 Single Parameter Case

Suppose the random vector $\mathbf{y}$ has a distribution indexed by a single parameter θ, and let the corresponding density function be $p\left(\mathbf{y}|\theta\right)$. Fisher's information (also known as Fisher's expected information, or expected information) about θ, represented as $I(\theta)$, is defined to be

$$I(\theta) = E_{\mathbf{y}}\left[\left(\frac{dl}{d\theta}\right)^2\right], \tag{3.14}$$

where the notation emphasizes that expectations are taken with respect to the marginal distribution of $\mathbf{y}$. Hereinafter, $l = l(\theta) = l(\theta|\mathbf{y})$ will be used indistinctly for the log-likelihood of θ. This measure of information must be positive because it involves an average of positive random variables, the square of the first derivatives of the log-likelihood with respect to θ.

If the elements of $\mathbf{y}$ are drawn independently of each other, the log-likelihood is expressible as

$$l = \text{constant} + \ln p(\mathbf{y}|\theta) = \text{constant} + \sum_{i=1}^{n} \ln p(y_i|\theta). \tag{3.15}$$

Hence, the amount of information in a sample of size n can be written as

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{3.16}$$

where

$$I_i(\theta) = E\left[\left(\frac{d \ln p(y_i|\theta)}{d\theta}\right)^2\right]$$

and the expectation is taken over the marginal distribution of y_i. If the observations have independent and identical distributions, the result in (3.16) indicates that the amount of information in a sample of size n is exactly n times larger than that contained in a single draw from the distribution. This is because $I_i(\theta)$ is constant from observation to observation in this case.

Information accrues additively, irrespective of whether the observations are distributed independently or not. In the general case, it can be verified that, for a sample of size n,

$$\begin{aligned} I(\theta) = I_1(\theta) + I_{2.1}(\theta) + ... + I_{i.i-1,i-2,...,1}(\theta) + ... + I_{n-1.n-2,...,1}(\theta) \\ + I_{n.n-1,n-2,...,1}(\theta), \end{aligned} \tag{3.17}$$

where $I_{i.i-1,i-2,...,1}(\theta)$ is the information contributed by a sample from the distribution with density $p(y_i|y_{i-1}, y_{i-2}, ..., y_2, y_1, \theta)$. If the random variables are independent, the amount of information about θ is larger than otherwise.

Example 3.1 *Information in a sample of size two*

Consider a sample of size 2 consisting of draws from independently distributed random variables X and Y with p.d.f. $p(x|\theta)$ and $g(y|\theta)$. From

(3.14), the information about θ in the sample is

$$\begin{aligned} I(\theta) &= E\left[\frac{d}{d\theta}\ln p(x|\theta) + \frac{d}{d\theta}\ln g(y|\theta)\right]^2 \\ &= E\left[\frac{d}{d\theta}\ln p(x|\theta)\right]^2 + E\left[\frac{d}{d\theta}\ln g(y|\theta)\right]^2 \\ &\quad +2\,E\left[\frac{d}{d\theta}\ln p(x|\theta)\right]E\left[\frac{d}{d\theta}\ln g(y|\theta)\right]. \end{aligned} \tag{3.18}$$

The last line follows because X and Y are independent, and therefore, the expectation of the product is equal to the product of the expectations. As will be shown below, the terms in the third line are equal to zero. Letting $I_X(\theta) = E[\frac{d}{d\theta}\ln p(x|\theta)]^2$, (3.18) shows that the information in the sample is given by the sum of the information contributed by each sample point

$$I(\theta) = I_X(\theta) + I_Y(\theta).$$

■

Example 3.2 *Information under independent sampling from a normal distribution*

In the normal model with likelihood (3.10), the first derivative is given in (3.11). Upon squaring it, the term $(\overline{y} - m)^2$ is encountered and this has expectation σ^2/n. Hence

$$I(\theta) = \frac{n^2}{\sigma^4}\frac{\sigma^2}{n} = \frac{n}{\sigma^2},$$

so the information is proportional to sample size. ■

Example 3.3 *Information with correlated draws from a normal distribution*

Two lactation milk yields are available from the same cow. Their distribution is identical, having the same unknown mean μ and known variance σ^2, but there is a known correlation ρ between records, induced by factors that are common to all lactations of a cow. Here a bivariate normal sampling model may be appropriate, in which case the joint density of the two yields is

$$p(\mathbf{y}|\mu,\sigma^2,\rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)\,\sigma^4}}\exp\left\{-\frac{1}{2(1-\rho^2)\,\sigma^2}Q\right\},$$

where

$$Q = \left[(y_1-\mu)^2 + (y_2-\mu)^2 - 2\rho(y_1-\mu)(y_2-\mu)\right].$$

With μ being the only unknown parameter, the log-likelihood, apart from a constant is

$$l\left(\mu|\sigma^2,\rho,\mathbf{y}\right) = -\frac{(y_1-\mu)^2 + (y_2-\mu)^2 - 2\rho(y_1-\mu)(y_2-\mu)}{2(1-\rho^2)\,\sigma^2}.$$

After taking derivatives with respect to μ, squaring the result and taking expectations, the information about μ can be shown to be equal to

$$I(\mu) = \frac{2}{(1+\rho)\sigma^2}$$

and this is smaller than the information under independence in Example 3.2 by a factor $(1+\rho)^{-1}$. When the correlation is perfect, the information is equal to that obtained from a single sample drawn from $N(\mu, \sigma^2)$. ■

The definition of information given in (3.14) involves conceptual repeated sampling, as the expectation operator indicates averaging with respect to the distribution $[\mathbf{y}|\theta]$ or, equivalently, over all possible values that the random vector $\mathbf{y}$ can take at a fixed, albeit unknown, value of θ. This suggests that a more precise term is expected information, to make a distinction with the observed information resulting from a single realization of the random process. Hence, the observed information is $(dl/d\theta)^2$. The observed information is a function both of $\mathbf{y}$ and θ, whereas the expected information is a function of θ only.

The observed information represents the curvature (see Example 3.4 below) of the observed log-likelihood for the given data $\mathbf{y}$, whereas the expected information is an average curvature over realizations of $\mathbf{y}$. In this sense, the observed information is to be preferred (Efron and Hinkley, 1978). Typically, because the true value of the parameter is unknown, expected and observed information are approximated by replacing θ with the maximum likelihood estimate. The observed information evaluated at $\widehat{\theta}$ represents the curvature at $\widehat{\theta}$; a large curvature is associated with a strong peak, intuitively indicating less uncertainty about θ. In Example 3.2 the expected information was found to be n/σ^2, whereas the observed information is the square of (3.11). Note that the expected information does not depend on the unknown m, whereas the observed information involves the two parameters of the model $\left(m, \sigma^2\right)$ plus the sample average $\overline{y}$, which is the ML estimator in this case. Often, the expected information is algebraically simpler than the observed information.

A word of warning is particularly apposite here. Although one speaks of information about a parameter contained in a data point, the concept is mainly justified by the fact that, in many cases, under regularity conditions to be specified below, the inverse of the information is the smallest asymptotic variance obtainable by an estimator. In other words, the concept of information has an asymptotic justification and is meaningful provided regularity conditions are satisfied (Lehmann, 1999).

3.5.2 Alternative Representation of Information

Recall that $l = \ln p(\mathbf{y}|\theta)$, and let l' and l'' denote the first and second derivatives of the log-likelihood with respect to θ. Here it is shown that

Fisher's expected information can also be expressed as

$$\begin{aligned} I(\theta) &= -E_{\mathbf{y}}(l'') \\ &= -\int l'' p(\mathbf{y}|\theta)\, d\mathbf{y} \end{aligned} \tag{3.19}$$

and, thus,

$$\begin{aligned} I(\theta) &= E_{\mathbf{y}}(l')^2 \\ &= E_{\mathbf{y}}(-l''). \end{aligned}$$

Since $p(\mathbf{y}|\theta)$ is a density function, it follows that

$$\int p(\mathbf{y}|\theta)\, d\mathbf{y} = 1.$$

Therefore,

$$\frac{d}{d\theta}\int p(\mathbf{y}|\theta)\, d\mathbf{y} = 0.$$

If the derivative can be taken inside the integral sign, it follows that

$$\begin{aligned} 0 &= \int \frac{d}{d\theta} p(\mathbf{y}|\theta)\, d\mathbf{y} \\ &= \int l' p(\mathbf{y}|\theta)\, d\mathbf{y} \\ &= E_{\mathbf{y}}(l'). \end{aligned} \tag{3.20}$$

Thus, the expected value of the score is equal to zero. If a second differentiation also can be taken under the integral sign, then we have

$$\begin{aligned} 0 &= \int \frac{d}{d\theta} l' p(\mathbf{y}|\theta)\, d\mathbf{y} \\ &= \int l'' p(\mathbf{y}|\theta)\, d\mathbf{y} + \int (l')^2 p(\mathbf{y}|\theta)\, d\mathbf{y} \end{aligned}$$

leading directly to (3.19).

Example 3.4 *Curvature as a measure of information*
Another way of visualizing information derives from the definition of curvature at a given point of the log-likelihood function $l(\theta)$. The curvature at the point θ is (Stein, 1977)

$$c(\theta) = \frac{l''(\theta)}{\left[1 + l'(\theta)^2\right]^{\frac{3}{2}}}.$$

Since $l'\left(\widehat{\theta}\right) = 0$, the curvature at the maximum of the function is given by the second derivative

$$c\left(\widehat{\theta}\right) = l''\left(\widehat{\theta}\right).$$

When the curvature is large, the sample of data points clearly toward the value $\widehat{\theta}$. On the other hand, if the curvature is small, there is ambiguity since a range of values of θ leads to almost the same value of the likelihood. ■

Example 3.5 *A quadratic approximation to the log-likelihood*
Consider a second order Taylor series expansion of $l(\theta) = \ln L(\theta|\mathbf{y})$ around $\widehat{\theta}$

$$\begin{aligned} l(\theta) &\approx l\left(\widehat{\theta}\right) + l'\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right) + \frac{1}{2}l''\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^2 \\ &= l\left(\widehat{\theta}\right) - \frac{1}{2}J\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^2, \end{aligned}$$

where $J\left(\widehat{\theta}\right) = -l''\left(\widehat{\theta}\right)$ is the observed information. Therefore,

$$\frac{L(\theta|\mathbf{y})}{L\left(\widehat{\theta}|\mathbf{y}\right)} \approx \exp\left[-\frac{1}{2}J\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^2\right]. \tag{3.21}$$

It is important to note that in (3.21), θ is a fixed parameter, and $\widehat{\theta}$ varies in conceptual replications.
In Subsection 3.7.3, it will be shown that under regularity conditions, a slight variant of the following asymptotic approximation can be established

$$\widehat{\theta} \sim N\left(\theta, J\left(\widehat{\theta}\right)^{-1}\right),$$

which means that, approximately,

$$p\left(\widehat{\theta}\right) \approx (2\pi)^{-\frac{1}{2}}\left|J\left(\widehat{\theta}\right)\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}J\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^2\right]. \tag{3.22}$$

Again, this is an approximation to the sampling density of $\widehat{\theta}$, with θ fixed. Using (3.21) in (3.22), we obtain

$$p\left(\widehat{\theta}\right) \approx (2\pi)^{-\frac{1}{2}}\left|J\left(\widehat{\theta}\right)\right|^{\frac{1}{2}} \frac{L(\theta|\mathbf{y})}{L\left(\widehat{\theta}|\mathbf{y}\right)}, \tag{3.23}$$

which is a more accurate approximation than (3.22). A slightly modified version of this formula due to Barndorff-Nielsen (1983) is so precise that Efron (1998) refers to it as the "magic formula". ■

3.5.3 Mean and Variance of the Score Function

The mean and variance of the score function are needed to establish some properties of the ML estimator, as discussed later. The score function in a single parameter model is $S(\theta|\mathbf{y}) = l'$. As seen in connection with (3.11), the score is a function of the data, so it must be a random variable. Usually, the distribution of the score is unknown, although in (3.11) it possesses a normal distribution, by virtue of being a linear function of **y,** a normally distributed vector. In (3.20), it was shown that the score has zero expectation. Here we show that the variance of the score is equal to Fisher's expected information. Thus

$$l' = S(\theta|\mathbf{y}) \sim [0, I(\theta)]. \tag{3.24}$$

The variance of the score, by definition, is

$$Var(l') = E(l')^2 - E^2(l') = E(l')^2 = I(\theta) \tag{3.25}$$

which follows from the definition of information as given in (3.14) and (3.19). Note, from the derivation in (3.20), that it is assumed that the derivative can be moved inside of the integral sign.

Example 3.6 *Mean and variance of the score function in the normal distribution*

When sampling n observations independently from a normal distribution with mean μ and variance σ^2, the score, as given in (3.11), is

$$\frac{dl(\mu|\sigma^2, \mathbf{y})}{d\mu} = \frac{n(\overline{y} - \mu)}{\sigma^2}.$$

Consequently the score is a function of the data, as stated before. For any particular sample, it can be positive or negative, depending on the value of $\overline{y} - \mu$. However, over an infinite number of samples drawn from the distribution of $\mathbf{y}$, its average value is

$$E_y\left[\frac{n(\overline{y} - \mu)}{\sigma^2}\right] = \frac{n}{\sigma^2} E_y[(\overline{y} - \mu)] = 0$$

and it has variance

$$Var\left[\frac{n(\overline{y} - \mu)}{\sigma^2}\right] = \frac{n}{\sigma^2}.$$

This is precisely the information about μ, as found in Example 3.2. This completes the verification that the distribution of the score has parameters as in (3.24). In addition, this distribution is normal in the present example, as pointed out previously. ■

3.5.4 Multiparameter Case

The results given above generalize to a vector of parameters $\boldsymbol{\theta}$ in a straightforward manner. The expected information matrix is defined to be:

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\mathbf{y}}\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] = -E_{\mathbf{y}}\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\right), \tag{3.26}$$

where $l = \ln(\boldsymbol{\theta}|\mathbf{y})$ is now a function of a vector-valued parameter, $\partial l/\partial \boldsymbol{\theta}$ is a vector of first partial derivatives of the log-likelihood with respect to each of the elements of $\boldsymbol{\theta}$, and $\partial^2 l/\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'$ is the matrix (whose dimension is equal to the number of elements in $\boldsymbol{\theta}$) of second derivatives of the log-likelihood with respect to the parameters. The equivalent of (3.24) is that now there is a score vector $S(\boldsymbol{\theta}|\mathbf{y})$ having the multivariate distribution

$$S(\boldsymbol{\theta}|\mathbf{y}) \sim [\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})]. \tag{3.27}$$

The concept of information is more subtle in the multiparameter than in the single parameter case, unless the information matrix is diagonal. In general, in a multiparameter model, the ith diagonal element of $\mathbf{I}(\boldsymbol{\theta})$ cannot be interpreted literally as information about the ith element of θ. Often, reference needs to be made to the entire information matrix.

To illustrate, consider the sampling model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ where $\boldsymbol{\beta}$ is unknown, $\mathbf{X}$ is a known matrix of explanatory variables and $\mathbf{V}$ is a nonsingular variance–covariance matrix, assumed known. The likelihood function is expressible as

$$L(\boldsymbol{\beta}|\mathbf{V}, \mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]. \tag{3.28}$$

The score vector is

$$S(\boldsymbol{\beta}|\mathbf{y}) = \frac{\partial\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.29}$$

The observed information matrix is equal to the expected information matrix in this case, because the second derivatives do not involve $\mathbf{y}$. Here

$$\begin{aligned}\mathbf{I}(\boldsymbol{\beta}) &= E\left[-\frac{\partial^2 l}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'}\right] = E\left[-\frac{\partial S(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}'}\right] \\ &= E\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right) = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}.\end{aligned} \tag{3.30}$$

From (3.26) and (3.29), the information matrix can also be calculated as

$$\begin{aligned}\mathbf{I}(\boldsymbol{\beta}) &= E\left[\left(\frac{\partial l}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\beta}}\right)'\right] \\ &= E\left[\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{X}\right] \\ &= \mathbf{X}'\mathbf{V}^{-1}E\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right]\mathbf{V}^{-1}\mathbf{X} \\ &= \mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}.\end{aligned}$$

The preceding result follows because the variance–covariance matrix of $\mathbf{y}$, by definition, is $\mathbf{V} = E\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right]$. Also, note from (3.29) that

$$E\left[S\left(\boldsymbol{\beta}|\mathbf{y}\right)\right] = \mathbf{X}'\mathbf{V}^{-1}E\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}$$

and that

$$\begin{aligned} Var\left[S\left(\boldsymbol{\beta}|\mathbf{y}\right)\right] &= \mathbf{X}'\mathbf{V}^{-1}\left[Var\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right]\mathbf{V}^{-1}\mathbf{X}' \\ &= \mathbf{X}'\mathbf{V}^{-1}\,Var\left(\mathbf{y}\right)\mathbf{V}^{-1}\mathbf{X}' = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}. \end{aligned}$$

This verifies that the score has a distribution with mean $\mathbf{0}$ and a variance–covariance matrix equal to the expected information. Here, this distribution is multivariate normal, with dimension equal to the number of elements in $\boldsymbol{\beta}$. This is because the score vector in (3.29) is a linear transformation of the data vector $\mathbf{y}$.

Example 3.7 *Linear regression*
In a simple linear regression model, it is postulated that the observations are linked to an intercept β_0 and to a slope parameter β_1 via the relationship

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

where x_i $(i = 1, 2, ..., n)$ are known values of an explanatory variable and $e_i \sim N\left(0, \sigma^2\right)$ is a residual. The distributions of the n residuals are assumed to be mutually independent. The parameter vector is $[\beta_0, \beta_1, \sigma^2]$. The likelihood function can be written as

$$L\left(\beta_0, \beta_1, \sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

and the corresponding log-likelihood, apart from an additive constant, is

$$l\left(\beta_0, \beta_1, \sigma^2|\mathbf{y}\right) = -\frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

The score vector is given by

$$\begin{bmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \frac{\partial l}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) \\ \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y_i - \beta_0 - \beta_1 x_i\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 \end{bmatrix}.$$

Setting the score vector to zero and solving simultaneously for the unknown parameters gives explicit solutions to the ML equations. The ML estimators

are

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n},$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the solutions to the matrix equation

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}.$$

Explicitly,

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

and

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2}.$$

The first two columns of the 3×3 matrix of negative second derivatives of the log-likelihood with respect to the parameters, or observed information matrix, are

$$\begin{bmatrix} \sigma^{-2} n & \sigma^{-2} \sum_{i=1}^{n} x_i \\ \sigma^{-2} \sum_{i=1}^{n} x_i & \sigma^{-2} \sum_{i=1}^{n} x_i^2 \\ \sigma^{-4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) & \sigma^{-4} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \end{bmatrix}$$

and the last column

$$\begin{bmatrix} \sigma^{-4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \sigma^{-4} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \\ -\left(2\sigma^4\right)^{-1} n + \sigma^{-6} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{bmatrix}.$$

It is easy to verify that the expected value of each of the elements of the score vector is null. For example,

$$E\left(\frac{\partial l}{\partial \beta_1}\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i E\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

and

$$E\left(\frac{\partial l}{\partial \sigma^2}\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} E\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$
$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} n\sigma^2 = 0.$$

Further, the expected information matrix is given by

$$\mathbf{I}(\theta) = \sigma^{-2}\begin{bmatrix} n & \sum_{i=1}^{n} x_i & 0 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & 0 \\ 0 & 0 & \left(2\sigma^2\right)^{-1} n \end{bmatrix}.$$

Note that the elements $(1,2)$ and $(2,1)$ of this matrix are not null. This illustrates that in a multiparameter model it is often more sensible to speak about joint information on a set of parameters, rather than about information on individual parameters themselves. Also, observe that the elements $(1,3)$, $(3,1)$, $(2,3)$, and $(3,2)$ are null. This has a connection with the distribution of the ML estimator of $\boldsymbol{\theta}$, as discussed later. ■

3.5.5 Cramér–Rao Lower Bound

The concept of expected information can be used to determine a lower bound for the variance of an estimator of the scalar parameter θ. Let $\widehat{\theta} = T(\mathbf{y})$ be a function of a sample $\mathbf{y}$ drawn from the distribution $[\mathbf{y}|\theta]$; this function, usually called a statistic, is employed for estimating θ. Also, let

$$E\left(\widehat{\theta}\right) = m(\theta),$$

where $m(\theta)$ is a function of θ; for example, m can be the identity operator. If $m(\theta) = \theta$, the estimator is said to be unbiased for θ. The Cramér–Rao inequality (e.g., Casella and Berger, 1990) states that

$$Var\left(\widehat{\theta}\right) \geq \frac{\left[m'(\theta)\right]^2}{I(\theta)}, \tag{3.31}$$

where $m'(\theta) = dm(\theta)/d\theta$, assuming $m(\theta)$ is differentiable. If the variance of $\widehat{\theta}$ attains the right-hand side of (3.31), then the estimator is said to have minimum variance.

A particular case is when $\widehat{\theta}$ is unbiased, that is, when $E(\widehat{\theta}) = \theta$. Here, the derivative of $m(\theta) = \theta$ with respect to θ is equal to 1. Then the Cramér–Rao lower bound for the variance of an unbiased estimator is

$$Var\left(\widehat{\theta}\right) \geq \left[I(\theta)\right]^{-1}. \tag{3.32}$$

This states that an unbiased estimator cannot have a variance that is lower than the inverse of Fisher's information measure. If

$$Var\left(\widehat{\theta}\right) = \left[I\left(\theta\right)\right]^{-1},$$

then $\widehat{\theta}$ is said to be a minimum variance unbiased estimator.

In general, as in (3.31), the lower bound depends on the expectation of the estimator and on the distribution of the observations, because $I\left(\theta\right)$ depends on $p\left(\mathbf{y}|\theta\right)$. In order to prove (3.31), use is made of the Cauchy–Schwarz inequality (Stuart and Ord, 1991), as given below.

Cauchy–Schwarz Inequality

Let u and v be any two random variables and let c be any constant. Then $\left(u - cv\right)^2$ is positive or null, and so is its expectation. Hence

$$E\left(u - cv\right)^2 = E\left(u^2\right) + c^2 E\left(v^2\right) - 2cE\left(uv\right) \geq 0.$$

In this expression, arbitrarily choose

$$c = \frac{E\left(uv\right)}{E\left(v^2\right)}$$

to obtain

$$E\left(u^2\right) + \frac{E^2\left(uv\right)}{E\left(v^2\right)} - 2\frac{E^2\left(uv\right)}{E\left(v^2\right)} \geq 0.$$

This leads directly to the Cauchy–Schwarz inequality

$$E^2\left(uv\right) \leq E\left(u^2\right) E\left(v^2\right). \tag{3.33}$$

If the random variables are expressed as deviations from their expectations, the above implies that

$$Cov^2\left(u, v\right) \leq Var\left(u\right) Var\left(v\right). \tag{3.34}$$

In addition, if these deviations are measured in standard deviation units of the corresponding variate, (3.34) implies that

$$Corr^2\left(u, v\right) \leq 1, \tag{3.35}$$

where $Corr(\cdot)$ denotes the coefficient of correlation. ■

Now let $u = \widehat{\theta}$, and $v = l'$, the score function. It has been established already that $E(l') = 0$ and $Var(l') = I(\theta)$. Hence

$$\begin{aligned} Cov\left(\widehat{\theta}, l'\right) &= \int \left[\widehat{\theta}\ \frac{d \ln p(\mathbf{y}|\theta)}{d\theta}\right] p(\mathbf{y}|\theta)\, d\mathbf{y} \\ &= \int \widehat{\theta}\ \frac{dp(\mathbf{y}|\theta)}{d\theta} d\mathbf{y} \\ &= \frac{d}{d\theta} \int \widehat{\theta}\ p(\mathbf{y}|\theta)\, d\mathbf{y} \\ &= \frac{d}{d\theta} E\left(\widehat{\theta}\right) = \frac{d}{d\theta} m(\theta) = m'(\theta). \end{aligned} \tag{3.36}$$

Applying the Cauchy–Schwarz inequality, as in (3.34),

$$Cov^2\left(\widehat{\theta}, l'\right) = \left[m'(\theta)\right]^2 \leq Var\left(\widehat{\theta}\right) I(\theta).$$

Rearrangement of this expression leads directly to the Cramér–Rao lower bound given in (3.31). Note that the proof requires interchange of integration and differentiation, as seen in connection with (3.36). There are situations in which it is not possible to do this; typically, when the limits of integration depend on the parameter. An example is provided by the uniform distribution

$$X \sim Un(0, \theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Integration with respect to $p(x|\theta)$ is over the range $0 \leq x \leq \theta$, which includes θ.

Example 3.8 *Cramér–Rao lower bound in the linear regression model*
Return to the linear regression Example 3.7, and consider finding the Cramér–Rao lower bound for an estimator of the variance. It was seen that $I\left(\sigma^2\right) = n/2\sigma^4$. Hence, any unbiased estimator of the variance ($\widehat{v}$, *say*) must be such that

$$Var(\widehat{v}) \geq \frac{2\sigma^4}{n}.$$

The ML estimator of σ^2 for this example can be written as

$$\begin{aligned} \widehat{\sigma^2} &= \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2}{n} \\ &= \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2}{\sigma^2} \frac{\sigma^2}{n} \\ &= \chi^2_{n-2} \frac{\sigma^2}{n}, \end{aligned}$$

because the residual sum of squares

$$\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)^2\sim\chi^2_{n-2}\sigma^2$$

has a scaled chi-square distribution with $n-2$ degrees of freedom. The expectation of the estimator is

$$E\left(\widehat{\sigma}^2\right)=\sigma^2\frac{n-2}{n},$$

so the estimator has a downward bias. It can readily be verified that the variance of the ML estimator of σ^2 satisfies the Cramér–Rao lower bound given by (3.31) and equal to $2\sigma^4\left(n-2\right)/n^2$. An unbiased estimator is

$$\widetilde{\sigma}^2=\widehat{\sigma}^2\frac{n}{n-2}=\chi^2_{n-2}\frac{\sigma^2}{n-2},$$

with variance

$$Var\left(\widetilde{\sigma}^2\right)=\frac{2\sigma^4}{n-2}.$$

Hence, this unbiased estimator does not attain the Cramér–Rao lower bound for $\widehat{v}$ given above. ■

This example suggests that it would be useful to find a condition under which an unbiased estimator $\widehat{\theta}$ reaches the lower bound. Suppose that the score can be written as a linear function of $\widehat{\theta}$ as follows:

$$l'=a\left(\theta\right)\left(\widehat{\theta}-\theta\right), \tag{3.37}$$

where $a\left(\theta\right)$ is some constant that does not involve the observations. Then,

$$\begin{aligned}Cov\left(\widehat{\theta},l'\right)&=\int\widehat{\theta}\;a\left(\theta\right)\left(\widehat{\theta}-\theta\right)p\left(\mathbf{y}|\theta\right)d\mathbf{y}\\&=a\left(\theta\right)Var\left(\widehat{\theta}\right),\end{aligned}$$

and

$$Var\left(l'\right)=I\left(\theta\right)=a^2\left(\theta\right)Var\left(\widehat{\theta}\right).$$

The Cramér–Rao inequality in (3.31), in view of (3.37), can be written as

$$\begin{aligned}Cov^2\left(\widehat{\theta},l'\right)&=a^2\left(\theta\right)Var^2\left(\widehat{\theta}\right)\leq Var\left(\widehat{\theta}\right)I\left(\theta\right)\\&=Var\left(\widehat{\theta}\right)a^2\left(\theta\right)Var\left(\widehat{\theta}\right).\end{aligned}$$

The inequality becomes an equality, so the Cramér–Rao lower bound is attained automatically for an unbiased estimator $\widehat{\theta}$ provided the score can

be written as in (3.37). For example, in a normal sampling process with unknown mean and known variance, the score is, as given in (3.11),

$$\frac{dl\left(\mu|\sigma^2,\mathbf{y}\right)}{d\mu}=\frac{n\left(\overline{y}-\mu\right)}{\sigma^2}.$$

Because this has the required form, with $\theta=\mu$, $a\left(\theta\right)=n/\sigma^2$ and $\widehat{\theta}=\overline{y}$, it follows that the sample mean is the minimum variance unbiased estimator of μ.

In a multiparameter situation, the condition for an unbiased estimator to attain the lower bound is written as

$$l'=\mathbf{A}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right), \tag{3.38}$$

where l is the score vector and $\mathbf{A}\left(\boldsymbol{\theta}\right)$ is a matrix that may be a function of the parameters, but not of the data. For example, under the sampling model $\mathbf{y}\sim N\left(\mathbf{X}\boldsymbol{\beta},\mathbf{V}\right)$, with known variance–covariance matrix, it was seen in (3.29) that

$$\mathbf{l}'=S\left(\boldsymbol{\beta}|\mathbf{y}\right)=\mathbf{X}'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right),$$

and if $\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$ exists, the score can be written as

$$S\left(\boldsymbol{\beta}|\mathbf{y}\right)=\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}-\boldsymbol{\beta}\right],$$

which is in the form of (3.38), with $\boldsymbol{\theta}=\boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}}=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, and $\mathbf{A}\left(\boldsymbol{\theta}\right)=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)$. Hence, $\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is the minimum variance unbiased estimator of $\boldsymbol{\beta}$.

3.6 Sufficiency

The concept of sufficiency was developed by Fisher in the early 1920s (Fisher, 1920, 1922). Let $\mathbf{y}$ represent data from the sampling model $p\left(\mathbf{y}|\theta\right)$. An estimator $T\left(\mathbf{y}\right)$ of a parameter θ is said to be sufficient if the conditional distribution of the data $\mathbf{y}$, given $T\left(\mathbf{y}\right)$, is independent of θ. This implies that $T\left(\mathbf{y}\right)$ contains as much information about θ as the data itself. Obviously, when $T\left(\mathbf{y}\right)=\mathbf{y}$, then $T\left(\mathbf{y}\right)$ is sufficient. But this is not very useful because the idea of sufficiency is to reduce dimensionality of $\mathbf{y}$ without losing information about θ.

The definition of sufficiency is model dependent. That is, if $T\left(\mathbf{y}\right)$ is sufficient for θ under a certain probability model, it may not be so under another probability model.

Given a model, ML estimators are sufficient. To see this, assume that the likelihood can be factorized into a term that depends on θ and a second

term that does not

$$\begin{aligned} L(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta) \\ &= g(T(\mathbf{y})|\theta)\, h(\mathbf{y}), \end{aligned} \tag{3.39}$$

where the equality must hold for all $\mathbf{y}$ and θ. The solution of the equation

$$\frac{dL(\theta|\mathbf{y})}{d\theta} = 0$$

is the same as that of equation

$$\frac{dg(T(\mathbf{y})|\theta)}{d\theta} = 0.$$

Therefore, the ML estimator must be a function of the sufficient statistic $T(\mathbf{y})$, if the latter exists. Similarly, if

$$T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_r(\mathbf{y})$$

are jointly sufficient statistics in a model with parameters

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_r]',$$

then the ML estimators of these parameters are functions only of $T_1(\mathbf{y})$, $T_2(\mathbf{y}), \ldots, T_r(\mathbf{y})$ and are, thus, jointly sufficient themselves.

Sufficient statistics are not unique. If a statistic T is sufficient for a parameter θ, then a one-to-one transformation function of T is also sufficient.

Example 3.9 *A sample from a normal distribution with known variance* The log-likelihood based on a sample from a normal distribution with unknown mean and known variance can be put, from (3.10), as

$$l(\mu|\sigma^2, \mathbf{y}) = \text{constant} - \frac{n(\overline{y} - \mu)^2}{2\sigma^2} - \frac{\sum_i (y_i - \overline{y})^2}{2\sigma^2}.$$

Clearly, the first term is in the form $g(T(\mathbf{y})|\mu)$, and the second term is independent of μ. In this example, the sufficient statistic for μ is $T(\mathbf{y}) = \overline{y}$, which is the ML estimator of this parameter. ■

3.7 Asymptotic Properties: Single Parameter Models

An important reason why ML estimation is often advocated in the statistical literature is because the method possesses some properties that are deemed desirable, in some sense. In general, these properties can be shown

to hold only when, given n independent draws from the same distribution, sample size increases beyond all bounds, i.e., when $n \to \infty$. Such properties, termed asymptotic, are consistency, efficiency and asymptotic normality. A mathematically rigorous discussion of these properties is beyond the scope of this book. However, because application of ML estimation typically requires recourse to asymptotic properties, some results based on the classical, first-order asymptotic theory are presented, to enhance understanding of this method of estimation.

3.7.1 Probability of the Data Given the True Value of the Parameter

Suppose n observations are drawn independently from the distribution $[y|\theta_0]$, where θ_0 is the true value of the parameter θ, at some point of the parameter space. The joint density of the observations under this distribution is then

$$p\left(\mathbf{y}|\theta_0\right) = \prod_{i=1}^{n} p\left(y_i|\theta_0\right).$$

Likewise, let

$$p\left(\mathbf{y}|\theta\right) = \prod_{i=1}^{n} p\left(y_i|\theta\right)$$

be the joint density under any other value of the parameter. Then, under certain conditions described in, e.g., Lehmann and Casella (1998),

$$\lim_{n\to\infty} \Pr\left\{\prod_{i=1}^{n} p\left(y_i|\theta_0\right) > \prod_{i=1}^{n} p\left(y_i|\theta\right)\right\} = 1. \tag{3.40}$$

This result can be interpreted in the following manner: as the sample gets larger, the density (or probability if the random variable is discrete) of the observations at θ_0 exceeds that at any other value of θ with high probability. In order to prove the above, consider the statement

$$\frac{\prod_{i=1}^{n} p\left(y_i|\theta\right)}{\prod_{i=1}^{n} p\left(y_i|\theta_0\right)} < 1.$$

This is equivalent to

$$\ln\left[\frac{p\left(\mathbf{y}|\theta\right)}{p\left(\mathbf{y}|\theta_0\right)}\right] = \sum_{i=1}^{n} \ln\left[\frac{p\left(y_i|\theta\right)}{p\left(y_i|\theta_0\right)}\right] < 0$$

and to

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{p\left(y_i|\theta\right)}{p\left(y_i|\theta_0\right)}\right] < 0. \tag{3.41}$$

Then, by the law of large numbers,

$$\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\left[\frac{p\left(y_i|\theta\right)}{p\left(y_i|\theta_0\right)}\right]\right\}=E\left\{\ln\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]\right\}.$$

In an infinitely large sample the condition in (3.41) would become

$$E\left\{\ln\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]\right\}<0. \tag{3.42}$$

By Jensen's inequality (see below):

$$E\left\{\ln\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]\right\}<\ln E\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]. \tag{3.43}$$

Now

$$\begin{aligned}E\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]&=\int\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}p\left(y|\theta_0\right)dy\\&=\int p\left(y|\theta\right)dy=1.\end{aligned}$$

so

$$\ln E\left[\frac{p\left(y|\theta\right)}{p\left(y|\theta_0\right)}\right]=0. \tag{3.44}$$

In view of (3.43), with the right-hand side as in (3.44), inequality (3.42) follows, thus proving the statement in (3.40). As pointed out by Lehmann and Casella (1998), (3.40) suggests that if in large samples the ML estimator of θ were close to θ_0, it would constitute a reasonable estimator, generating the observed data with near maximum probability.

Jensen's Inequality

Jensen's inequality (Rao, 1973; Casella and Berger, 1990) states that if X is a random variable, and $g(X)$ is a concave function ("holds water"), then

$$g\left[E(X)\right]\leq E\left[g(X)\right]. \tag{3.45}$$

If $g(X)$ is convex ("spills water")

$$E\left[g(X)\right]\leq g\left[E(X)\right]. \tag{3.46}$$

A function is said to be convex if its second derivative with respect to the variable is negative or null throughout. For example, $\ln\left(X\right)$ is convex. Under convexity, the function lies below all its tangent lines (Casella and Berger, 1990). Hence, it must be true that

$$g(X)\leq l(X)=a+bX,$$

where $l(X)$ is a line tangent to $g(X)$ at X. Taking expectations of the inequality

$$E\left[g(X)\right] \leq E\left[l(X)\right] = l\left[E\left(X\right)\right] = g\left[E\left(X\right)\right],$$

thus proving (3.46). The first equality arises because the expectation of a linear function is equal to the linear function of the expectation, and the second, because l is the tangent at $E\left(X\right)$.

A similar argument can be employed for a concave upward function ("holds water"); in this case the function lies above the tangent lines. ■

3.7.2 Consistency

Suppose that $\widehat{\theta}_n$ is an estimator of the parameter θ based on random variables $Y_1, Y_2, \ldots, Y_n$. If, as n increases, the sampling distribution of $\widehat{\theta}_n$ becomes more and more concentrated around θ, then $\widehat{\theta}_n$ is said to be consistent. More formally, the sequence of estimators $\{\widehat{\theta}_n\}$ is consistent if $\{\widehat{\theta}_n\}$ converges in probability to the constant θ. That is,

$$\Pr\left(\left|\widehat{\theta}_n - \theta\right| < \epsilon\right) \to 1 \quad \text{as } n \to \infty$$

for each $\epsilon > 0$ and each θ. Although convergence refers to a sequence of estimators, one writes informally that $\widehat{\theta}_n$ is a consistent estimator of θ.

The consistency property of the ML estimator (sketched below) states that, as $n \to \infty$, the solution to the ML equation $l'\left(\theta\right) = 0$ has a root $\widehat{\theta}_n = f(\mathbf{y})$ tending to the true value θ_0 with probability 1. For this to hold, the likelihood must be differentiable with respect to $\theta \in \Omega$, and the observations must be i.i.d.. Following Lehmann and Casella (1998), where more technical detail is given, suppose that a is small enough so that $(\theta_0 - a, \theta_0 + a) \in \Omega$, and consider the event

$$\begin{aligned} S_n &= \left\{\left[p\left(\mathbf{y}|\theta_0\right) > p\left(\mathbf{y}|\theta_0 - a\right)\right] \cap \left[p\left(\mathbf{y}|\theta_0\right) > p\left(\mathbf{y}|\theta_0 + a\right)\right]\right\} \\ &= \left\{\left[l\left(\theta_0\right) > l\left(\theta_0 - a\right)\right] \cap \left[l\left(\theta_0\right) > l\left(\theta_0 + a\right)\right]\right\}. \end{aligned}$$

By virtue of (3.40), it must be true that, as $n \to \infty$, then $\Pr\left(S_n\right) \to 1$. This implies that within the interval considered there exists a value $\theta_0 - a < \widehat{\theta}_n < \theta_0 + a$ at which $l\left(\theta\right)$ has a local maximum. Hence, the first-order condition $l'\left(\widehat{\theta}_n\right) = 0$ would be satisfied. It follows that for any $a > 0$ sufficiently small, it must be true that

$$\Pr\left(\left|\widehat{\theta}_n - \theta_0\right| < a\right) \to 1. \tag{3.47}$$

An alternative motivation follows from (3.40). Note that

$$\lim_{n\to\infty} \Pr\left\{\ln \prod_{i=1}^{n} p\left(y_i|\theta_0\right) > \ln \prod_{i=1}^{n} p\left(y_i|\theta\right)\right\} = 1$$

can be restated as

$$\lim_{n\to\infty} \Pr\left\{l\left(\theta_0|\mathbf{y}\right) > l\left(\theta|\mathbf{y}\right)\right\} = 1.$$

However, by definition of the ML estimator, it must be true that

$$l\left(\widehat{\theta}_n|\mathbf{y}\right) \geqq l\left(\theta_0|\mathbf{y}\right).$$

Together the two preceding expressions imply that as $n \to \infty$, $l\left(\widehat{\theta}_n|\mathbf{y}\right)$ must take the value $l\left(\theta_0|\mathbf{y}\right)$. This means that

$$\lim_{n\to\infty} \Pr\left(\widehat{\theta}_n = \theta_0\right) = 1,$$

which shows the consistency of $\widehat{\theta}_n$.

Feng and McCulloch (1996) have extended these results by proving consistency of the ML estimator when the true parameter value is on the boundary of the parameter space.

3.7.3 Asymptotic Normality and Efficiency

Lehmann and Casella (1998) give a formal statement of the asymptotic properties of the ML estimator $\widehat{\theta}_n$ of the parameter $\theta \in \Omega$, where Ω is the parameter space. These properties are attained subject to the following regularity conditions:

1. The parameter space Ω is an open interval (not necessarily finite). This guarantees that θ lies inside Ω and not on the boundaries. Unless this condition is satisfied, a generally valid Taylor expansion of $(\widehat{\theta} - \theta)$ is not possible. If Ω is a closed interval (such as $0 \leq h^2 \leq 1$), the theory holds for values of the parameters that do not include the boundary.

2. The support of the p.d.f. of the data does not depend on θ; that is, the set
$$A = \{\mathbf{y} : p\left(\mathbf{y}|\theta\right) > 0\}$$
is independent of θ. The data are $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and the Y_i are i.i.d. with p.d.f. $p\left(\mathbf{y}|\theta\right)$ or with p.m.f. $\Pr\left(\mathbf{Y} = \mathbf{y}|\theta\right) = p\left(\mathbf{y}|\theta\right)$.

3. The distributions of the observations are distinct; that is,
$$F\left(\mathbf{y}|\theta_1\right) = F\left(\mathbf{y}|\theta_2\right)$$
implies $\theta_1 = \theta_2$. In other words, the parameter must be identifiable. In linear models, this condition is typically referred to as estimability (Searle, 1971).

4. The density $p(\mathbf{y}|\theta)$ is three times differentiable with respect to θ and the third derivative is continuous in θ.

5. The integral $\int p(\mathbf{y}|\theta)\, d\mathbf{y}$ can be differentiated three times under the integral sign. This condition implies $E[l'(\theta|\mathbf{y})] = 0$ and that

$$E[-l''(\theta|\mathbf{y})] = E[l'(\theta|\mathbf{y})]^2 = I(\theta).$$

6. The Fisher information satisfies $0 < I(\theta) < \infty$.

7. There exists a function $M(\mathbf{y})$ (whose expectation is finite) such that third derivatives are bounded as follows:

$$\left| \frac{d^3}{(d\theta)^3} [\ln p(\mathbf{y}|\theta)] \right| \leq M(\mathbf{y})$$

for all $\mathbf{y}$ in A and for θ near θ_0, where θ_0 is the true value of the parameter.

Under the above conditions, Lehmann and Casella (1998) prove that $\widehat{\theta}_n$ is asymptotically ($n \to \infty$) normal

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{D} N\left[0, \frac{1}{I_1(\theta_0)}\right] \tag{3.48}$$

where $I_1(\theta_0) = I(\theta_0)/n$ is the amount of information about θ contained in a sample of size 1 and $I(\theta_0)$ is the information about θ contained in the data. The notation "$\xrightarrow{D}$" means convergence in distribution; it was encountered in Chapter 2, Section 2.2.4. See Casella and Berger (1990) and Lehmann (1999) for a careful development of this concept. As pointed out in Chapter 2, (3.48) means that the sequence of random variables $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$, as $n \to \infty$, has a sequence of cumulative distribution functions (c.d.f.) that converges to the c.d.f. of an $N\left[0, \frac{1}{I_1(\theta_0)}\right]$ random variable. Expression (3.48) is often interpreted as:

$$\widehat{\theta}_n \sim N(\theta_0, I(\theta_0)). \tag{3.49}$$

The basic elements of the proof are elaborated below.

Let $Y_1, Y_2, ..., Y_n$ be i.i.d. random variables from a distribution having true parameter value θ_0. The maximized log-likelihood is then $l\left(\widehat{\theta}_n\right)$, and expanding the corresponding score function about θ_0 yields

$$l'\left(\widehat{\theta}_n\right) \approx l'(\theta_0) + l''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right) + \frac{1}{2} l'''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)^2. \tag{3.50}$$

The first, second, and third derivatives are functions of both θ_0 and $\mathbf{y}$. Assuming that $\widehat{\theta}_n$ is the maximizer of $l(\theta)$, then $l'\left(\widehat{\theta}_n\right) = 0$. Expression (3.50) can be rearranged, after multiplying both sides by $\sqrt{n}$, as

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \frac{\sqrt{n}\, l'(\theta_0)}{-l''(\theta_0) - \frac{1}{2} l'''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)}.$$

Dividing the numerator and denominator by n yields

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \frac{n^{-1/2}\, l'(\theta_0)}{-n^{-1} l''(\theta_0) - (2n)^{-1} l'''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)}. \tag{3.51}$$

Now we examine the limiting behavior of the terms in (3.51), as $n \to \infty$, considering the three terms in the numerator and denominator of (3.51) separately.

(1) First, note that

$$\begin{aligned}
\frac{1}{\sqrt{n}} l'(\theta_0) &= \sqrt{n}\frac{1}{n}\left\{\frac{d}{d\theta}\left[\ln p(\mathbf{y}|\theta)\right]\right\}_{\theta=\theta_0} \\
&= \sqrt{n}\frac{1}{n}\left\{\frac{d}{d\theta}\sum_{i=1}^{n}\ln\left[p(y_i|\theta)\right]\right\}_{\theta=\theta_0} \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{p'(y_i|\theta_0)}{p(y_i|\theta_0)}.
\end{aligned} \tag{3.52}$$

By being expressible as a sum of independent random variables, this function should have an asymptotically normal distribution (central limit theorem). Its expectation is

$$E\left[\frac{1}{\sqrt{n}} l'(\theta_0)\right] = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} E\left[\frac{p'(y_i|\theta_0)}{p(y_i|\theta_0)}\right] = 0 \tag{3.53}$$

this being so because

$$E\left[\frac{p'(y_i|\theta_0)}{p(y_i|\theta_0)}\right] = \int p'(y_i|\theta_0)\, dy_i = \frac{d}{d\theta}\int p(y_i|\theta_0)\, dy_i = 0$$

under the condition that differentiation and integration are interchangeable, as noted earlier. Also, from (3.25),

$$\begin{aligned}
Var\left[\frac{1}{\sqrt{n}} l'(\theta_0)\right] &= \frac{1}{n} Var\left[l'(\theta_0)\right] \\
&= \frac{1}{n} I(\theta_0) = \frac{n I_1(\theta_0)}{n} = I_1(\theta_0).
\end{aligned} \tag{3.54}$$

From (3.53), (3.54), and the central limit theorem, it follows that

$$\frac{1}{\sqrt{n}} l'(\theta_0) \xrightarrow{D} N[0, I_1(\theta_0)]. \tag{3.55}$$

(2) The second term of interest in (3.51) is

$$\begin{aligned} -\frac{1}{n} l''(\theta_0) &= -\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d}{d\theta} \left[\frac{p'(y_i|\theta)}{p(y_i|\theta)} \right] \right\}_{\theta=\theta_0} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{[p'(y_i|\theta_0)]^2 - p(y_i|\theta_0) p''(y_i|\theta_0)}{p^2(y_i|\theta_0)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{[p'(y_i|\theta_0)]^2}{p^2(y_i|\theta_0)} - \frac{1}{n} \sum_{i=1}^{n} \frac{p''(y_i|\theta_0)}{p(y_i|\theta_0)}. \end{aligned} \tag{3.56}$$

As $n \to \infty$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \frac{[p'(y_i|\theta_0)]^2}{p^2(y_i|\theta_0)} &\to E\left\{ \frac{[p'(y_i|\theta_0)]^2}{p^2(y_i|\theta_0)} \right\} \\ &= E\left[\frac{d}{d\theta} \ln p(y_i|\theta) \right]^2_{\theta=\theta_0} = I_1(\theta_0) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \frac{p''(y_i|\theta_0)}{p(y_i|\theta_0)} &\to E\left[\frac{p''(y_i|\theta_0)}{p(y_i|\theta_0)} \right] = \int p''(y_i|\theta_0)\, dy_i \\ &= \left[\frac{d^2}{d\theta} \int p(y_i|\theta)\, dy_i \right]_{\theta=\theta_0} = 0. \end{aligned}$$

Hence, in probability, we have that (3.56)

$$-\frac{1}{n} l''(\theta_0) \to I_1(\theta). \tag{3.57}$$

(3) The third term in (3.51) is

$$-\frac{1}{2n} l'''(\theta_0) \left(\widehat{\theta}_n - \theta_0 \right).$$

Note that

$$\frac{1}{n} l'''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d^3}{(d\theta)^3} [\ln p(y_i|\theta)] \right\}_{\theta=\theta_0}.$$

Based on the regularity condition 7, suppose that the third derivatives are bounded as follows:

$$\left| \frac{d^3}{(d\theta)^3} [\ln p(\mathbf{y}|\theta)] \right| \le M(\mathbf{y})$$

for all $\mathbf{y}$ in A and for θ near θ_0. Then

$$\left|\frac{1}{n}l'''(\theta_0)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\left\{\frac{d^3}{(d\theta)^3}\left[\ln p(y_i|\theta)\right]\right\}_{\theta=\theta_0}\right|$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}M(y_i).$$

As $n \to \infty$, $\left|n^{-1}l'''(\theta_0)\right|$ must be smaller than or equal to $E[M(y)]$, which in condition 7 above, is assumed to be finite. Hence, as sample size goes to infinity, the term

$$-\frac{1}{2n}l'''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right) \to 0, \tag{3.58}$$

this being so because $\widehat{\theta}_n \to \theta_0$ and the third derivatives are bounded in the preceding sense.

Considering (3.55), collecting (3.57) and (3.58), then the expression of interest in (3.51), that is,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \frac{n^{-1/2}l'(\theta_0)}{-n^{-1}l''(\theta_0) - (2n)^{-1}l'''(\theta_0)\left(\widehat{\theta}_n - \theta_0\right)}$$

behaves in the limit as

$$\frac{n^{-1/2}l'(\theta_0)}{I_1(\theta_0)} \sim N\left[0, \frac{1}{I_1(\theta_0)}\right]. \tag{3.59}$$

This indicates that if

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \sim N\left[0, \frac{1}{I_1(\theta_0)}\right] \tag{3.60}$$

then, as $n \to \infty$, the ML estimator can be written as the random variable

$$\widehat{\theta}_n = \theta_0 + \frac{1}{\sqrt{nI_1(\theta_0)}}N(0,1)), \tag{3.61}$$

where $nI_1(\theta_0) = I(\theta_0)$. Hence, as $n \to \infty$:

(1) $E\left(\widehat{\theta}_n\right) = \theta_0$, so the ML estimator is asymptotically unbiased.

(2) $Var\left(\widehat{\theta}_n\right) = [nI_1(\theta_0)]^{-1} = [I(\theta_0)]^{-1}$, so it reaches the Cramér-Rao lower bound, i.e., it has minimum variance among asymptotically unbiased estimators of θ. When this limit is reached, the estimator is said to be efficient.

(3) Informally stated, as in (3.49), it is said that the ML estimator has the asymptotic distribution $\widehat{\theta}_n \backsim N\left[\theta_0, I^{-1}(\theta_0)\right]$. This facilitates inferences, although these are valid strictly for a sample having infinite size.

(4) As seen earlier, $\widehat{\theta}_n$ is consistent, reaching the true value θ_0 when $n \to \infty$.

3.8 Asymptotic Properties: Multiparameter Models

Consider now the situation where the distribution of the observations is indexed by parameters $\boldsymbol{\theta}$. This vector can have distinct components, some of primary interest and others incidental, often referred to as nuisance parameters. The asymptotic properties of the ML estimator extend rather naturally, and constitute multidimensional counterparts of (3.60) and (3.61). Establishing these properties is technically more involved than in the single parameter case (Lehmann and Casella, 1998), so only an informal account will be given here. Similar to (3.50), the score vector $\mathbf{l}'\left(\widehat{\boldsymbol{\theta}}_n\right)$ evaluated at $\widehat{\boldsymbol{\theta}}_n$, the ML estimator based on n i.i.d. samples, can be expanded about the true parameter value $\boldsymbol{\theta}_0$ as

$$\mathbf{l}'\left(\widehat{\boldsymbol{\theta}}_n\right) \approx \mathbf{l}'\left(\boldsymbol{\theta}_0\right) + \left[\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right).$$

Derivatives higher than second-order are ignored in the above expansion. The score vector must be null when evaluated at a stationary point, so rearranging and multiplying both sides by $\sqrt{n}$ yields

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \approx \left[-\frac{1}{n}\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \frac{1}{\sqrt{n}}\mathbf{l}'\left(\boldsymbol{\theta}_0\right). \tag{3.62}$$

The random vector:

$$\frac{1}{\sqrt{n}}\mathbf{l}'\left(\boldsymbol{\theta}_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{l}_i'\left(\boldsymbol{\theta}_0\right) \overset{D}{\to} N\left[\mathbf{0}, \mathbf{I}_1\left(\boldsymbol{\theta}_0\right)\right] \tag{3.63}$$

by the central limit theorem, because it involves the sum of many i.i.d. random score vectors $\mathbf{l}_i'\left(\boldsymbol{\theta}_0\right)$. Likewise, as $n \to \infty$, the random matrix

$$\left[-\frac{1}{n}\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left\{\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial^2 l_i\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]\right\}^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \to \mathbf{I}_1^{-1}\left(\boldsymbol{\theta}_0\right).$$

Hence, because $n^{-1/2}\mathbf{l}'\left(\boldsymbol{\theta}_0\right)$ converges in distribution to an $N\left[\mathbf{0}, \mathbf{I}_1\left(\boldsymbol{\theta}\right)\right]$ process and

$$\left[-\frac{1}{n}\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

converges to the constant $\mathbf{I}_1^{-1}\left(\boldsymbol{\theta}_0\right)$, the multivariate version of Slutsky's theorem (Casella and Berger, 1990) yields

$$\left[-\frac{1}{n}\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \frac{1}{\sqrt{n}}\mathbf{l}'\left(\boldsymbol{\theta}_0\right) \overset{D}{\to} \mathbf{I}_1^{-1}\left(\boldsymbol{\theta}_0\right) N\left[\mathbf{0}, \mathbf{I}_1\left(\boldsymbol{\theta}\right)\right]. \tag{3.64}$$

Using this in (3.62), it follows that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{D} N\left[\mathbf{0}, \mathbf{I}_1^{-1}\left(\boldsymbol{\theta}_0\right)\right]. \tag{3.65}$$

Hence, in large samples, the ML estimator can be written (informally) as

$$\widehat{\boldsymbol{\theta}}_n \sim N\left[\boldsymbol{\theta}_0, \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right], \tag{3.66}$$

so the ML estimator is asymptotically unbiased, efficient, and multivariate normal. Distribution (3.66) gives the basis for the solution of many inferential problems in genetics via large sample theory. Note, however, that the asymptotic distribution depends on the unknown $\boldsymbol{\theta}_0$. In practice, one proceeds by using the approximate distribution $\widehat{\boldsymbol{\theta}}_n \sim N[\widehat{\boldsymbol{\theta}}_n, \mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}}_n)]$. For a large sample, this may be accurate enough, in view of the consistency property of $\widehat{\boldsymbol{\theta}}_n$. It should be clear, however, that this approximation does not take into account the error associated with the estimation of $\boldsymbol{\theta}$.

3.9 Functional Invariance of Maximum Likelihood Estimators

The property of functional invariance states that if $\widehat{\theta}$ is the ML estimator of θ, then the ML estimator of the function $f(\theta)$ is $f(\widehat{\theta})$. The property is motivated using the linear regression model of Example 3.7, and a more formal presentation is given subsequently.

3.9.1 Illustration of Functional Invariance

In Example 3.7 the ML estimators of β_0 and β_1 were found to be $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. If the function β_0/β_1 exists, then its ML estimator is:

$$\frac{\widehat{\beta}_0}{\widehat{\beta}_1} = \frac{\left(\overline{y} - \widehat{\beta}_1 \overline{x}\right)\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}.$$

We proceed to verify that this is indeed the case. The log-likelihood, assuming independent sampling from a normal distribution, apart from an additive constant, is

$$l\left(\beta_0, \beta_1, \sigma^2 | \mathbf{y}\right) = -\frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

Define $\eta = \beta_0/\beta_1$, supposing the slope coefficient cannot be null. The sampling scheme remains unchanged, but the model is formulated now in terms of the one-to-one reparameterization

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \rightarrow \begin{bmatrix} \eta \\ \beta_1 \end{bmatrix}.$$

This means that it is possible to go from one parameterization to another in a unique manner. For example, a reparameterization to β_0 and β_1^2 would not be one-to-one because β_1 and $-\beta_1$ yield the same value of β_1^2. The log-likelihoods under the alternative parameterizations are related according to

$$\begin{aligned} l\left(\beta_0, \beta_1, \sigma^2 | \mathbf{y}\right) &= l\left(\eta, \beta_1, \sigma^2 | \mathbf{y}\right) \\ &= -\frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \eta\beta_1 - \beta_1 x_i\right)^2. \end{aligned}$$

The score vector under the new parameterization is

$$\begin{bmatrix} \partial l/\partial\eta \\ \partial l/\partial\beta_1 \\ \partial l/\partial\sigma^2 \end{bmatrix} = \begin{bmatrix} \dfrac{\beta_1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\left(y_i - \eta\beta_1 - \beta_1 x_i\right) \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\left(y_i - \eta\beta_1 - \beta_1 x_i\right)\left(\eta + x_i\right) \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}\left(y_i - \eta\beta_1 - \beta_1 x_i\right)^2 \end{bmatrix}.$$

Setting its three elements simultaneously to zero, and solving for the unknown parameters gives the ML estimators. From the first equation, one obtains directly

$$\widehat{\eta} = \frac{\left(\overline{y} - \widehat{\beta}_1\overline{x}\right)}{\widehat{\beta}_1}.$$

The second equation leads to

$$\begin{aligned} &\widehat{\eta}\sum_{i=1}^{n}\left[y_i - \overline{y} - \widehat{\beta}_1\left(x_i - \overline{x}\right)\right] + \sum_{i=1}^{n}\left(x_i y_i - \widehat{\eta}\widehat{\beta}_1 x_i - \widehat{\beta}_1 x_i^2\right) \\ &= \sum_{i=1}^{n}\left[x_i y_i - \left(\overline{y} - \widehat{\beta}_1\overline{x}\right)x_i - \widehat{\beta}_1 x_i^2\right] = 0, \end{aligned}$$

which, when solved for $\widehat{\beta}_1$, gives as solution

$$\widehat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

This is identical to the ML estimator found under the initial parameterization, as shown in Example 3.7. The third equation leads to the ML estimator of σ^2

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\eta}\widehat{\beta}_1 - \widehat{\beta}_1 x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

as found before. Finally, from the reparameterized model, one can estimate β_0 as $\widehat{\beta}_0 = \widehat{\eta}\widehat{\beta}_1$.

It can be verified that the ML estimators of β_0 and β_1 are unbiased. For example,

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E\left(\frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\right) \\ &= \frac{E\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \left(n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \beta_1, \end{aligned}$$

and the same is true for β_0. This should not be construed as an indication that ML estimators are always unbiased; it was seen in Example 3.8 that the ML estimator of σ^2 has a downward bias. It is not obvious how to obtain an unbiased estimator of $\eta = \beta_0/\beta_1$, but the functional invariance property of the ML estimator leads directly to ML estimation of this ratio. The estimator is biased, as shown below.

Example 3.10 *Bias of a ratio of estimators*
Suppose we wish to evaluate, in the linear regression model,

$$E\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}\right).$$

Following Goodman and Hartley (1958) and Raj (1968), one can write

$$\begin{aligned} Cov\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}, \widehat{\beta}_1\right) &= E\left(\widehat{\beta}_0\right) - E\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}\right) E\left(\widehat{\beta}_1\right) \\ &= \beta_0 - E\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}\right)\beta_1. \end{aligned}$$

Rearranging,

$$E\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}\right) = \frac{\beta_0}{\beta_1} - \frac{Cov\left(\frac{\widehat{\beta}_0}{\widehat{\beta}_1}, \widehat{\beta}_1\right)}{\beta_1},$$

which shows that unless $Cov\left(\widehat{\beta}_0/\widehat{\beta}_1, \widehat{\beta}_1\right)$ is null, the ML estimator of β_0/β_1 is biased. This bias, $E\left(\widehat{\beta}_0/\widehat{\beta}_1\right) - \beta_0/\beta_1$, expressed in units of standard deviation, is

$$\begin{aligned}\frac{Bias\left(\widehat{\beta}_0/\widehat{\beta}_1\right)}{\sqrt{Var\left(\widehat{\beta}_0/\widehat{\beta}_1\right)}} &= -\frac{Corr\left(\widehat{\beta}_0/\widehat{\beta}_1, \widehat{\beta}_1\right)\sqrt{Var\left(\widehat{\beta}_1\right)}}{\beta_1} \\ &= -Corr\left(\widehat{\beta}_0/\widehat{\beta}_1, \widehat{\beta}_1\right) CV\left(\widehat{\beta}_1\right),\end{aligned}$$

where $CV(\cdot)$ denotes coefficient of variation. Hence

$$\frac{\left|Bias\left(\widehat{\beta}_0/\widehat{\beta}_1\right)\right|}{\sqrt{Var\left(\widehat{\beta}_0/\widehat{\beta}_1\right)}} \leq CV\left(\widehat{\beta}_1\right)$$

which gives an upper bound for the absolute value of the bias in units of standard deviation. From the form of the ML estimator of β_1 given in Example 3.7, it can be deduced that

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

Thus,

$$CV\left(\widehat{\beta}_1\right) = \frac{\sigma}{\beta}\sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}}.$$

This indicates that the absolute standardized bias of the ratio estimator can be reduced by increasing the dispersion of the values of the explanatory variable x, as measured by

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2.$$

■

3.9.2 Invariance in a Single Parameter Model

In the preceding section, it was seen that for a one-to-one reparameterization, the same results are obtained irrespective of whether the likelihood function is maximized with respect to θ or $\eta = f(\theta)$. In the presentation below it is assumed that the transformation is one-to-one, in the sense that $\theta = f^{-1}(\eta)$. However, the principle of invariance can be extended to hold for any transformation (Mood et al., 1974; Cox and Hinkley, 1974).

The explanation of the principle of invariance given here is based on the fact that two alternative parameterizations must generate the same probability distribution of $\mathbf{y}$. Hence

$$\begin{aligned} &p(\mathbf{y}|\theta, \theta \in \Omega) = p\left[\mathbf{y}|f^{-1}(\eta), \quad \eta \in \Omega^*\right] \\ &= p(\mathbf{y}|\eta, \eta \in \Omega^*) = p\left[\mathbf{y}|f(\theta), \quad \theta \in \Omega\right], \end{aligned} \tag{3.67}$$

where Ω^* is the parameter space of η. A similar relationship must then hold for the log-likelihoods

$$l(\theta|\mathbf{y}) = l\left[f^{-1}(\eta)|\mathbf{y}\right] = l(\eta|\mathbf{y}) = l\left[f(\theta)|\mathbf{y}\right]. \tag{3.68}$$

Then, if $\widehat{\eta}$ is the maximizer of $l(\eta|\mathbf{y})$, it must be true that $\widehat{\eta} = f\left(\widehat{\theta}\right)$, because $l(\theta|\mathbf{y})$ and $l(\eta|\mathbf{y})$ have the same maximum and the relationship is one-to-one. Under the new parameterization, the score can be written as

$$\frac{dl(\eta|\mathbf{y})}{d\eta} = \frac{dl(\theta|\mathbf{y})}{d\eta} = \frac{dl(\theta|\mathbf{y})}{d\theta}\frac{d\theta}{d\eta}. \tag{3.69}$$

Setting this equation to zero leads to $\widehat{\eta}$, the ML estimator of η. In order to evaluate the observed information about η contained in the sample, differentiation of (3.69) with respect to η yields

$$\begin{aligned} &\frac{d^2 l(\eta|\mathbf{y})}{(d\eta)^2} = \frac{d}{d\eta}\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\frac{d\theta}{d\eta}\right] \\ &= \left\{\frac{d}{d\eta}\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]\right\}\frac{d\theta}{d\eta} + \left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]\frac{d^2\theta}{(d\eta)^2} \\ &= \left\{\left[\frac{d^2 l(\theta|\mathbf{y})}{(d\theta)^2}\right]\frac{d\theta}{d\eta}\right\}\frac{d\theta}{d\eta} + \left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]\frac{d^2\theta}{(d\eta)^2} \\ &= \left[\frac{d^2 l(\theta|\mathbf{y})}{(d\theta)^2}\left(\frac{d\theta}{d\eta}\right)^2 + \frac{dl(\theta|\mathbf{y})}{d\theta}\frac{d^2\theta}{(d\eta)^2}\right]. \end{aligned} \tag{3.70}$$

The expected information about the new parameter η contained in the sample is then

$$\begin{aligned} I(\eta) &= E\left[\frac{d^2 l(\theta|\mathbf{y})}{(d\theta)^2}\right]\left(\frac{d\theta}{d\eta}\right)^2 + E\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]\frac{d^2\theta}{(d\eta)^2} \\ &= E\left[\frac{d^2 l(\theta|\mathbf{y})}{(d\theta)^2}\right]\left(\frac{d\theta}{d\eta}\right)^2 = I(\theta)\left(\frac{d\theta}{d\eta}\right)^2 \\ &= I\left[f^{-1}(\eta)\right]\left[\frac{df^{-1}(\eta)}{d\eta}\right]^2. \end{aligned} \tag{3.71}$$

This being so because $E\left[dl(\theta|\mathbf{y})/d\theta\right] = 0$, as seen in (3.20). Now, using the general result

$$\frac{dx}{dy} = \frac{1}{dy/dx},$$

the (asymptotic) variance of the ML estimator of the transformed parameter is

$$\begin{aligned} Var(\eta) &= [I(\eta)]^{-1} \\ &= [I(\theta)]^{-1}\left[\frac{d\eta}{d\theta}\right]^2 = Var(\theta)\left[\frac{d\eta}{d\theta}\right]^2 \end{aligned} \tag{3.72}$$

which would be evaluated at $\theta = \widehat{\theta}$. Expression (3.72) can give erroneous results when f is not monotone, as shown in the following example.

Example 3.11 *A binomially distributed random variable*
For a binomially distributed random variable X (number of "successes" in n trials), the likelihood is

$$L(\theta|x,n) \propto \theta^x (1-\theta)^{n-x}$$

and the log-likelihood, ignoring an additive constant, is

$$l(\theta|x,n) = x\ln\theta + (n-x)\ln(1-\theta).$$

Setting the first differential of the log-likelihood with respect to θ equal to zero and solving for θ yields the closed-form ML estimator

$$\widehat{\theta} = \frac{x}{n}.$$

Since, by definition, X results from the sum of n independent Bernoulli trials, each with variance $\theta(1-\theta)$, the variance of $\widehat{\theta}$ is

$$Var\left(\widehat{\theta}\right) = \frac{\theta(1-\theta)}{n}.$$

Imagine that interest focuses on the odds ratio $\eta = f(\theta) = \theta/(1-\theta)$. Since

$$\frac{d\eta}{d\theta} = \frac{1}{(1-\theta)^2},$$

using (3.72), we obtain

$$Var(\widehat{\eta}) = \frac{\widehat{\theta}}{n\left(1-\widehat{\theta}\right)^3}.$$

In this case, the transformation f is one-to-one. Suppose instead that there is interest in the quantity $\omega = g(\theta) = \theta(1-\theta)$. Now,

$$\frac{d\omega}{d\theta} = 1 - 2\theta.$$

Then (3.72) yields, for the asymptotic variance of the ML of ω,

$$Var(\widehat{\omega}) = \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}\left(1-2\widehat{\theta}\right)^2,$$

which underestimates the variance rather drastically if $\widehat{\theta} = 1/2$. The source of this problem is that g is not one-to-one. A way around this problem was discussed in Section 2.2.4 of Chapter 2. ■

3.9.3 Invariance in a Multiparameter Model

Let the distribution of the random vector $\mathbf{y}$ be indexed by a parameter $\boldsymbol{\theta}$ having more than one element. Consider the one-to-one transformation $\boldsymbol{\eta} = \mathbf{f}(\boldsymbol{\theta})$ such that the inverse function $\boldsymbol{\theta} = \mathbf{f}^{-1}(\boldsymbol{\eta})$ exists, and suppose that the likelihood is differentiable with respect to $\boldsymbol{\eta}$ at least twice**.** The relationship between likelihoods given in (3.68) carries directly to the multiparameter situation. The score vector in the reparameterized model, after equation (3.69), is

$$\frac{\partial l(\boldsymbol{\eta}|\mathbf{y})}{\partial \boldsymbol{\eta}} = \frac{\partial \boldsymbol{\theta}'}{\partial \boldsymbol{\eta}} \frac{\partial l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}, \tag{3.73}$$

where $\partial\boldsymbol{\theta}'/\partial\boldsymbol{\eta} = \{\partial\theta_j/\partial\eta_i\}$ is a matrix of order $p \times p$, p is the number of elements in $\boldsymbol{\theta}$, and subscripts i and j refer to row and column, respectively. Thus, the jth row of $\partial\boldsymbol{\theta}'/\partial\boldsymbol{\eta}$ looks as follows:

$$\left[\frac{\partial\theta_1}{\partial\eta_j}, \frac{\partial\theta_2}{\partial\eta_j}, \ldots, \frac{\partial\theta_p}{\partial\eta_j}\right].$$

The expected information matrix, following (3.71), is

$$\mathbf{I}(\boldsymbol{\eta}) = \frac{\partial\boldsymbol{\theta}'}{\partial\boldsymbol{\eta}}\mathbf{I}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\eta}'} = \frac{\partial\left[\mathbf{f}^{-1}(\boldsymbol{\eta})\right]'}{\partial\boldsymbol{\eta}}\mathbf{I}\left[\mathbf{f}^{-1}(\boldsymbol{\eta})\right]\frac{\partial\left[\mathbf{f}^{-1}(\boldsymbol{\eta})\right]}{\partial\boldsymbol{\eta}'}, \tag{3.74}$$

and the expression equivalent to (3.72) is

$$Var\left(\widehat{\boldsymbol{\eta}}\right) = \left.\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} Var\left(\widehat{\boldsymbol{\theta}}\right) \left.\frac{\partial \boldsymbol{\eta}'}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{3.75}$$

Example 3.12 *A reparameterized linear regression model*
The matrix in (3.74) is illustrated using the linear regression model. The original parameterization was in terms of $\boldsymbol{\theta}' = [\beta_0, \beta_1, \sigma^2]$. The new parameterization consists of the vector

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \beta_0/\beta_1 \\ \beta_1 \\ \sigma^2 \end{bmatrix}$$

with inverse

$$\boldsymbol{\theta} = \mathbf{f}^{-1}(\boldsymbol{\eta}) = \begin{bmatrix} \eta_1\eta_2 \\ \eta_2 \\ \eta_3 \end{bmatrix}.$$

The 3×3 matrix defined after (3.73) would be

$$\frac{\partial \boldsymbol{\theta}'}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \partial\theta_1/\partial\eta_1 & \partial\theta_2/\partial\eta_1 & \partial\theta_3/\partial\eta_1 \\ \partial\theta_1/\partial\eta_2 & \partial\theta_2/\partial\eta_2 & \partial\theta_3/\partial\eta_2 \\ \partial\theta_1/\partial\eta_3 & \partial\theta_2/\partial\eta_3 & \partial\theta_3/\partial\eta_3 \end{bmatrix} = \begin{bmatrix} \eta_2 & 0 & 0 \\ \eta_1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.76}$$

Differentiating the log-likelihood twice under the new parameterization, multiplying by -1, and taking expectations, yields the information matrix

$$\mathbf{I}(\boldsymbol{\eta}) = \frac{1}{\sigma^2} \begin{bmatrix} n\eta_2^2 & \eta_2\left(n\eta_1 + \sum_{i=1}^{n} x_i\right) & 0 \\ \eta_2\left(n\eta_1 + \sum_{i=1}^{n} x_i\right) & \sum_{i=1}^{n}(\eta_1 + x_i)^2 & 0 \\ 0 & 0 & \dfrac{n}{2\eta_3} \end{bmatrix}.$$

It can be verified that the same result is obtained from the matrix product given in (3.74), employing the matrix (3.76). ■

4

Further Topics in Likelihood Inference

4.1 Introduction

This chapter continues the discussion on likelihood inference. First, two commonly used numerical methods for obtaining ML estimates when the likelihood equations do not have a closed form or are difficult to solve are introduced. (A third method, the Expectation–Maximization algorithm, often referred to as the EM-algorithm, is discussed in Chapter 9). This is followed by a discussion of the traditional test of hypotheses entrenched in the Neyman–Pearson theory. The classical first-order asymptotic distribution of the likelihood ratio is derived assuming that the standard regularity conditions are satisfied, and examples of likelihood ratio tests are given.

The presence of so-called nuisance parameters has been one of the major challenges facing the likelihood paradigm. How does one make inferences about the parameters of interest in the presence of nuisance parameters without overstating precision? Many different approaches have been suggested for dealing with this problem and some of these are briefly discussed and illustrated here. The chapter ends with examples involving the multinomial model and the analysis of ordered categorically distributed data.

4.2 Computation of Maximum Likelihood Estimates

As noted earlier, the ML equations (the score vector)

$$\mathbf{l}'(\boldsymbol{\theta}) = \mathbf{0}$$

may not have an explicit solution, so numerical methods must be used to solve them. Maximization of a likelihood function (or minimization of any induced objective function) can be viewed as a nonlinear optimization problem. A plethora of algorithms can be used for this purpose. Treatises in numerical analysis, such as Dahlquist and Björck (1974), Dennis and Schnabel (1983), and Hager (1988) can be consulted. Here, a sketch is presented of two widely used algorithms employed in connection with likelihood inference: the Newton–Raphson method and Fisher's scoring procedure.

The Newton–Raphson Procedure

Consider expanding the score vector about a trial value $\boldsymbol{\theta}^{[t]}$, where $t = 0, 1, 2, \ldots$ denotes an iterate number. A linear approximation to the score gives

$$\mathbf{l}'(\boldsymbol{\theta}) \approx \mathbf{l}'\left(\boldsymbol{\theta}^{[t]}\right) + \left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{[t]}\right).$$

In the vicinity of a stationary point, it must be true that $\mathbf{l}'(\boldsymbol{\theta}) \approx \mathbf{0}$. Using this in the preceding expression, and solving for $\boldsymbol{\theta}$ at each step t, gives the sequence of updated values

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} + \left[-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}} \mathbf{l}'\left(\boldsymbol{\theta}^{[t]}\right), \quad t = 0, 1, 2, \ldots. \tag{4.1}$$

The difference $\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}$ is called a correction. The iterative process is continued until the correction is null, corresponding to the situation where $\mathbf{l}'(\boldsymbol{\theta}^{[t]}) = \mathbf{0}$. This procedure is known as the Newton–Raphson algorithm; Fisher's scoring method uses $\mathbf{I}(\boldsymbol{\theta})$ instead of the observed information matrix $-\partial^2 l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'$. Typically, the method of scoring requires fewer calculations, because many expressions vanish or simplify in the process of taking expectations. However, it may converge at a slower rate. The two methods may not converge at all and, even if they do, there is no assurance that a global maximum would be located. This is not surprising, as the methods search for stationarity without reference to the possible existence of multiple maxima. This is a potential problem in models having many parameters, and it is expected to occur more frequently when sample sizes are small, as the likelihood may have several "peaks and valleys".

Quadratic Convergence Property

The Newton–Raphson procedure has the property of converging quadratically to a stationary point. Dahlquist and Björck (1974) provide a proof for the single parameter case, and this is presented in more detail here.

Let the root of $l'(\theta) = 0$ be $\widehat{\theta}$. Hence, $l''(\widehat{\theta})$ cannot be a null matrix and $l''(\theta)$ cannot be null for all θ near (in some sense) $\widehat{\theta}$. Let the error of the trial values at iterates t and $t+1$ be $\epsilon^{[t]} = \theta^{[t]} - \widehat{\theta}$ and $\epsilon^{[t+1]} = \theta^{[t+1]} - \widehat{\theta}$, respectively. A second-order Taylor series expansion of the score evaluated at $\widehat{\theta}$ (which must be equal to zero, because $\widehat{\theta}$ is a root), about iterate value $\theta^{[t]}$, gives

$$l'\left(\widehat{\theta}\right) \approx l'\left(\theta^{[t]}\right) + l''\left(\theta^{[t]}\right)\left(\widehat{\theta} - \theta^{[t]}\right) + \tfrac{1}{2} l'''\left(\theta^{[t]}\right)\left(\widehat{\theta} - \theta^{[t]}\right)^2.$$

Because $l''\left(\theta^{[t]}\right)$ is not null, one can write

$$\widehat{\theta} - \theta^{[t]} + \frac{l'\left(\theta^{[t]}\right)}{l''\left(\theta^{[t]}\right)} = \frac{-\frac{1}{2} l'''\left(\theta^{[t]}\right)\left(\widehat{\theta} - \theta^{[t]}\right)^2}{l''\left(\theta^{[t]}\right)}. \tag{4.2}$$

Now, because of the form of the Newton–Raphson iteration in (4.1), expression (4.2) is equivalent to

$$\widehat{\theta} - \theta^{[t+1]} = \frac{-\frac{1}{2} l'''\left(\theta^{[t]}\right)\left(\widehat{\theta} - \theta^{[t]}\right)^2}{l''\left(\theta^{[t]}\right)}.$$

Writing the above in terms of the errors about the root, one obtains

$$\epsilon^{[t+1]} = \frac{l'''\left(\theta^{[t]}\right)}{2l''\left(\theta^{[t]}\right)}\left(\epsilon^{[t]}\right)^2. \tag{4.3}$$

This indicates that the error at iterate t is proportional to the square of the error at the preceding iteration, so the method is said to be quadratically convergent. Now

$$\lim_{\theta^{[t]} \to \widehat{\theta}} \left[\frac{\epsilon^{[t+1]}}{\left(\epsilon^{[t]}\right)^2}\right] = \lim_{\theta^{[t]} \to \widehat{\theta}} \left[\frac{l'''\left(\theta^{[t]}\right)}{2l''\left(\theta^{[t]}\right)}\right] = \frac{l'''\left(\widehat{\theta}\right)}{2l''\left(\widehat{\theta}\right)}. \tag{4.4}$$

Also,

$$\lim_{\theta^{[t]} \to \widehat{\theta}} \left[\left|\frac{\epsilon^{[t+1]}}{\left(\epsilon^{[t]}\right)^2}\right|\right] = \lim_{\theta^{[t]} \to \widehat{\theta}} \left[\frac{\left|\epsilon^{[t+1]}\right|}{\left(\epsilon^{[t]}\right)^2}\right] = \frac{1}{2}\frac{\left|l'''\left(\widehat{\theta}\right)\right|}{\left|l''\left(\widehat{\theta}\right)\right|} = C$$

and C, called the "asymptotic error constant", will be nonnull whenever $l'''\left(\widehat{\theta}\right) \neq 0$. For a sample of size n, the second derivative of the log-likelihood, with respect to θ, is equal to

$$l''(\theta) = \sum_{i=1}^{n} \frac{d^2}{(d\theta)^2} \ln p(y_i|\theta)$$

and, for large n, this converges in probability to $-I(\theta) = -nI_1(\theta)$. Hence, C goes to zero as the sample size increases, provided the third derivatives are bounded; recall that this was an assumption made when proving consistency of the ML estimator. This suggests a relatively faster convergence rate at larger values of n, given that the algorithm converges, as assumed here.

The Newton–Raphson algorithm always converges to a root of the ML equation provided that the starting value $\theta^{[0]}$ is sufficiently close to such a root (Dahlquist and Björck, 1974). Using (4.3),

$$\widehat{\theta} - \theta^{[1]} = \left(\widehat{\theta} - \theta^{[0]}\right) \left[\left(\widehat{\theta} - \theta^{[0]}\right) \frac{l'''\left(\theta^{[0]}\right)}{2l''\left(\theta^{[0]}\right)}\right].$$

Thus, if

$$\left|\left(\widehat{\theta} - \theta^{[0]}\right) \frac{l'''\left(\theta^{[0]}\right)}{2l''\left(\theta^{[0]}\right)}\right| < 1$$

the next approximation is closer to the root than the starting value. Let

$$\frac{1}{2}\left|\frac{l'''(\theta)}{l''(\theta)}\right| \leq m \quad \text{for all } \theta \in I$$

with $m > 0$ and where I is a region near $\widehat{\theta}$, and suppose that

$$\left|\left(\widehat{\theta} - \theta^{[0]}\right) m\right| = \left|\epsilon^{[0]} m\right| < 1.$$

Then, from (4.3),

$$\left|\epsilon^{[t+1]}\right| = \left|\frac{l'''\left(\theta^{[t]}\right)}{2l''\left(\theta^{[t]}\right)}\right| \left(\epsilon^{[t]}\right)^2 \leq m\left(\epsilon^{[t]}\right)^2$$

so

$$\left|m\epsilon^{[t+1]}\right| \leq \left(m\epsilon^{[t]}\right)^2,$$

and

$$\begin{aligned}\left|\epsilon^{[t+1]}\right| &\leq \frac{1}{m}\left(m\epsilon^{[t]}\right)^2 = \frac{1}{m}\left|m\epsilon^{[t]}\right|^2 \\ &\leq \frac{1}{m}\left(m\epsilon^{[t-1]}\right)^{2^2} = \frac{1}{m}\left|m\epsilon^{[t-1]}\right|^{2^2} \\ &\leq \frac{1}{m}\left(m\epsilon^{[t-2]}\right)^{2^3} = \frac{1}{m}\left|m\epsilon^{[t-2]}\right|^{2^3} \\ &\leq \frac{1}{m}\left(m\epsilon^{[0]}\right)^{2^{t+1}}. \end{aligned} \tag{4.5}$$

Therefore, as $t \to \infty$, then $\left|\epsilon^{[t+1]}\right| \to 0$, and the algorithm converges toward $\widehat{\theta}$. Similarly, let $\theta^{[0]} = \widehat{\theta}+\epsilon^{[0]}$, $\theta^{[1]} = \widehat{\theta}+\epsilon^{[1]}, \ldots, \theta^{[t]} = \widehat{\theta}+\epsilon^{[t]}$ be the sequence of iterates produced by the Newton–Raphson algorithm. If $\theta^{[0]}$ belongs to the region I, this is equivalent to the statement

$$\left[\widehat{\theta} - \left|\epsilon^{[0]}\right|, \widehat{\theta} + \left|\epsilon^{[0]}\right|\right] \in I. \tag{4.6}$$

The next iterate, $\theta^{[1]}$, must fall in the set

$$I_1 = \left[\widehat{\theta} - \left|\epsilon^{[1]}\right|, \widehat{\theta} + \left|\epsilon^{[1]}\right|\right]. \tag{4.7}$$

Now, by (4.5),

$$\widehat{\theta} - \left|\epsilon^{[1]}\right| \geq \widehat{\theta} - \frac{1}{m}\left(m\epsilon^{[0]}\right)^2 = \widehat{\theta} - m\left|\epsilon^{[0]}\right|^2 = \widehat{\theta} - \left|m\epsilon^{[0]}\right|\left|\epsilon^{[0]}\right|.$$

By assumption, $\left|m\epsilon^{[0]}\right| < 1$, so it must be true that

$$\widehat{\theta} - \left|\epsilon^{[1]}\right| \geq \widehat{\theta} - \left|\epsilon^{[0]}\right|. \tag{4.8}$$

Similarly,

$$\widehat{\theta} + \left|\epsilon^{[1]}\right| \leq \widehat{\theta} + \frac{1}{m}\left(m\epsilon^{[0]}\right)^2 = \widehat{\theta} + \left|m\epsilon^{[0]}\right|\left|\epsilon^{[0]}\right|$$

so

$$\widehat{\theta} + \left|\epsilon^{[1]}\right| \leq \widehat{\theta} + \left|\epsilon^{[0]}\right|. \tag{4.9}$$

From (4.8) and (4.9)

$$\widehat{\theta} - \left|\epsilon^{[0]}\right| \leq \widehat{\theta} - \left|\epsilon^{[1]}\right| < \widehat{\theta} + \left|\epsilon^{[1]}\right| \leq \widehat{\theta} + \left|\epsilon^{[0]}\right|.$$

Then the region of values I_1 of the first iterate is such that $I_1 \subset I$. Using this argument repeatedly leads to

$$I_t \subset I_{t-1} \subset I_{t-2} \subset \cdots \subset I_2 \subset I_1 \subset I. \tag{4.10}$$

This indicates that all iterates stay within the initial region, and that as $t \to \infty$, I_∞ should contain a single value, $\widehat{\theta}$, proving convergence of the iterative scheme (provided iteration starts in the vicinity of the root).

4.3 Evaluation of Hypotheses

A frequently encountered problem is the need for evaluating or testing a hypothesis about the state of a biological system. In genetics, for example, in a large randomly mated population, in the absence of mutation, migration, or selection, genotypic frequencies are expected to remain constant generation after generation. This is called the Hardy–Weinberg equilibrium law (i.e., Crow and Kimura, 1970) and it may be of interest to test if this hypothesis holds, given a body of data. A discussion of how such tests can be constructed from a frequentist point of view based on a ML analysis is presented below. The presentation is introductory, and only classical, first-order asymptotic results are discussed. This field has been undergoing rapid developments. The reader is referred to books such as Barndorff-Nielsen and Cox (1994) and Severini (2000) for an account of the modern approach to likelihood inference.

4.3.1 Likelihood Ratio Tests

The use of ratios between likelihoods obtained under different models or hypotheses is widespread in genetics. The theoretical basis of tests constructed from likelihood ratios was developed by Neyman and Pearson (1928). The asymptotic distribution of twice the logarithm of a maximum likelihood ratio statistic was derived by Wilks (1938). Cox and Hinkley (1974) and Stuart and Ord (1991) give a heuristic account of this classical theory, and this is followed closely here. In order to pose the problem in a sufficiently general framework, partition the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix},$$

where $\boldsymbol{\theta}_1$ is an $r \times 1$ vector of parameters involved in the hypothesis, having at least one element, and $\boldsymbol{\theta}_2$ is an $s \times 1$ vector of supplementary or nuisance parameters, with zero or more components. Let $p = r + s$. Some hypotheses to be contrasted can be formulated as

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$$

versus

$$H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10},$$

where $\boldsymbol{\theta}_{10}$ is the value of the parameter under the null hypothesis. It is important here to note the nesting structure of $\boldsymbol{\theta}$: in one of the hypotheses or models, elements of $\boldsymbol{\theta}$ take a fixed value, but are free to vary in the other.

The likelihood ratio test is based on the fact that the likelihood maximized under H_1 must be at least as large as that under H_0. This is so

because there is always the possibility that there may be a higher likelihood at values of the parameters violating the restrictions imposed by the null hypothesis. Let

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\boldsymbol{\theta}}_1 \\ \widehat{\boldsymbol{\theta}}_2 \end{bmatrix}$$

be the unrestricted ML estimator, and let

$$\widetilde{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\theta}_{10} \\ \widetilde{\boldsymbol{\theta}}_2 \end{bmatrix}$$

be the estimator under H_0. In general, $\widetilde{\boldsymbol{\theta}}_2$, which is the ML estimator of $\boldsymbol{\theta}_2$, for fixed $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$, will not coincide with $\widehat{\boldsymbol{\theta}}_2$. The ratio of maximized likelihoods is

$$LR = \frac{L\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2 | \mathbf{y}\right)}{L\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 | \mathbf{y}\right)}, \quad 0 \leq LR \leq 1. \tag{4.11}$$

Values of LR close to 1 suggest plausibility of H_0, because the restriction imposed by the hypothesis does not lower the likelihood in an appreciable manner. Since LR is a function of $\mathbf{y}$, it is a random variable having some sampling distribution. Now, because large values of LR suggest that H_0 holds, one would reject the null hypothesis whenever the LR is below a critical threshold t. For a test with rejection rate (under H_0) α, the value of the threshold is given by the solution to the integral equation (Stuart and Ord, 1991)

$$\int_0^{t_\alpha} p\left(LR\right) dLR = \alpha, \tag{4.12}$$

where $p\left(LR\right)$ is the density of the distribution of the likelihood ratio under H_0. In general, this distribution is unrecognizable and must be approximated. However, there are instances in which this distribution can be identified. Examples of the two situations follow.

Before embarking on these examples, we mention briefly an alternative, "pure likelihoodist", approach to inference. Arguments in favor of this approach can be found in Edwards (1992) and in Royall (1997). Rather than maximizing the likelihood and studying the distribution of the ML estimator in conceptual replications, adherents to this school draw inferences from the likelihood or from the likelihood ratio only, with the data fixed. The justification is to be found in what Hacking (1965) termed the law of likelihood:

> "If one hypothesis, H_1, implies that a random variable X takes the value x with probability $f_1\left(x\right)$, while another hypothesis, H_2, implies that the probability is $f_2\left(x\right)$, then the observation $X = x$ is evidence supporting H_1 over H_2 if $f_1\left(x\right) > f_2\left(x\right)$,

and the likelihood ratio $f_1(x)/f_2(x)$ measures the strength of that evidence."

Likelihood ratios close to 1 represent weak evidence, and extreme ratios represent strong evidence. The problem is finding a benchmark value k that would give support to H_1 over H_2 analogous to the traditional p values equal to 0.05 and 0.01. Values of $k = 8$ (representing "fairly strong") and $k = 32$ (representing "strong") have been proposed and Royall (1997) gives a rationale for these choices. We will not pursue this subject further. Instead the reader is referred to the works mentioned above, where arguments in favor of this approach can be found.

Example 4.1 *Likelihood ratio in a Gaussian linear model*
Suppose that a vector of observations is drawn from the multivariate distribution

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\sigma^2\right),$$

where $\boldsymbol{\beta}(p \times 1)$ and σ^2 are unknown parameters and $\mathbf{V}$ is a known matrix. A hypothesis of interest may be H_0: $\boldsymbol{\beta} = \boldsymbol{\alpha}$, so σ^2 is the incidental parameter here. The likelihood under H_1 is

$$L\left(\boldsymbol{\beta},\sigma^2|\mathbf{y}\right) = (2\pi)^{-\frac{N}{2}}\left|\mathbf{V}\sigma^2\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right],$$

where N is the order of $\mathbf{y}$. In the absence of any restriction, the ML estimators can be found to be

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

and

$$\widehat{\sigma^2} = \frac{1}{N}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

The likelihood under H_0 is

$$L\left(\sigma^2|\boldsymbol{\beta}=\boldsymbol{\alpha},\mathbf{y}\right) = (2\pi)^{-\frac{N}{2}}\left|\mathbf{V}\sigma^2\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)\right]$$

and the corresponding ML estimator of σ^2 is

$$\widehat{\widehat{\sigma}}^2 = \frac{1}{N}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right).$$

Note that

$$\begin{aligned}
&\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)\\
&= \left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{X}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right)\right]'\mathbf{V}^{-1}\left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{X}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right)\right]\\
&= \left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)+\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right)\\
&= N\widehat{\sigma^2}+\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}\right).
\end{aligned}$$

Hence

$$\widehat{\widehat{\sigma}}^2 = \widehat{\sigma^2} + \frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)}{N}.$$

Using (4.11) the ratio of maximized likelihoods is then

$$\begin{aligned} LR &= \frac{L\left(\widehat{\widehat{\sigma}}^2 | \boldsymbol{\beta} = \boldsymbol{\alpha}, \mathbf{y}\right)}{L\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2} | \mathbf{y}\right)} \\ &= \left[1 + \frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)}{N\widehat{\sigma^2}}\right]^{-\frac{N}{2}}. \end{aligned}$$

Now, under the null hypothesis (Searle, 1971),

$$\frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\right)}{\sigma^2} \sim \chi^2_p$$

and, further, this random variable can be shown to be distributed independently of

$$\frac{N\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{N-p}.$$

Using these results, the likelihood ratio is expressible as

$$LR = \left[1 + \frac{\sigma^2\chi^2_p}{\sigma^2\chi^2_{N-p}}\right]^{-\frac{N}{2}} = \left[1 + \frac{p\frac{\chi^2_p}{p}}{(N-p)\frac{\chi^2_{N-p}}{N-p}}\right]^{-\frac{N}{2}}.$$

In addition (Searle, 1971),

$$F_{p,N-p} = \frac{\chi^2_p/p}{\chi^2_{N-p}/(N-p)}$$

defines an F-distributed random variable with p and $n-p$ degrees of freedom. Hence

$$LR = \left[1 + \frac{p}{(N-p)} F_{p,N-p}\right]^{-\frac{N}{2}}.$$

Thus, we see that the LR decreases monotonically as F increases, so

$$\Pr\left(LR \leqq t_\alpha\right) = \Pr\left(F \geqq F_{\alpha,p,N-p}\right),$$

where $F_{\alpha,p,N-p}$ is a critical value defined by

$$\int_0^{F_{\alpha,p,N-p}} p\left(F_{p,N-p}\right) dF = 1 - \alpha.$$

Here the distribution of the LR statistic is known exactly. ■

Example 4.2 *The Behrens–Fisher problem*
Suppose samples are drawn from two populations having distinct variances σ_1^2 and σ_2^2. The sampling model is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X}_1\boldsymbol{\beta}_1 \\ \mathbf{X}_2\boldsymbol{\beta}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{I}_1\sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2\sigma_2^2 \end{bmatrix} \right).$$

A hypothesis of interest may be: H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$. Under $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ the likelihood is

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2 | \mathbf{y}\right) \propto \prod_{i=1}^{2} \left(\sigma_i^2\right)^{-\frac{N_i}{2}} \exp\left[-\frac{1}{2\sigma_i^2} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i\right)' \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i\right) \right],$$

where N_i is the order of the data vector $\mathbf{y}_i$. The maximizers of the unrestricted likelihood are

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i &= \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i{}'\mathbf{y}_i, \quad i = 1, 2, \\ \widehat{\sigma^2} &= \frac{1}{N_i} \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)' \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right), \quad i = 1, 2, \end{aligned}$$

and the maximized likelihood is

$$L\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\sigma_1^2}, \widehat{\sigma_2^2} | \mathbf{y}\right) \propto \prod_{i=1}^{2} \left(\widehat{\sigma_i^2}\right)^{-\frac{N_i}{2}} \exp\left[-\frac{N_i}{2} \right].$$

Under H_0:

$$L\left(\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2 | \mathbf{y}\right) \propto \prod_{i=1}^{2} \left(\sigma_i^2\right)^{-\frac{N_i}{2}} \exp\left[-\frac{1}{2\sigma_i^2} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right) \right].$$

After differentiation of the log-likelihood with respect to the parameters, setting the resulting equations to zero gives the system

$$\begin{aligned} \widehat{\widehat{\sigma}}_i^2 &= \frac{1}{N_i} \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\widehat{\boldsymbol{\beta}}}\right)' \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\widehat{\boldsymbol{\beta}}}\right), \quad i = 1, 2, \\ \widehat{\widehat{\boldsymbol{\beta}}} &= \left(\sum_{i=1}^{2} \frac{\mathbf{X}_i'\mathbf{X}_i}{\widehat{\widehat{\sigma}}_i^2} \right)^{-1} \left(\sum_{i=1}^{2} \frac{\mathbf{X}_i{}'\mathbf{y}_i}{\widehat{\widehat{\sigma}}_i^2} \right) \end{aligned}$$

which is not explicit in $\widehat{\widehat{\boldsymbol{\beta}}}$, so it must be solved iteratively. The likelihood ratio is

$$LR = \prod_{i=1}^{2} \left[\frac{\widehat{\widehat{\sigma}}_i^2}{\widehat{\sigma^2}} \right]^{-\frac{N_i}{2}} = \prod_{i=1}^{2} \left[\frac{\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\widehat{\boldsymbol{\beta}}}\right)' \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\widehat{\boldsymbol{\beta}}}\right)}{\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i\right)' \left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i\right)} \right]^{-\frac{N_i}{2}}.$$

Because $\widehat{\widehat{\boldsymbol{\beta}}}$ cannot be written explicitly, the distribution of the LR is difficult or impossible to arrive at without using approximations. ■

Approximating the Distribution of the Likelihood Ratio

The derivation of the asymptotic distribution of the likelihood ratio test presented below is based on first-order asymptotic results where Taylor expansions play a central role. Here it is important to respect the conditions for the differentiability of functions and the rate of convergence of the terms that are ignored relative to the rate of convergence of those that are kept. The reader is referred to Lehmann and Casella (1998) for a careful treatment of this subject.

In (4.11) the likelihood ratio was defined as

$$LR = \frac{L\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right)}{L\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right)}. \tag{4.13}$$

Minus twice the log-likelihood ratio (sometimes called the deviance) is equal to

$$\begin{aligned} -2\ln LR &= -2\left[l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right) - l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right)\right] \\ &= 2\left[l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right) - l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right)\right]. \end{aligned} \tag{4.14}$$

In these expressions, $\widetilde{\boldsymbol{\theta}}_2$ is the ML estimator of $\boldsymbol{\theta}_2$ under H_0 and $\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2\right)$ are the ML estimators of $\boldsymbol{\theta}$ under H_1.

The asymptotic properties of (4.14) are derived as follows. First, expand the log-likelihood of $\boldsymbol{\theta}$, $l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)$, in a Taylor series about the ML estimates $\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2\right)$. Since first derivatives evaluated at the ML estimates $\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2\right)$ are zero, this yields

$$\begin{aligned} l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right) &\approx l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right) \\ &+ \frac{1}{2}\left(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1\right)' \left.\frac{\partial^2 l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial\boldsymbol{\theta}_1\,\partial\boldsymbol{\theta}_1'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \left(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1\right) \\ &+ \frac{1}{2}\left(\boldsymbol{\theta}_2 - \widehat{\boldsymbol{\theta}}_2\right)' \left.\frac{\partial^2 l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial\boldsymbol{\theta}_2\,\partial\boldsymbol{\theta}_2'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \left(\boldsymbol{\theta}_2 - \widehat{\boldsymbol{\theta}}_2\right) \\ &+ \left(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1\right)' \left.\frac{\partial^2 l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial\boldsymbol{\theta}_1\,\partial\boldsymbol{\theta}_2'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \left(\boldsymbol{\theta}_2 - \widehat{\boldsymbol{\theta}}_2\right). \end{aligned} \tag{4.15}$$

Define

$$-\left.\frac{\partial^2 l\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j|\mathbf{y}\right)}{\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_j'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{J}_{ij}\left(\widehat{\boldsymbol{\theta}}\right), \tag{4.16}$$

which is the ijth block of the observed information matrix $\mathbf{J}\left(\widehat{\boldsymbol{\theta}}\right)$. (This should not be confused with a similar symbol used to define the Jacobian

in (2.4) and (2.30)). We will need

$$\begin{bmatrix} \mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{J}^{12}\left(\widehat{\boldsymbol{\theta}}\right) \\ \mathbf{J}^{21}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{J}^{22}\left(\widehat{\boldsymbol{\theta}}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{11}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{J}_{12}\left(\widehat{\boldsymbol{\theta}}\right) \\ \mathbf{J}_{21}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right) \end{bmatrix}^{-1}$$
$$= \left[\mathbf{J}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1},$$

which is an expression for the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$. Substituting (4.16) in (4.15) gives

$$l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) \approx l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 | \mathbf{y}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)' \mathbf{J}_{11}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)$$
$$-\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right)' \mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right) - \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)' \mathbf{J}_{12}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right). \tag{4.17}$$

Under H_1, (4.17) evaluated at $\widehat{\boldsymbol{\theta}}$ is equal to $l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 | \mathbf{y}\right)$.

The log-likelihood (4.17) under H_0, which is the log-likelihood for $\boldsymbol{\theta}_2$, with $\boldsymbol{\theta}_1$ fixed at $\boldsymbol{\theta}_{10}$, is equal to

$$l\left(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2 | \mathbf{y}\right) \approx l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 | \mathbf{y}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)' \mathbf{J}_{11}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)$$
$$-\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right)' \mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right) - \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)' \mathbf{J}_{12}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right). \tag{4.18}$$

It is easy to obtain $\widetilde{\boldsymbol{\theta}}_2$, the ML estimator of $\boldsymbol{\theta}_2$, given $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$, from (4.18). Taking partial derivatives with respect to $\boldsymbol{\theta}_2$, setting these equal to zero and solving for $\boldsymbol{\theta}_2$ yields

$$\widetilde{\boldsymbol{\theta}}_2 = \widehat{\boldsymbol{\theta}}_2 + \left[\mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} \mathbf{J}_{21}\left(\widehat{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right), \tag{4.19}$$

whose asymptotic covariance matrix is equal to

$$Var\left(\widetilde{\boldsymbol{\theta}}_2\right) = \mathbf{J}^{22}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{J}^{21}\left(\widehat{\boldsymbol{\theta}}\right)\left[\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} \mathbf{J}^{12}\left(\widehat{\boldsymbol{\theta}}\right)$$
$$= \left[\mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}.$$

After some algebra, the log-likelihood (4.18), evaluated at $\boldsymbol{\theta}_2 = \widetilde{\boldsymbol{\theta}}_2$, can be shown to be equal to

$$l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2 | \mathbf{y}\right) \approx l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 | \mathbf{y}\right)$$
$$-\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)' \left[\mathbf{J}_{11}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{J}_{12}\left(\widehat{\boldsymbol{\theta}}\right)\left(\mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\right)^{-1} \mathbf{J}_{21}\left(\widehat{\boldsymbol{\theta}}\right)\right]$$
$$\times \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right), \tag{4.20}$$

which is not a function of $\boldsymbol{\theta}_2$. Recalling that

$$\mathbf{J}_{11}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{J}_{12}\left(\widehat{\boldsymbol{\theta}}\right)\left(\mathbf{J}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\right)^{-1}\mathbf{J}_{21}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1},$$

(4.20) can be written

$$l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right) \approx l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)'\left[J^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right). \tag{4.21}$$

It follows that $-2\ln LR$ has the form

$$\begin{aligned} &2\left[l\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2|\mathbf{y}\right) - l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right)\right] \\ &\approx \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)'\left[\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) \\ &= \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)'\left[n\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right). \end{aligned} \tag{4.22}$$

As $n \to \infty$, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)$ converges to $N\left(\mathbf{0}, \mathbf{I}^{11}\left(\boldsymbol{\theta}\right)\right)$ and $n\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)$ to $\mathbf{I}^{11}\left(\boldsymbol{\theta}\right)$, the covariance matrix of the limiting marginal distribution of

$$\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right).$$

Therefore, as $n \to \infty$, $-2\ln LR$ converges to a chi-square distribution, with noncentrality parameter equal to zero under H_0, and with $r = \dim\left(\boldsymbol{\theta}_1\right)$ degrees of freedom

$$-2\ln\left(LR\right)|H_0 \sim \chi^2_r. \tag{4.23}$$

Note that this limiting chi-square distribution does not involve the nuisance parameter $\boldsymbol{\theta}_2$.

The test criterion decreases monotonically as the LR increases. Thus, if χ^2_r exceeds a certain critical value, this corresponds to a significant lowering of the likelihood under H_0 and, thus, to rejection. If H_1 holds

$$-2\ln\left(LR\right)|H_1 \sim \chi^2_{r,\lambda} \tag{4.24}$$

where $\lambda = \left[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\right]'\left[\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}\left[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\right]$ is a noncentrality parameter (this is clearly null when $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$). When $s > 0$ (the number of nuisance parameters) the computation of LR requires two maximizations: one under H_0 and another under H_1.

The developments in this section make it clear that a comparison between two models by means of the likelihood ratio test assumes a common parameter $\boldsymbol{\theta}$ that is allowed to take specific values under either model. The test does not make sense otherwise. There must be a nested structure whereby the reduced model is embedded under the unrestricted model. If this requirement is not satisfied, the asymptotic theory described does not

hold. A modification of the likelihood ratio criterion for dealing with tests involving nonnested models has been proposed by Cox (1961, 1962); more recent and important generalizations can be found in Vuong (1989) and Lo et al. (2001).

The asymptotic results presented above assume asymptotic normality of the ML estimator. Recently, Fan et al. (2000) provided a proof of the asymptotic chi-square distribution of the log-likelihood ratio statistic, for cases where the ML estimator is not asymptotically normal.

As a final warning, with many nuisance parameters, notably when their number is large relative to the number of observations, use of the above theory to discriminate between models can give misleading results.

Power of the Likelihood Ratio Test

When designing experiments, it is important to assess the power of the test. This is the probability of rejecting H_0, given that H_1 is true or, equivalently, of accepting that the parameter value is $\boldsymbol{\theta}_1$, instead of $\boldsymbol{\theta}_{10}$. For example, a genetic experiment may be carried out to evaluate linkage between a genetic marker and a QTL. A design question could be: How many individuals need to be scored such that the hypothesis that the recombination rate is, say, equal to 10% $(\boldsymbol{\theta}_{10})$, is rejected with probability P? Under H_0, the noncentrality parameter is 0, and the probability of rejection for a test of size α would be computed using (4.12) and (4.23) as

$$\Pr\left(\text{rejecting } H_0 | H_0\right) = \int_{t_\alpha}^{\infty} p\left(\chi_1^2\right) d\chi_1^2 = \alpha, \tag{4.25}$$

where t_α is a critical value of a central chi-square distribution on one degree of freedom. If H_1 holds, the power would be computed as

$$\Pr\left(\text{rejecting } H_0 | H_1\right) = \int_{t_\alpha}^{\infty} p\left(\chi_{1,\lambda}^2\right) d\chi_{1,\lambda}^2. \tag{4.26}$$

For the linkage experiment considered here, the noncentrality parameter would be $\lambda = (\theta - 0.10)^2 I(\theta)$ and the calculation can be carried out for any desired θ.

Example 4.3 *Likelihood ratio test of Hardy–Weinberg proportions*
Consider a segregating locus with two alleles: B and b. Suppose there are three observable phenotypes corresponding to individuals with genotypes BB, Bb, and bb. A random sample is drawn from a population. The data consist of the vector $\mathbf{y}' = [y_{BB}, y_{Bb}, y_{bb}]$, where y_{BB}, y_{Bb}, and y_{bb} are the observed number of individuals with genotype BB, Bb, and bb, respectively, in a sample of size $n = y_{BB} + y_{Bb} + y_{bb}$. Let the unknown probabilities of drawing a BB, Bb, or bb individual be θ_{BB}, θ_{Bb}, and θ_{bb}, respectively. If the sample size is fixed by design, and individuals are drawn independently and with replacement, it may be reasonable to compute the probability of

observing $\mathbf{y}$ using the multinomial distribution. This, of course, would ignore knowledge about family aggregation that may exist in the population. For example, if some parents are both bb, their progeny must be bb, necessarily, so the random sampling model discussed here would not take this into account. Under multinomial sampling, the probability of observing $\mathbf{y}$ is

$$p\left(y_{BB}, y_{Bb}, y_{bb}\right) = \frac{y!}{y_{BB}!\, y_{Bb}!\, y_{bb}!} \left(\theta_{BB}\right)^{y_{BB}} \left(\theta_{Bb}\right)^{y_{Bb}} \left(\theta_{bb}\right)^{y_{bb}}$$

so the likelihood is

$$L\left(\theta_{BB}, \theta_{Bb}, \theta_{bb}\right) \propto \left(\theta_{BB}\right)^{y_{BB}} \left(\theta_{Bb}\right)^{y_{Bb}} \left(\theta_{bb}\right)^{y_{bb}}.$$

The model imposes the parametric restriction $\theta_{BB} + \theta_{Bb} + \theta_{bb} = 1$ so there are only two "free" parameters governing the distribution. Also, $y_{bb} = n - y_{BB} - y_{Bb}$, say. The likelihood is then reexpressible as

$$L\left(\theta_{BB}, \theta_{Bb}\right) \propto \theta_{BB}^{y_{BB}} \theta_{Bb}^{y_{Bb}} \left(1 - \theta_{BB} - \theta_{Bb}\right)^{n - y_{BB} - y_{Bb}}.$$

Differentiation of the log-likelihood with respect to the two θ's, and setting the derivatives to zero, gives the relationships

$$\widehat{\theta}_{BB} = \frac{y_{BB}}{n - y_{BB} - y_{Bb}} \left(1 - \widehat{\theta}_{BB} - \widehat{\theta}_{Bb}\right),$$
$$\widehat{\theta}_{Bb} = \frac{y_{Bb}}{n - y_{BB} - y_{Bb}} \left(1 - \widehat{\theta}_{BB} - \widehat{\theta}_{Bb}\right).$$

Summing these two equations yields

$$\widehat{\theta}_{BB} + \widehat{\theta}_{Bb} = \frac{y_{BB} + y_{Bb}}{n - y_{BB} - y_{Bb}} \left(1 - \widehat{\theta}_{BB} - \widehat{\theta}_{Bb}\right).$$

Because of the parametric relationship (probabilities of all possible disjoint events sum to one), the invariance property of the ML estimates yields $\widehat{\theta}_{bb} = 1 - \widehat{\theta}_{BB} - \widehat{\theta}_{Bb}$. Using this above

$$1 - \widehat{\theta}_{bb} = \frac{y_{BB} + y_{Bb}}{n - y_{BB} - y_{Bb}} \widehat{\theta}_{bb},$$

from which: $\widehat{\theta}_{bb} = y_{bb}/n$. Similarly, it can be established that $\widehat{\theta}_{BB} = y_{BB}/n$ and $\widehat{\theta}_{Bb} = y_{Bb}/n$. In the absence of restrictions (other than those imposed by probability theory) on the values of the parameters, i.e., under H_1, the maximized likelihood is

$$L\left(\widehat{\theta}_{BB}, \widehat{\theta}_{Bb}, \widehat{\theta}_{bb}\right) \propto \left(\frac{y_{BB}}{n}\right)^{y_{BB}} \left(\frac{y_{Bb}}{n}\right)^{y_{Bb}} \left(\frac{y_{bb}}{n}\right)^{y_{bb}}.$$

A genetic hypothesis (denoted as H_0 here) is that the population is in Hardy–Weinberg equilibrium (Crow and Kimura, 1970; Weir, 1996). Under

H_0, the genotypic distribution is expected to obey the parametric relationship

$$\begin{aligned}\theta_{BB} &= \left(\theta_{BB} + \tfrac{1}{2}\theta_{Bb}\right)^2 = \theta_B^2,\\ \theta_{Bb} &= 2\theta_B\left(1-\theta_B\right),\\ \theta_{bb} &= \left(1-\theta_B\right)^2,\end{aligned}$$

where $\theta_B = \theta_{BB} + \theta_{Bb}/2$ is called the frequency of allele B in the population. This is the total probability of drawing an allele B at this locus, that is, the sum of:
a) the probability of drawing an individual with genotype BB (θ_{BB}) times the probability of obtaining a B from BB, which is a certain event; plus
b) the probability of drawing a Bb (θ_{Bb}), times the probability of obtaining B from a heterozygote $\left(\frac{1}{2}\right)$.
Hence, under Hardy–Weinberg equilibrium the probability distribution of the observations is governed by a single parameter, θ_B. The likelihood under H_0 is

$$L\left(\theta_B\right) \propto \left(\theta_B^2\right)^{y_{BB}} \left[\theta_B\left(1-\theta_B\right)\right]^{y_{Bb}} \left[\left(1-\theta_B\right)^2\right]^{y_{bb}}.$$

Differentiation of the log-likelihood with respect to θ_B and setting to zero gives an explicit solution that can be expressed as a function of the ML estimators under H_1:

$$\hat{\hat{\theta}}_B = \frac{2y_{BB} + y_{Bb}}{2y} = \hat{\theta}_{BB} + \frac{1}{2}\hat{\theta}_{Bb}.$$

Because of the parametric relationships, under H_0 the ML estimators of the probabilities of genotypes in the population (or genotypic frequencies, in the population genetics literature) are:

$$\begin{aligned}\hat{\hat{\theta}}_{BB} &= \hat{\hat{\theta}}_B^2 = \left(\hat{\theta}_{BB} + \tfrac{1}{2}\hat{\theta}_{Bb}\right)^2,\\ \hat{\hat{\theta}}_{Bb} &= 2\left(\hat{\theta}_{BB} + \tfrac{1}{2}\hat{\theta}_{Bb}\right)\left(\hat{\theta}_{bb} + \tfrac{1}{2}\hat{\theta}_{Bb}\right),\\ \hat{\hat{\theta}}_{bb} &= \left(\hat{\theta}_{bb} + \tfrac{1}{2}\hat{\theta}_{Bb}\right)^2.\end{aligned}$$

Using (4.23), the statistic for the likelihood ratio test of the hypothesis that the population is in Hardy–Weinberg equilibrium, is

$$-2\ln(LR) = -2\ln\left\{\left[\frac{\left(\widehat{\theta}_{BB}+\frac{1}{2}\widehat{\theta}_{Bb}\right)^2}{\widehat{\theta}_{BB}}\right]^{y_{BB}} \times\left[\frac{2\left(\widehat{\theta}_{BB}+\frac{1}{2}\widehat{\theta}_{Bb}\right)\left(\widehat{\theta}_{bb}+\frac{1}{2}\widehat{\theta}_{Bb}\right)}{\widehat{\theta}_{Bb}}\right]^{y_{Bb}}\left[\frac{\left(\widehat{\theta}_{bb}+\frac{1}{2}\widehat{\theta}_{Bb}\right)^2}{\widehat{\theta}_{bb}}\right]^{y_{Bb}}\right\}.$$

Asymptotically, this has a central χ^2_1 distribution. In the general setting described in Section 4.3, the parameter vector $\boldsymbol{\theta}$ has $r+s$ identifiable parameters under H_1; here $r+s=2$ instead of 3 because there is a redundant parameter. Under H_0 there is a single estimable parameter. This provides the basis for the single degree of freedom of the distribution of the LR statistic.

It is possible to move from the unrestricted to the restricted model by setting a single function of parameters to zero. For example, letting

$$d = \theta_{Bb} - 2\left(\theta_{BB}+\tfrac{1}{2}\theta_{Bb}\right)\left(\theta_{bb}+\tfrac{1}{2}\theta_{Bb}\right)$$

it follows that a test of the Hardy–Weinberg equilibrium is equivalent to a test of the hypothesis H_0: $d=0$. Hence, the models under the two hypotheses differ by a single parameter which, upon setting to 0, can produce the null model as a "nested" version of the unrestricted model. ■

4.3.2 Confidence Regions

The asymptotic distribution of ML estimators enables one to obtain interval inferences about the "true" value of the parameters. These inferences are expressed in terms of confidence regions. There is a close relationship between the techniques needed here and those described for the evaluation of hypotheses. Hence, construction of confidence regions is dealt with only briefly.

Before we do so, a comment about the interpretation of confidence intervals is in order. Confidence intervals are defined in terms of the distribution of the random variable $\mathbf{y}$, the data, and as such the confidence interval is also a random variable. Given a realization of $\mathbf{y}$, the confidence interval will either contain the true value of the parameter or not. If the confidence region is $1-\alpha$ (see below), then under repeated sampling, $100(1-\alpha)\,\%$ of the intervals will contain the true value of the parameter. Thus the confidence interval is not a probability statement about the parameter but about the random interval.

Given that a model or a hypothesis is true, the score has, asymptotically, a normal distribution with null mean and covariance matrix $\mathbf{I}(\boldsymbol{\theta})$. Hence, a confidence region of size $1-\alpha$ can be constructed from the property that

$$\Pr\left[\mathbf{l}'(\boldsymbol{\theta})' \mathbf{I}^{-1}(\boldsymbol{\theta}) \mathbf{l}'(\boldsymbol{\theta}) \leq t_\alpha\right] = 1-\alpha, \tag{4.27}$$

where t_α is the critical value of a χ_p^2 random variable. Values of $\boldsymbol{\theta}$ outside the region are viewed as being unlikely, but not in a probabilistic sense, because values of the parameters cannot be assigned probabilities. Note that if the statement involves a single parameter, the confidence region can be formed from a standard normal distribution, this being so because the random variable

$$\frac{\left[l'(\theta)\right]^2}{I(\theta)} \sim \chi_1^2$$

so

$$\frac{l'(\theta)}{\sqrt{I(\theta)}} \sim N(0,1).$$

Another way of constructing confidence regions is based on the asymptotic distribution of the ML estimator, as given in (3.66) of the previous chapter. The confidence region stems from the fact that, asymptotically, the distribution under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mathbf{I}(\boldsymbol{\theta}_0) \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \sim \chi_p^2. \tag{4.28}$$

In the single parameter case, the region is defined by the appropriate $\alpha/2$ percentiles of a $N(0,1)$ distribution. This is so because

$$\frac{\widehat{\theta} - \theta_0}{\sqrt{I^{-1}(\theta_0)}} = \left(\widehat{\theta} - \theta_0\right)\sqrt{I(\theta_0)} \sim N(0,1).$$

Similarly, a confidence region can be developed from the asymptotic distribution of the likelihood ratio as given in (4.23). A confidence region here is a set of values of $\boldsymbol{\theta}$ in the neighborhood of the maximum (Cox and Snell, 1989). With a single parameter, the asymptotic distribution of the likelihood ratio statistic $-2\ln(LR)$ is χ_1^2 under H_0. Noting that a χ_1^2 random variable can be generated by squaring a standard normal deviate, an equivalent form of generating a confidence region is to use $\sqrt{-2\ln(LR)} \sim N(0,1)$ if the ML estimator is larger than the null value θ_0, and $-\sqrt{-2\ln(LR)} \sim N(0,1)$ otherwise. This is because if x and $-x$ are realized values from a $N(0,1)$ process, these two generate the same realized value from a χ_1^2 distribution, in a two-to-one mapping.

In passing, we mention that the "pure likelihoodist" computes confidence regions based on quantiles derived from the likelihood ratio directly, without invoking its distribution over replications of the data. A tutorial on the topic can be found in Meeker and Escobar (1995).

This section finishes with a brief description of two tests that are also based on classical, first-order asymptotic likelihood theory.

4.3.3 Wald's Test

As stated in the previous chapter, in large samples, the distribution of the ML estimator can be written (informally) as $\widehat{\boldsymbol{\theta}} \sim N\left[\boldsymbol{\theta}_0, \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right]$, where $\boldsymbol{\theta}_0$ is the true value of the parameter $\boldsymbol{\theta}$. Then, asymptotically, the quadratic form

$$\lambda_W = \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right)' \left[\mathbf{I}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) \sim \chi_r^2, \tag{4.29}$$

where $\widehat{\boldsymbol{\theta}}_1$ is the ML estimator of $\boldsymbol{\theta}_1$ under H_1, and χ_r^2 is a central chi-square random variable with r degrees of freedom (the number of elements in $\boldsymbol{\theta}_1$). Result (4.29) is an immediate application of (1.101). If λ_W is "too large", the hypothesis is rejected. Approximate confidence regions can be readily constructed from (4.29).

In (4.29), $\mathbf{I}^{11}\left(\widehat{\boldsymbol{\theta}}\right)$ is the top left element of the inverse of Fisher's expected information matrix. That is

$$\mathbf{I}\left(\widehat{\boldsymbol{\theta}}\right) = -E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \begin{bmatrix} \mathbf{I}_{11}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}_{12}\left(\widehat{\boldsymbol{\theta}}\right) \\ \mathbf{I}_{21}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}_{22}\left(\widehat{\boldsymbol{\theta}}\right) \end{bmatrix}, \tag{4.30}$$

and

$$\begin{bmatrix} \mathbf{I}_{11}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}_{12}\left(\widehat{\boldsymbol{\theta}}\right) \\ \mathbf{I}_{21}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}_{22}\left(\widehat{\boldsymbol{\theta}}\right) \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}^{11}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}^{12}\left(\widehat{\boldsymbol{\theta}}\right) \\ \mathbf{I}^{21}\left(\widehat{\boldsymbol{\theta}}\right) & \mathbf{I}^{22}\left(\widehat{\boldsymbol{\theta}}\right) \end{bmatrix}.$$

Therefore,

$$\left[\mathbf{I}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} = \mathbf{I}_{11}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{I}_{12}\left(\widehat{\boldsymbol{\theta}}\right)\left[\mathbf{I}_{22}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1}\mathbf{I}_{21}\left(\widehat{\boldsymbol{\theta}}\right).$$

Rather than using the inverse of Fisher's information, other estimators of $\mathbf{I}^{11}\left(\boldsymbol{\theta}\right)$, which are consistent under H_0, can be used. Thus, an alternative form of Wald's statistic uses the inverse of the observed information, $\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)$, in (4.29), which retrieves (4.22). This makes it clear that the Wald statistic is based on a quadratic approximation to $-2\ln LR$.

Due to the relationship between a chi-squared variable and a normal variable, for scalar θ, the $100\left(1-\alpha\right)\%$ confidence interval based on the Wald statistic is given by the well known formula

$$\widehat{\theta} \pm z_{\alpha/2} I^{-1/2}\left(\widehat{\theta}\right).$$

4.3.4 Score Test

The score test (or Lagrange multiplier test, as called by econometricians) is also based on a quadratic approximation to the log-likelihood ratio. It was

proposed by Rao (1947) and uses the asymptotic properties of the score (3.63). The score statistic is

$$\mathbf{l}'_{\theta_1}\left(\boldsymbol{\theta}_{10},\widetilde{\boldsymbol{\theta}}_2\right)'\mathbf{I}^{11}\left(\boldsymbol{\theta}_{10},\widetilde{\boldsymbol{\theta}}_2\right)\mathbf{l}'_{\theta_1}\left(\boldsymbol{\theta}_{10},\widetilde{\boldsymbol{\theta}}_2\right), \tag{4.31}$$

which, in view of (3.63) and of (1.101), follows a χ^2_r distribution. The null hypothesis is rejected for large values of (4.31). In (4.31),

$$\mathbf{l}'_{\theta_1}\left(\boldsymbol{\theta}_{10},\widetilde{\boldsymbol{\theta}}_2\right)=\left.\frac{\partial l\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial\boldsymbol{\theta}_1}\right|_{\substack{\boldsymbol{\theta}_1=\boldsymbol{\theta}_{10}\\ \boldsymbol{\theta}_2=\widetilde{\boldsymbol{\theta}}_2}}.$$

Contrary to the likelihood ratio test and the Wald test, (4.31) requires only one maximization (to compute the ML estimate under H_0, $\widetilde{\boldsymbol{\theta}}_2$). Approximate confidence regions can also be constructed from (4.31).

The three tests described here (likelihood ratio, Wald's, and score test) are asymptotically equivalent and are all first-order approximations. Which test to use depends on the situation (Lehmann, 1999), although Meeker and Escobar (1995) and Pawitan (2000) argue in favor of the likelihood ratio. There are two reasons. First, the Wald and score statistics are convenient only if the log-likelihood is well-approximated by a quadratic function. Second, the Wald statistic has an important disadvantage relative to the likelihood ratio test: it is not transformation invariant. However, asymptotically, when the likelihood is a quadratic function of the parameter, the tests are equivalent. To illustrate this in the case of a single parameter, write the scaled log-likelihood (i.e., the logarithm of (3.21)) as

$$l\left(\theta\right)=k\left(\theta-\widehat{\theta}\right)^2.$$

Therefore,

$$\frac{dl\left(\theta\right)}{d\theta}=2k\left(\theta-\widehat{\theta}\right)$$

and

$$\frac{d^2l\left(\theta\right)}{d^2\theta}=-I\left(\theta\right)=2k.$$

Assume that under the null hypothesis, $H_0:\theta=\theta_0$ and the alternative hypothesis is $H_1:\theta\neq\theta_0$. Then the test based on the likelihood ratio is

$$\begin{aligned}2\left[l\left(\widehat{\theta}\right)-l\left(\theta_0\right)\right]&=2\left[k\left(\widehat{\theta}-\widehat{\theta}\right)^2-k\left(\theta_0-\widehat{\theta}\right)^2\right]\\&=-2k\left(\theta_0-\widehat{\theta}\right)^2.\end{aligned}$$

The Wald test is obtained from (4.29):

$$\left(\widehat{\theta}-\theta_0\right)^2 I^{-1}\left(\theta\right)_{\theta=\widehat{\theta}}=-2k\left(\theta_0-\widehat{\theta}\right)^2.$$

The score test is obtained from (4.31):

$$\frac{\left[dl\left(\theta_0\right)/d\theta\right]^2}{I\left(\theta_0\right)} = \frac{\left[2k\left(\theta_0 - \widehat{\theta}\right)\right]^2}{-2k} = -2k\left(\theta_0 - \widehat{\theta}\right)^2.$$

4.4 Nuisance Parameters

Recall that a statistical model typically includes parameters of primary inferential interest (denoted as $\boldsymbol{\theta}_1$ here), plus additional parameters ($\boldsymbol{\theta}_2$) that are necessary to index completely the distributions of all random variables entering into a probability model. The additional parameters are called nuisance parameters. As pointed out by Edwards (1992), one would wish to make statements about the values of $\boldsymbol{\theta}_1$ without reference to the values of $\boldsymbol{\theta}_2$. If $\boldsymbol{\theta}_2$ were known, there would be no difficulty, as one would write the likelihood of $\boldsymbol{\theta}_1$, with $\boldsymbol{\theta}_2$ replaced by its true value. In the absence of such knowledge, a possibility would be to infer $\boldsymbol{\theta}_1$ at each of a series of possible values of $\boldsymbol{\theta}_2$. This is unsatisfactory, because it does not give guidance about the plausibility of each of the values of the nuisance parameters. A more appealing option is to estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ jointly, as if both were of primary interest, using the machinery for analysis of likelihoods developed so far. Unfortunately, when making inferences about $\boldsymbol{\theta}_1$, this modus operandi does not take into account the fact that part of the information contained in the data must be used to estimate $\boldsymbol{\theta}_2$. For example, in a nonlinear regression model, the vector of parameters of the expectation function may be of primary interest, with the residual variance playing the role of a nuisance parameter. Another example is that of a model with two means and two variances, where the inferential interest centers on, say, σ_2/μ_1, with the remaining parameters acting as nuisances. In any case, it is not obvious how to deal with nuisance parameters in likelihood-based inference and many solutions have been proposed. This area has been undergoing rapid development (Kalbfleisch and Sprott, 1970, 1973; Barndorff-Nielsen, 1986, 1991; Cox and Reid, 1987; McCullagh and Nelder, 1989; Efron, 1993; Barndorff-Nielsen and Cox, 1994; Severini, 1998). Useful reviews can be found in Reid (1995) and Reid (2000). The recent book of Severini (2000) is a good starting point to study modern likelihood methods. Here we only discuss the use of marginal and profile likelihoods.

The subject is introduced below with an example that illustrates the loss of efficiency in the estimation of parameters of interest due to the presence of nuisance parameters. The material is taken from Lehmann (1999).

4.4.1 Loss of Efficiency Due to Nuisance Parameters

Consider a model depending on p parameters $\boldsymbol{\theta}$ with elements

$$\theta_1, \ldots, \theta_p.$$

Let $\mathbf{I}(\boldsymbol{\theta})$ denote the information matrix, with typical element $I_{ij}(\boldsymbol{\theta})$ and with inverse $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$ with typical element $I^{ij}(\boldsymbol{\theta})$. Assuming that the necessary regularity conditions hold, then from (3.65),

$$\sqrt{n}\left(\widehat{\theta}_1 - \theta_1\right), \ldots, \sqrt{n}\left(\widehat{\theta}_k - \theta_p\right)$$

has a joint multivariate distribution with mean $(0, 0, \ldots, 0)'$ and covariance matrix $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$. In particular,

$$\sqrt{n}\left(\widehat{\theta}_j - \theta_j\right) \to N\left(0, I^{jj}(\boldsymbol{\theta})\right),$$

where $I^{jj}(\boldsymbol{\theta})$ is the element in row j and column j of $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$. On the other hand, if $\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p$ are known, from (3.55),

$$\sqrt{n}\left(\widehat{\theta}_j - \theta_j\right) \to N\left(0, [I_{jj}(\boldsymbol{\theta})]^{-1}\right).$$

The loss in efficiency due to the presence of nuisance parameters can be studied via the relationship between $I^{jj}(\boldsymbol{\theta})$ and $[I_{jj}(\boldsymbol{\theta})]^{-1}$. Consider the case $p = 2$ and $j = 1$. The inverse of the 2×2 matrix $\mathbf{I}(\boldsymbol{\theta})$ is

$$[\mathbf{I}(\boldsymbol{\theta})]^{-1} = \begin{bmatrix} I_{22}(\boldsymbol{\theta})/\Delta & -I_{12}(\boldsymbol{\theta})/\Delta \\ -I_{12}(\boldsymbol{\theta})/\Delta & I_{11}(\boldsymbol{\theta})/\Delta \end{bmatrix},$$

where $\Delta = I_{11}(\boldsymbol{\theta}) I_{22}(\boldsymbol{\theta}) - [I_{12}(\boldsymbol{\theta})]^2$. Since $\mathbf{I}(\boldsymbol{\theta})$ is positive definite, Δ is positive. This implies that $I_{11}(\boldsymbol{\theta}) I_{22}(\boldsymbol{\theta}) \geq \Delta$ which is equivalent to

$$I^{11}(\boldsymbol{\theta}) = \frac{I_{22}(\boldsymbol{\theta})}{\Delta} \geq [I_{11}(\boldsymbol{\theta})]^{-1}, \tag{4.32}$$

with equality holding when $I_{12}(\boldsymbol{\theta}) = 0$. The conclusion is that, even in an asymptotic scenario, unless the estimators are asymptotically independent, the asymptotic variance of the estimator of the parameter of interest is larger in models containing unknown nuisance parameters.

4.4.2 Marginal Likelihoods

The marginal likelihood approach is based on the construction of a likelihood for the parameters of interest from a "suitably chosen subset of the data vector" (McCullagh and Nelder, 1989). It is desirable to choose this subset to be as large as possible, to minimize any loss of information. As

indicated by McCullagh and Nelder (1989), the method does not always work satisfactorily because general rules do not seem to exist. Even in cases where it appears to work acceptably (e.g., elimination of location parameters in a Gaussian linear model), it is difficult to evaluate whether or not there is a loss of information in the process of the eliminating parameters. This is illustrated below.

Consider the sampling model studied in Example 4.1, and suppose the parameter of primary interest is $\theta_1 = \sigma^2$ whereas the nuisance parameter is the location vector $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$. A likelihood that does not involve $\boldsymbol{\beta}$ can be constructed from a vector of fitted residuals, also called "error" contrasts, as suggested by Patterson and Thompson (1971). Let

$$\begin{aligned}\mathbf{w} &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{y} \\ &= \mathbf{M}\mathbf{y} \end{aligned} \tag{4.33}$$

for $\mathbf{M} = [\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}]$ of order $N \times N$. If $\mathbf{y}$ is normal, $\mathbf{w}$ must be normal by virtue of it being a linear combination of normal variables. Further

$$E(\mathbf{w}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \tag{4.34}$$

and:

$$Var(\mathbf{w}) = \mathbf{M}\mathbf{V}\mathbf{M}'\sigma^2. \tag{4.35}$$

Observe that

$$rank[Var(\mathbf{w})] \leq rank[\mathbf{M}]$$

and that $\mathbf{M}$ is an idempotent matrix. Using properties of idempotent matrices plus cyclical commutation under the trace operator (Searle, 1982)

$$\begin{aligned} rank[\mathbf{M}] &= tr[\mathbf{M}] = tr\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\right] \\ &= N - tr\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right] \\ &= N - p, \end{aligned}$$

where p is the order of $\boldsymbol{\beta}$. Hence, $Var(\mathbf{w})$ has deficient rank and

$$\mathbf{w} \sim SN\left(\mathbf{0}, \mathbf{M}\mathbf{V}\mathbf{M}'\sigma^2\right), \tag{4.36}$$

where SN denotes a singular normal distribution having a covariance matrix of rank $N - p$ (Searle, 1971). This means that p of the elements of $\mathbf{w}$ are redundant and can be expressed as a linear function of the $N - p$ linearly independent combinations. Put

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_L \\ \mathbf{w}_R \end{bmatrix} = \begin{bmatrix} \mathbf{M}_L \\ \mathbf{M}_R \end{bmatrix}\mathbf{y} = \begin{bmatrix} \mathbf{M}_L\mathbf{y} \\ \mathbf{M}_R\mathbf{y} \end{bmatrix},$$

where $\mathbf{w}_L$ ($\mathbf{w}_R$) stands for a linearly independent (redundant) part of $\mathbf{w}$, and $\mathbf{M}_L$ ($\mathbf{M}_R$) denotes the corresponding partition of $\mathbf{M}$.

Consider now the distribution of $\mathbf{w}_L$. From (4.34) and (4.35), this distribution $\mathbf{w}_L$ must be normal with a nonsingular covariance matrix

$$\mathbf{w}_L \sim N\left(\mathbf{0}, \mathbf{M}_L \mathbf{V} \mathbf{M}_L' \sigma^2\right). \tag{4.37}$$

This distribution depends on σ^2 only, so it is "free" of the nuisance parameter $\boldsymbol{\beta}$. Hence, a likelihood function devoid of $\boldsymbol{\beta}$ can be constructed based on $\mathbf{w}_L$. This marginal likelihood can be written as

$$\begin{aligned} L\left(\sigma^2\right) &\propto \left|\mathbf{M}_L \mathbf{V} \mathbf{M}_L' \sigma^2\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} \mathbf{w}_L' \left(\mathbf{M}_L \mathbf{V} \mathbf{M}_L\right)^{-1} \mathbf{w}_L\right\} \\ &\propto \left(\sigma^2\right)^{-\frac{N-p}{2}} \exp\left\{-\frac{1}{2\sigma^2} \mathbf{y}' \mathbf{M}_L' \left(\mathbf{M}_L \mathbf{V} \mathbf{M}_L'\right)^{-1} \mathbf{M}_L \mathbf{y}\right\}. \end{aligned} \tag{4.38}$$

A result in linear algebra (Searle et al., 1992) states that if $\mathbf{M}_L \mathbf{X} = \mathbf{0}$ and $\mathbf{V}$ is positive definite, two conditions met here, then

$$\mathbf{M}_L' \left(\mathbf{M}_L \mathbf{V} \mathbf{M}_L'\right)^{-1} \mathbf{M}_L = \mathbf{V}^{-1} \mathbf{M}$$

and this holds for any $\mathbf{M}_L$ having full row rank. Using the preceding in (4.38), the marginal log-likelihood function is

$$\begin{aligned} l\left(\sigma^2\right) &= -\frac{N-p}{2} \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \mathbf{y}' \mathbf{V}^{-1} \mathbf{M} \mathbf{y} \\ &= -\frac{N-p}{2} \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \left(\mathbf{y}' \mathbf{V}^{-1} \mathbf{y} - \widehat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}\right) \\ &= -\frac{N-p}{2} \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right), \end{aligned} \tag{4.39}$$

where $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$. Maximization of (4.39) with respect to σ^2 gives, as the marginal ML estimator,

$$\widetilde{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)}{N-p} = \frac{S_y}{N-p}. \tag{4.40}$$

Writing

$$\widetilde{\sigma}^2 = \frac{S_y}{\sigma^2} \frac{\sigma^2}{N-p},$$

and noting that $S_y/\sigma^2 \sim \chi^2_{N-p}$, it follows that the marginal ML estimator has a scaled chi-square distribution, with mean and variance

$$E\left(\widetilde{\sigma}^2\right) = \frac{\sigma^2}{N-p}(N-p) = \sigma^2, \tag{4.41}$$

$$Var\left(\widetilde{\sigma}^2\right) = \left(\frac{\sigma^2}{N-p}\right)^2 2(N-p) = \frac{2\sigma^4}{N-p}. \tag{4.42}$$

This should be contrasted with the ML estimator:

$$\widehat{\sigma^2} = \frac{S_y}{N} \tag{4.43}$$

that also has a scaled chi-square distribution with mean and variance

$$E\left(\widehat{\sigma^2}\right) = \sigma^2 \frac{N-p}{N}, \tag{4.44}$$

$$Var\left(\widehat{\sigma^2}\right) = \frac{2\sigma^4}{N}. \tag{4.45}$$

The marginal ML estimator $\widetilde{\sigma}^2$ is unbiased, whereas $\widehat{\sigma^2}$ has a downward bias. This bias can be severe if p is large relative to N. On the other hand, the ML estimator is more precise (lower variance) than the estimator based on a marginal likelihood, suggesting that some information is lost in the process of eliminating the nuisance parameter $\boldsymbol{\beta}$. This can be checked by computing the information about σ^2 contained in the marginal likelihood (4.38). Differentiating the marginal log-likelihood twice with respect to σ^2, and multiplying by -1, gives the observed information

$$-\frac{N-p}{2\sigma^4} + \frac{S_y}{\sigma^6}.$$

The expected information is

$$I_M\left(\sigma^2\right) = -\frac{N-p}{2\sigma^4} + \frac{(N-p)\,\sigma^2}{\sigma^6} = \frac{N-p}{2\sigma^4}. \tag{4.46}$$

Using a similar procedure, the information from the full likelihood can be found to be equal to

$$I\left(\sigma^2\right) = \frac{N}{2\sigma^4}. \tag{4.47}$$

A comparison between (4.46) and (4.47) indicates that there is a loss of information in the process of eliminating the nuisance parameter, at least in the situation considered here. The loss of information can be serious in a model where p/N is large. The corresponding asymptotic distributions are then $\widetilde{\sigma}^2 \sim N\left(0, 2\sigma^4/(N-p)\right)$ and $\widehat{\sigma^2} \sim N\left(0, 2\sigma^4/N\right)$. This fact would seem to favor the ML estimator. However, for very large N (relative to p) the two distributions are expected to differ by little.

These asymptotic distributions do not give guidance on how to choose between the estimators when samples are finite. Here, it was shown that the two ML estimators have distributions that are multiples of chi-square random variables but with different means and variances. Consider a comparison based on the mean squared error criterion. The mean squared error

of an estimator $\widehat{\theta}$ is:

$$\begin{aligned} E\left(\widehat{\theta}-\theta\right)^2 &= E\left[\widehat{\theta}-E\left(\widehat{\theta}\right)+E\left(\widehat{\theta}\right)-\theta\right]^2 \\ &= E\left[\widehat{\theta}-E\left(\widehat{\theta}\right)\right]^2+\left[E\left(\widehat{\theta}\right)-\theta\right]^2 \\ &= Var\left(\widehat{\theta}\right)+Bias^2\left(\widehat{\theta}\right), \end{aligned} \tag{4.48}$$

where $Bias\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right)-\theta$ gives the expected deviation of the average of the realized value of the estimators from the true parameter value in a process of conceptual repeated sampling. The marginal ML estimator has null bias (see equation 4.41) so its mean squared error is equal to its variance, as given in (4.42). Using expressions (4.44) and (4.45), the corresponding mean squared error of the ML estimator is

$$E\left(\widehat{\sigma^2}-\sigma^2\right) = \frac{2\sigma^4}{N-p}\left[1-k+k^2\frac{N-p}{2}\right],$$

where $k = p/N$. From this, conditions can be found under which one of the two estimators is "better" than the other in terms of mean squared error. For example, if $p = 1$, that is, if the model has a single location parameter, the ML estimator has a smaller mean squared error than the estimator based on marginal likelihood throughout the parameter space. For this particular model and loss function (mean squared error), the marginal ML estimator is said to be inadmissible, because it is known that a better estimator exists for all values of σ^2. Unfortunately, these calculations are not possible in more complicated models, for example, Gaussian mixed effects models with unknown variance components. In general, the choice of the likelihood to be maximized is not obvious.

The idea of using a subset of the data (error contrasts) to make inferences about variance components in a linear model was suggested by Patterson and Thompson (1971). These authors used the term restricted likelihood instead of marginal likelihood, and called the resulting estimates REML for short. It is unclear how this idea can be generalized to other parameters of a linear or nonlinear model. The Bayesian approach, on the other hand, provides a completely general form of elimination of nuisance parameters. This will be discussed in the following chapter.

4.4.3 Profile Likelihoods

When a model has nuisance parameters, it is possible to define a likelihood that can be used almost invariably, but not without pitfalls. It is known as a profile likelihood. In order to introduce the concept, consider a model with parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_2$ is regarded as a vector of nuisance

parameters. Observe that ML estimation equations must satisfy

$$\frac{\partial l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial \boldsymbol{\theta}_1} = \mathbf{0},$$
$$\frac{\partial l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)}{\partial \boldsymbol{\theta}_2} = \mathbf{0}.$$

The first equation defines the ML estimator of $\boldsymbol{\theta}_1$ at a fixed value of the nuisance parameter, and vice-versa for the second equation. Let $\widehat{\boldsymbol{\theta}}_{2|1}$ be the partial ML estimator of the nuisance parameter obtained from the second equation, that is, with $\boldsymbol{\theta}_1$ fixed. This "partial" estimator depends only on the data and $\boldsymbol{\theta}_1$. The profile log-likelihood of $\boldsymbol{\theta}_1$ is obtained by replacing $\boldsymbol{\theta}_2$ by $\widehat{\boldsymbol{\theta}}_{2|1}$ in the likelihood function, and is defined as

$$l_P\left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_{2|1}|\mathbf{y}\right). \tag{4.49}$$

Note that (4.49) has the same form as the numerator of the likelihood ratio (4.11), which can also be viewed as a profile likelihood ratio. Expression (4.49) is a function of $\boldsymbol{\theta}_1$ only, and its maximizer must be such that

$$\frac{\partial l_P\left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_{2|1}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}_1} = \mathbf{0}.$$

It follows that the maximizer of the profile likelihood must be identical to the ML estimator of $\boldsymbol{\theta}_1$ obtained from $l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)$, this being so because the first of the two ML estimating equations given above is satisfied by $\widehat{\boldsymbol{\theta}}_{2|1}$.

In Subsection 4.3.1, the nuisance parameter $\boldsymbol{\theta}_2$ was estimated conditionally on a given value of $\boldsymbol{\theta}_1$. This conditional estimator, labelled $\widetilde{\boldsymbol{\theta}}_2$, was replaced in $l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right)$ and the latter evaluated at $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$. This led to expression (4.21) for $l\left(\boldsymbol{\theta}_{10}, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right)$. If instead, $\boldsymbol{\theta}_1$ is left free to vary, one obtains

$$l_P\left(\boldsymbol{\theta}_1, \widetilde{\boldsymbol{\theta}}_2|\mathbf{y}\right) \approx \text{constant} - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)' \left[\mathbf{J}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right), \tag{4.50}$$

which is an asymptotic first-order approximation to the profile log-likelihood. Then the standard results that led to (3.66) hold here also, and we can write (informally),

$$\widehat{\boldsymbol{\theta}}_1 \sim N\left[\boldsymbol{\theta}_1, \mathbf{I}^{11}\left(\widehat{\boldsymbol{\theta}}\right)\right]. \tag{4.51}$$

In view of (4.32), this shows that the asymptotic variance in the presence of nuisance parameters is larger than when these are absent or assumed known, and the profile likelihood accounts for this extra uncertainty.

However, a word of caution is necessary. Although it would seem that the ML machinery can be applied in a straightforward manner using a

profile likelihood, this is not always the case. The profile likelihood is not proportional to the density function of a random variable. In large samples, replacing $\boldsymbol{\theta}_2$ by its ML estimate has relatively little consequences for inferences involving $\boldsymbol{\theta}_1$. However, if the dimension of $\boldsymbol{\theta}_2$ is large relative to sample size, a situation in which it would be difficult to argue asymptotically, the profile log-likelihood can be misleading when interpreted as a log-likelihood function (McCullagh and Nelder, 1989; Cox and Snell, 1989).

Example 4.4 *Profile likelihoods in a linear model*
Let the joint distribution of the observations be

$$\mathbf{y} \sim N\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \mathbf{V}\sigma^2\right)$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and σ^2 are the unknown parameters and $\mathbf{V}$ is a known matrix. The likelihood function is

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) = (2\pi)^{-\frac{N}{2}} \left|\mathbf{V}\sigma^2\right|^{-\frac{1}{2}}$$
$$\times \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)\right].$$

It will be shown here how different profile likelihoods are constructed.

(a) Suppose that the nuisance parameter is σ^2 and that inferences are sought about the location vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The partial ML estimator of σ^2 is

$$\overrightarrow{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)}{N}.$$

Replacing this in the likelihood gives the profile likelihood of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$L_P\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \mathbf{y}\right) \propto \left(\overrightarrow{\sigma}^2\right)^{-\frac{N}{2}} \exp\left[-\frac{N}{2}\right] \propto \left(\overrightarrow{\sigma}^2\right)^{-\frac{N}{2}}$$

and this does not depend on the nuisance parameter σ^2. The corresponding profile log-likelihood, ignoring the constant, is

$$l_P\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \mathbf{y}\right) = -\frac{N}{2} \ln\left[\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2\right)\right].$$

Setting the first derivatives with respect to the unknowns to zero and rearranging gives as solution

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{y} \end{bmatrix},$$

retrieving the ML estimator, as expected. In this case the profile likelihood has the same asymptotic properties as the usual likelihood, this being so because $\overrightarrow{\sigma}^2$ is a consistent estimator of σ^2. This can be verified by noting

that $\mathbf{V}$ can be decomposed as $\mathbf{V} = \mathbf{L}\mathbf{L}'$ with nonsingular $\mathbf{L}$. Hence, for $\mathbf{y}^* = (\mathbf{L})^{-1}\mathbf{y}$, then

$$\mathbf{y}^* \sim \mathbf{N}\left[(\mathbf{L})^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2), \mathbf{I}\sigma^2\right],$$

so the transformed observations are independent. Putting $(\mathbf{L})^{-1}\mathbf{X}_i = \mathbf{X}_i^*$, and letting $\mathbf{x}_{ij}^{*\prime}$ denote the jth row of $\mathbf{X}_i^*$, it turns out that the partial ML estimator can be written as

$$\overrightarrow{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}\left(y_j^* - \mathbf{x}_{1j}^{*\prime}\boldsymbol{\beta}_1 - \mathbf{x}_{2j}^{*\prime}\boldsymbol{\beta}_2\right)^2.$$

It follows that as $N \to \infty$, then

$$\overrightarrow{\sigma}^2 \to \mathbf{E}\left[\left(y_j^* - \mathbf{x}_{1j}^{*\prime}\boldsymbol{\beta}_1 - \mathbf{x}_{2j}^{*\prime}\boldsymbol{\beta}_2\right)^2\right] = \sigma^2,$$

because of the law of large numbers.

(b) Let $\boldsymbol{\beta}_1$ now be of primary interest, with $\boldsymbol{\beta}_2$ and σ^2 acting as nuisance parameters. The partial ML estimators of $\boldsymbol{\beta}_2$ and σ^2 are

$$\overleftarrow{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1\right)$$

and

$$\overleftarrow{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\overleftarrow{\boldsymbol{\beta}}_2\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\overleftarrow{\boldsymbol{\beta}}_2\right)}{N}.$$

Note that $\overleftarrow{\boldsymbol{\beta}}_2 \sim N[\boldsymbol{\beta}_2, \left(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}\sigma^2]$ is the ML estimator of $\boldsymbol{\beta}_2$ applied to the "corrected" data $\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$, so it must converge in probability to $\boldsymbol{\beta}_2$. Hence, $\overleftarrow{\sigma}^2$ must be consistent as well. The profile likelihood for this case is

$$L_P(\boldsymbol{\beta}_1|, \mathbf{y}) \propto \left(\overleftarrow{\sigma}^2\right)^{-\frac{N}{2}}$$
$$\times \exp\left[-\frac{1}{2\overleftarrow{\sigma}^2}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\overleftarrow{\boldsymbol{\beta}}_2\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\overleftarrow{\boldsymbol{\beta}}_2\right)\right].$$

(c) The parameter of interest now is σ^2. The partial ML estimators of the nuisance parameters are $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, as with full ML. This is so because solving the full ML equations for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ does not require knowledge of σ^2. The profile likelihood is then

$$L_P\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{N}{2}}$$
$$\times \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right)\right].$$

It can be shown readily that the maximizer of this profile likelihood is identical to the ML estimator of σ^2, as it should be. The identity between the ML and the partial ML estimators of the nuisance parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ indicates that these must be consistent, so the profile likelihood can be used as a full likelihood to obtain inferences about σ^2. For example, the information measure based on the profile likelihood is

$$I_P\left(\sigma^2\right) = E\left[-\frac{\partial^2 \ln L_P\left(\sigma^2|\mathbf{y}\right)}{\left(\partial\sigma^2\right)^2}\right]$$

$$= -\frac{N}{2\sigma^4} + \frac{E\left[\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right)\right]}{\sigma^6}$$

$$= -\frac{N}{2\sigma^4} + \frac{(N-p)\,\sigma^2}{\sigma^6} = \frac{N}{2\sigma^4}\left(1 - \frac{2p}{N}\right).$$

The expectation taken above can be evaluated using (4.41). Note that this measure indicates that account is taken of the information lost in eliminating the nuisance parameters. Observe also that there is information, provided that $p < N/2$. It is of interest to compare the information in the profile likelihood with the information resulting from the marginal likelihood, as given in (4.46) and reexpressible as

$$I_M\left(\sigma^2\right) = \frac{N}{2\sigma^4}\left(1 - \frac{p}{N}\right).$$

If one accepts the marginal likelihood as the "correct" way of accounting for nuisance parameters, this example illustrates that profile likelihoods do not always do so in the same manner. ■

4.5 Analysis of a Multinomial Distribution

The multinomial sampling process was applied previously in connection with the evaluation of the hypothesis that a population is in Hardy–Weinberg equilibrium; see Example 4.3. Inferences about the parameters of a multinomial distribution will be dealt with here in a more general manner. The objective is to illustrate the application of ML methods to problems other than those arising in the classical linear model.

The data are counts n_i $(i = 1, 2, ..., T)$ in each of T mutually exclusive and exhaustive classes. The draws are made independently of each other from the same distribution, the only restriction being that the total number of draws, $n = \sum_{i=1}^{T} n_i$, is fixed by sampling. The unknown probability that an observation falls in class i is denoted as θ_i. These unknown parameters are subject to the natural restriction that $\sum_{i=1}^{T} \theta_i = 1$, because the probability that an observation falls in at least one class must be equal

to 1, and this observation cannot fall in two or more classes. As seen in Chapter 1, under multinomial sampling the joint probability function of the observations is

$$\Pr\left(n_1, n_2, ..., n_T | \theta_1, \theta_2, ..., \theta_T\right) = n! \prod_{i=1}^{T} \frac{\theta_i^{n_i}}{n_i!}.$$

Given n, there are only $T-1$ independent n_i's so this is a multivariate distribution of dimension $T-1$, having $T-1$ free parameters θ_i. When $T=2$, the multinomial distribution yields the binomial distribution as a particular case. In order to implement the ML machinery, the means and variances of the n_i's, as well as their covariances, are needed. These are:

$$E\left(n_i\right) = n\theta_i,$$

$$Var\left(n_i\right) = n\theta_i\left(1-\theta_i\right),$$

and

$$Cov\left(n_i, n_j\right) = -n\theta_i\theta_j.$$

The likelihood function is

$$L\left(\theta_1, \theta_2, ..., \theta_T | n_1, n_2, ..., n_T\right) \propto \prod_{i=1}^{T} \theta_i^{n_i}. \tag{4.52}$$

Because there are only $T-1$ free parameters and independent counts, $\theta_T = 1 - \sum_{i=1}^{T-1} \theta_i$ and $n_T = n - \sum_{i=1}^{T-1} n_i$. Employing this in the likelihood function and taking logs, the log-likelihood is, apart from an additive constant

$$\begin{aligned} &l\left(\theta_1, \theta_2, ..., \theta_{T-1} | n_1, n_2, ..., n_T\right) \\ &= \sum_{i=1}^{T-1} n_i \ln\left(\theta_i\right) + \left(n - \sum_{i=1}^{T-1} n_i\right) \ln\left(1 - \sum_{i=1}^{T-1} \theta_i\right). \end{aligned} \tag{4.53}$$

Differentiating with respect to θ_i gives the score

$$\begin{aligned} \frac{\partial l\left(\theta_1, \theta_2, ..., \theta_{T-1} | n_1, n_2, ..., n_T\right)}{\partial \theta_i} &= \frac{n_i}{\theta_i} - \frac{\left(n - \sum_{i=1}^{T-1} n_i\right)}{\left(1 - \sum_{i=1}^{T-1} \theta_i\right)} \\ &= \frac{n_i}{\theta_i} - \frac{n_T}{\theta_T}, \quad i = 1, 2, ..., T-1. \end{aligned} \tag{4.54}$$

Setting this to 0 gives

$$\widehat{\theta}_i = \frac{n_i}{n_T}\widehat{\theta}_T, \quad i = 1, 2, ..., T-1.$$

Because the ML estimates must be constrained to reside in the interior of the parameter space, it must be that $\sum_{i=1}^{T} \widehat{\theta}_i = 1$. Summing the above over the T ML estimates of probabilities gives

$$\widehat{\theta}_T = \frac{n_T}{n}.$$

Hence,

$$\widehat{\theta}_i = \frac{n_i}{n}, \quad i = 1, 2, ..., T, \tag{4.55}$$

so the proportions of observations falling in class i gives the ML estimates of the corresponding probability directly. It can be verified that the estimator is unbiased because

$$E\left(\widehat{\theta}_i\right) = E\left(\frac{n_i}{n}\right) = \frac{n\theta_i}{n} = \theta_i.$$

Differentiating (4.54) again with respect to θ_i (taking into account the fact that the last parameter is a function of the $T-1$ probabilities for the preceding categories):

$$\frac{\partial^2 l\left(\theta_1, \theta_2, ..., \theta_{T-1} | n_1, n_2, ..., n_T\right)}{\left(\partial\theta_i\right)^2} = -\frac{n_i}{\theta_i^2} - \frac{n_T}{\left(1 - \sum_{i=1}^{T-1} \theta_i\right)^2},$$

$$i = 1, 2, ..., T-1.$$

and

$$\frac{\partial^2 l\left(\theta_1, \theta_2, ..., \theta_{T-1} | n_1, n_2, ..., n_T\right)}{\partial\theta_i \, \partial\theta_j} = -\frac{n_T}{\left(1 - \sum_{i=1}^{T-1} \theta_i\right)^2},$$

$$i, j = 1, 2, ..., T-1, \; i \neq j.$$

The second derivatives multiplied by -1 give the elements of the information matrix about the free parameters of the model. Taking expectations yields Fisher's information matrix, with elements

$$I\left(i, i\right) = n\left(\frac{1}{\theta_i} + \frac{1}{\theta_T}\right), \quad i = 1, 2, ..., T-1, \tag{4.56}$$

and:

$$I\left(i, j\right) = \frac{n}{\theta_T}, \quad i, j = 1, 2, ..., T-1, \; i \neq j. \tag{4.57}$$

The information is proportional to sample size. The inverse of the information matrix gives the asymptotic variance covariance matrix of the ML estimates. This inverse can be shown to have elements

$$v\left(i, i\right) = \frac{1}{n}\theta_i\left(1 - \theta_i\right), \quad i = 1, 2, ..., T-1, \tag{4.58}$$

$$v\left(i, j\right) = -\frac{1}{n}\theta_i\theta_j, \quad i, j = 1, 2, ..., T-1, \; i \neq j. \tag{4.59}$$

These coincide with the exact variances and covariances, as given earlier. Hence, the ML estimator attains the Cramér–Rao lower bound, and is a minimum variance unbiased estimator in this case. The correlation between parameter estimates is

$$Corr\left(\widehat{\theta}_i, \widehat{\theta}_j\right) = -\sqrt{\frac{\theta_i \theta_j}{(1-\theta_i)(1-\theta_j)}}.$$

When $T = 2$, the multinomial model reduces to binomial sampling. The ML estimator of the probability of response in the first class, say, is the proportion of observations falling in this category. The sampling variance of the estimate is

$$Var\left(\widehat{\theta}\right) = \frac{\theta(1-\theta)}{n}.$$

In all cases, because the θ's are not known, parameter values must be replaced by the ML estimates to obtain an approximation to the asymptotic distribution. The goodness of this approximation is expected to improve as n increases.

Example 4.5 *Analysis of a trichotomy*

Suppose the data consist of $n = 100$ observations collected at random from a homogeneous populations of experimental chickens. Each bird is scored for the presence or absence of leg deformities and, within deformed individuals, there are two modalities. Hence, $T = 3$. Suppose the outcome of the experiment is $n_1 = 20$, $n_2 = 35$, and $n_3 = 45$. The objective is to obtain ML estimates of the prevalence of each of the three modalities, and of a non-linear function of the associated probabilities. Here we work with θ_1 and θ_2 as free parameters, because θ_3 is redundant. From (4.53) the log-likelihood, apart from a constant, is

$$l(\theta_1, \theta_2 | n_1, n_2, n_3) = 20 \ln(\theta_1) + 35 \ln(\theta_2) + 45 \ln(1 - \theta_1 - \theta_2).$$

Using (4.55) the ML estimates are $\widehat{\theta}_1 = .20$, $\widehat{\theta}_2 = .35$, and $\widehat{\theta}_3 = 1 - \widehat{\theta}_1 - \widehat{\theta}_2 = 1 - .20 - .35 = .45$. From (4.58), the asymptotic standard deviation of the estimate (equal to the standard deviation of the exact sampling distribution in this case) of $\widehat{\theta}_1$ can be estimated as

$$\sqrt{\frac{(.20)(.80)}{100}} = 0.04.$$

Based on the ML estimator, a confidence region based on two standard deviations is $.20 \pm .08$. The interval $(0.12 - 0.28)$ indicates that inferences about the true value of θ_1 are not very sharp in a sample of this size. How sharp a confidence region should be depends on the problem in question. A similar calculation can be carried out for the probabilities of falling into

any of the other two classes.
Suppose now that inferences are sought using the logit transform

$$w_i = \ln \frac{\theta_i}{1-\theta_i}$$

with inverse

$$\theta_i = \frac{\exp(w_i)}{1+\exp(w_i)}.$$

The logit is interpretable as a log-odds ratio, that is, the ratio between the probabilities of "response" and of "nonresponse" measured on a logarithmic scale. The logit is positive when the probability of "response" is larger than the probability of the complementary event, and negative otherwise. While the parameter space of θ_i is bounded between 0 and 1, that for the logit is $-\infty < w_i < \infty$. The ML estimator of the logit is

$$\widehat{w}_i = \ln \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i}.$$

Here,

$$\begin{aligned}
\widehat{w}_1 &= \ln \frac{.20}{.80} = -1.3863, \\
\widehat{w}_2 &= \ln \frac{.35}{.65} = -0.6190, \\
\widehat{w}_3 &= \ln \frac{.45}{.55} = -0.2007.
\end{aligned}$$

The variance of the asymptotic distribution of $\widehat{w}_i$ can be deduced from the variance–covariance matrix of the asymptotic distribution of the ML estimates of the probabilities. Because this is a multi-parameter problem, one needs to form the information matrix about the logits. Using (3.74), and working with two "free" logits only

$$\begin{aligned}
\mathbf{I}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}\right) &= \begin{bmatrix} \frac{d\theta_1}{dw_1} & \frac{d\theta_2}{dw_1} \\ \frac{d\theta 1}{dw_2} & \frac{d\theta_2}{dw_2} \end{bmatrix} \mathbf{I}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right) \begin{bmatrix} \frac{d\theta_1}{dw_1} & \frac{d\theta_1}{dw_2} \\ \frac{d\theta_2}{dw_1} & \frac{d\theta_2}{dw_2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{d\theta_1}{dw_1} & 0 \\ 0 & \frac{d\theta_2}{dw_2} \end{bmatrix} \mathbf{I}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right) \begin{bmatrix} \frac{d\theta_1}{dw_1} & 0 \\ 0 & \frac{d\theta_2}{dw_2} \end{bmatrix}.
\end{aligned}$$

The asymptotic variance of the ML estimates of the logits is then:

$$Var\begin{bmatrix} \widehat{w}_1 \\ \widehat{w}_2 \end{bmatrix} = \begin{bmatrix} \frac{d\theta_1}{dw_1} & 0 \\ 0 & \frac{d\theta_2}{dw_2} \end{bmatrix}^{-1} \mathbf{I}^{-1}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right) \begin{bmatrix} \frac{d\theta_1}{dw_1} & 0 \\ 0 & \frac{d\theta_2}{dw_2} \end{bmatrix}^{-1}.$$

The derivatives are:

$$\frac{d\theta_i}{dw_i} = \frac{\exp(w_i)}{1+\exp(w_i)}\left[1 - \frac{\exp(w_i)}{1+\exp(w_i)}\right] = \theta_i(1-\theta_i),$$

so, using the asymptotic covariance matrix of the ML estimates of the probabilities, with elements as in expressions (4.58) and (4.59),

$$Var\begin{bmatrix}\widehat{w}_1\\ \widehat{w}_2\end{bmatrix} = \frac{1}{n}\begin{bmatrix}\dfrac{1}{\theta_1(1-\theta_1)} & -\dfrac{1}{(1-\theta_1)(1-\theta_2)}\\ -\dfrac{1}{(1-\theta_1)(1-\theta_2)} & \dfrac{1}{\theta_2(1-\theta_2)}\end{bmatrix}.$$

The asymptotic correlation between the ML estimates of the logits is then:

$$Corr(\widehat{w}_1, \widehat{w}_2) = -\sqrt{\frac{\theta_1\theta_2}{(1-\theta_1)(1-\theta_2)}}.$$

In this example, the estimated variance–covariance matrix of the ML estimates of the logits is

$$\widehat{Var}\begin{bmatrix}\widehat{w}_1\\ \widehat{w}_2\end{bmatrix} = \frac{1}{100}\begin{bmatrix}\frac{1}{.16} & -\frac{1}{.07}\\ -\frac{1}{.07} & \frac{1}{.2275}\end{bmatrix} = \begin{bmatrix}.0625 & -.1429\\ -.1429 & .0440\end{bmatrix}.$$

■

Example 4.6 *Estimation of allele frequencies with inbreeding*

This example is adapted from Weir (1996). From population genetics theory, it is known that if a population is in Hardy–Weinberg equilibrium, and inbreeding is practiced, this causes a reduction in the frequency of heterozygotes and a corresponding increase in homozygosity. Consider a locus with segregating alleles A and a, so that three genotypes are possible: AA, Aa, and aa. As seen previously, if θ is the frequency of allele A, with Hardy–Weinberg equilibrium, the frequency of the three different genotypes is $\Pr(AA) = \theta^2$, $\Pr(Aa) = 2\theta(1-\theta)$, and $\Pr(aa) = (1-\theta)^2$. With inbreeding, the genotypic distribution changes to

$$\begin{aligned}\Pr(AA) &= \theta^2 + \theta(1-\theta)f,\\ \Pr(Aa) &= 2\theta(1-\theta)(1-f),\\ \Pr(aa) &= (1-\theta)^2 + \theta(1-\theta)f,\end{aligned}$$

where f is the inbreeding coefficient or, equivalently, the fractional reduction in heterozygosity. Suppose that n_{AA}, n_{Aa}, and n_{aa} individuals having

genotypes AA, Aa, and aa, respectively, are observed, and that the total number of observations is fixed by the sampling scheme. The unknown parameters are the allele frequency and the coefficient of inbreeding. If the individuals are sampled at random from the same conceptual population, a multinomial sampling model would be reasonable. The likelihood function can be expressed as

$$L(\theta, f) \propto \theta^{n_{AA}+n_{Aa}} [\theta + (1-\theta) f]^{n_{AA}} [(1-f)]^{n_{Aa}} \times [(1-\theta) + \theta f]^{n_{aa}} (1-\theta)^{n_{Aa}+n_{aa}}$$

where $1-\theta$ is the allelic frequency of a. The score vector has two elements, $\partial l/\partial\theta$ and $\partial l/\partial f$ where, as usual, l is the log-likelihood function. The elements are

$$\frac{\partial l}{\partial \theta} = \frac{n_{AA}+n_{Aa}}{\theta} + \frac{n_{AA}(1-f)}{\theta+(1-\theta)f} - \frac{n_{aa}(1-f)}{(1-\theta)+\theta f} - \frac{n_{Aa}+n_{aa}}{(1-\theta)}$$

and

$$\frac{\partial l}{\partial f} = \frac{n_{AA}(1-\theta)}{\theta+(1-\theta)f} - \frac{n_{Aa}}{(1-f)} + \frac{n_{aa}\theta}{(1-\theta)+\theta f}.$$

Setting the gradient to zero leads to a nonlinear system in θ and f that does not have an explicit solution. Hence, second derivatives are needed not only to complete the ML analysis, but to construct an iterative procedure for obtaining the estimates. The second derivatives are

$$\frac{\partial^2 l}{(\partial\theta)^2} = -\frac{n_{AA}+n_{Aa}}{\theta^2} - \frac{n_{AA}(1-f)^2}{[\theta+(1-\theta)f]^2} - \frac{n_{aa}(1-f)^2}{[(1-\theta)+\theta f]^2} - \frac{n_{Aa}+n_{aa}}{(1-\theta)^2},$$

$$\frac{\partial^2 l}{(\partial f)^2} = -\frac{n_{AA}(1-\theta)^2}{[\theta+(1-\theta)f]^2} - \frac{n_{Aa}}{(1-f)^2} - \frac{n_{aa}\theta^2}{[(1-\theta)+\theta f]^2},$$

and

$$\frac{\partial^2 l}{\partial\theta\partial f} = -\frac{n_{AA}}{\theta+(1-\theta)f} - \frac{n_{AA}(1-\theta)(1-f)}{[\theta+(1-\theta)f]^2} + \frac{n_{aa}}{(1-\theta)+\theta f} + \frac{n_{aa}(1-f)\theta}{[(1-\theta)+\theta f]^2} = -\frac{n_{AA}}{[\theta+(1-\theta)f]^2} + \frac{n_{aa}}{[(1-\theta)+\theta f]^2}.$$

Genotype	Phenotype	Observed counts	Frequency
AA	A	n_A	p_A^2
AO	A		$2p_Ap_O$
AB	AB	n_{AB}	$2p_Ap_B$
BB	B	n_B	p_B^2
BO	B		$2p_Bp_O$
OO	O	n_O	p_O^2

TABLE 4.1. Frequency of genotypes and phenotypes of ABO blood group data.

From the first and second derivatives, the Newton–Raphson algorithm can be formed as presented in (4.1). The expected second derivatives are obtained by replacing n_{AA}, n_{Aa}, and n_{aa} by their expectations. For example

$$E\left(n_{Aa}\right) = 2\theta\left(1-\theta\right)\left(1-f\right)\left(n_{AA}+n_{Aa}+n_{aa}\right).$$

To illustrate, suppose that $n_{AA} = 100$, $n_{Aa} = 200$, and $n_{aa} = 200$. The ML estimates are $\widehat{\theta} = .40$ and $\widehat{f} = .1667$, after round-off. From the Newton–Raphson algorithm, the observed information matrix can be estimated as

$$\widehat{\mathbf{I}}_o\left(\theta, f\right) = \left[\begin{array}{cc} -\dfrac{\partial^2 l}{\left(\partial\theta\right)^2} & -\dfrac{\partial^2 l}{\partial\theta\partial f} \\ -\dfrac{\partial^2 l}{\partial\theta\partial f} & -\dfrac{\partial^2 l}{\left(\partial f\right)^2} \end{array}\right]_{\substack{\theta=\widehat{\theta} \\ f=\widehat{f}}}.$$

This can be used in lieu of the expected information matrix to estimate the asymptotic variance–covariance matrix of the estimates. In this example

$$\widehat{\mathbf{I}}_o^{-1}\left(\theta, f\right) = \frac{1}{1000}\left[\begin{array}{cc} .28001 & .02777 \\ .02777 & 1.98688 \end{array}\right].$$

The asymptotic correlation between ML estimates is approximately .04. ■

Example 4.7 *ABO blood groups*

Consider the following blood group data in Table 4.1. With three alleles, A, B, and O there are six genotypes but only four phenotypic classes can be observed. The expected frequency of each genotype in the last column is derived assuming Hardy–Weinberg equilibrium. The problem at hand is to infer p_A, p_B and p_O, the frequency of alleles A, B, and O, respectively, subject to the constraint $p_A + p_B + p_O = 1$.

The observed data is $\mathbf{y}' = (n_A, n_{AB}, n_B, n_O)$. The log-likelihood is given by:

$$\begin{aligned} l\left(p_A, p_B|\mathbf{y}\right) \propto\ & n_A \ln\left[p_A\left(2-p_A-2p_B\right)\right] + n_{AB}\ln\left[2p_Ap_B\right] \\ & + n_B\ln\left[p_B\left(2-p_B-2p_A\right)\right] + 2n_O\ln\left[\left(1-p_A-p_B\right)\right]. \end{aligned} \tag{4.60}$$

Differentiating with respect to p_A and p_B yields the nonlinear system of equations

$$\frac{\partial l\left(p_A, p_B|\mathbf{y}\right)}{\partial p_A} = \frac{n_{AB}}{p_A} + \frac{n_A\left(2-2p_A-2p_B\right)}{p_A\left(2-p_A-2p_B\right)} - \frac{2n_B}{2-2p_A-p_B} - \frac{2n_O}{1-p_A-p_B}, \tag{4.61}$$

$$\frac{\partial l\left(p_A, p_B|\mathbf{y}\right)}{\partial p_B} = \frac{n_{AB}}{p_B} + \frac{n_B\left(2-2p_A-2p_B\right)}{p_B\left(2-2p_A-p_B\right)} - \frac{2n_A}{2-p_A-2p_B} - \frac{2n_O}{1-p_A-p_B}. \tag{4.62}$$

A solution can be obtained using Newton–Raphson. This requires the following second derivatives

$$\frac{\partial^2 l\left(p_A, p_B|\mathbf{y}\right)}{\left(\partial p_A\right)^2} = \frac{n_A\left(2-2p_A-2p_B\right)}{p_A\left(p_A+2p_B-2\right)^2} + \frac{2n_A}{p_A\left(p_A+2p_B-2\right)} - \frac{n_A\left(2p_A+2p_B-2\right)}{p_A^2\left(p_A+2p_B-2\right)} - \frac{n_{AB}}{p_A^2} - \frac{2n_O}{\left(p_A+p_B-1\right)^2} - \frac{4n_B}{\left(2p_A+p_B-2\right)^2},$$

$$\frac{\partial^2 l\left(p_A, p_B|\mathbf{y}\right)}{\left(\partial p_B\right)^2} = \frac{n_B\left(2-2p_A-2p_B\right)}{p_B\left(2p_A+p_B-2\right)^2} + \frac{2n_B}{p_B\left(2p_A+p_B-2\right)} - \frac{n_B\left(2p_A+2p_B-2\right)}{p_B^2\left(2p_A+p_B-2\right)} - \frac{n_{AB}}{p_B^2} - \frac{2n_O}{\left(p_A+p_B-1\right)^2} - \frac{4n_A}{\left(p_A+2p_B-2\right)^2},$$

$$\frac{\partial^2 l\left(p_A, p_B|\mathbf{y}\right)}{\partial p_A \partial p_B} = -\frac{2n_A}{\left(2-p_A-2p_B\right)^2} - \frac{2n_B}{\left(2-2p_A-p_B\right)^2} - \frac{2n_O}{\left(1-p_A-p_B\right)^2}.$$

Suppose the data are $n_A = 725$, $n_{AB} = 72$, $n_B = 258$, $n_O = 1073$. Using these expressions in (4.1) yields, at convergence, the ML estimates: $\widehat{p}_A = 0.2091$ and $\widehat{p}_B = 0.0808$. The observed information matrix evaluated at the ML estimates is

$$I\left(\widehat{p}_A, \widehat{p}_B|\mathbf{y}\right) = \begin{bmatrix} 23,215.5 & 5,031.14 \\ 5,031.14 & 56,009.4 \end{bmatrix},$$

resulting in an estimate of the asymptotic covariance matrix equal to:

$$Var\left(\widehat{p}_A, \widehat{p}_B|\mathbf{y}\right) = \left[I\left(\widehat{p}_A, \widehat{p}_B|\mathbf{y}\right)\right]^{-1} = 10^{-6}\begin{bmatrix} 43.930 & -3.946 \\ -3.946 & 18.209 \end{bmatrix}.$$

■

4.5.1 Amount of Information per Observation

An important problem in experimental genetics is the evaluation of designs for estimation of genetic parameters. In this context, a relevant question is the measurement of the average amount of information about a parameter per experimental unit included in the trial. It will be shown here how to assess this when the entity of interest is yet another parameter affecting the probabilities that govern the multinomial distribution. Subsequently, two examples are given where the target parameter is the recombination fraction between loci.

Under multinomial sampling, the likelihood is given in (4.52). Suppose now that the θ's depend on some scalar parameter α. To emphasize this dependency, the probabilities are denoted as $\theta_i(\alpha)$. The likelihood is then viewed as a function of α, and the score function is

$$\frac{dl(\alpha)}{d\alpha} = \sum_{i=1}^{T} \frac{n_i}{\theta_i(\alpha)} \frac{d\theta_i(\alpha)}{d\alpha}. \tag{4.63}$$

The amount of information about α contained in the sample is denoted as:

$$\begin{aligned} I_n(\alpha) &= E\left[-\frac{d^2 l(\alpha)}{(d\alpha)^2}\right] \\ &= -E\left\{\sum_{i=1}^{T}\left[-\frac{n_i}{\theta_i^2(\alpha)}\left(\frac{d\theta_i(\alpha)}{d\alpha}\right)^2 + \frac{n_i}{\theta_i(\alpha)}\frac{d^2\theta_i(\alpha)}{(d\alpha)^2}\right]\right\} \\ &= \left\{\sum_{i=1}^{T}\left[\frac{1}{\theta_i(\alpha)}\left(\frac{d\theta_i(\alpha)}{d\alpha}\right)^2 - \frac{d^2\theta_i(\alpha)}{(d\alpha)^2}\right]\right\} E\left[\frac{n_i}{\theta_i(\alpha)}\right]. \end{aligned} \tag{4.64}$$

Now, $E\left[n_i/\theta_i(\alpha)\right] = n\theta_i(\alpha)/\theta_i(\alpha) = n$. Using this in the preceding expression, distributing the sum, and noting that

$$\sum_{i=1}^{T} \frac{d^2\theta_i(\alpha)}{(d\alpha)^2} = \frac{d^2}{(d\alpha)^2}\sum_{i=1}^{T}\theta_i(\alpha) = \frac{d^2}{(d\alpha)^2}1 = 0,$$

one arrives at

$$I_n(\alpha) = n\sum_{i=1}^{T}\frac{1}{\theta_i(\alpha)}\left(\frac{d\theta_i(\alpha)}{d\alpha}\right)^2. \tag{4.65}$$

Hence, the expected amount of information per observation is

$$I_1(\alpha) = \sum_{i=1}^{T} \frac{1}{\theta_i(\alpha)} \left(\frac{d\theta_i(\alpha)}{d\alpha} \right)^2. \tag{4.66}$$

Example 4.8 *Estimating the recombination rate between two loci from matings involving coupling heterozygotes*

This example is patterned after Rao (1973). Consider two loci, each with two alleles. Let the alleles be A and a at the first locus, and B and b at the second locus. Suppose that individuals that are heterozygous at both loci are available, and that such heterozygotes originate from a cross between $AABB$ and $aabb$ parents. In this case, individuals are called coupling heterozygotes, and their genotype is indicated as AB/ab. This means that they came from AB and ab gametes only. Coupling heterozygotes are crossed inter se, and the objective is to estimate the probability of recombination (or recombination rate) between the two loci, denoted here as α. The distribution of gametes produced by coupling heterozygotes can be deduced by considering that recombinant gametes arise with probability α, whereas the complementary event (lack of recombination) has probability $1-\alpha$. Then the gametic distribution is

$$\Pr(AB) = \Pr(ab) = \frac{1-\alpha}{2}, \;\; \Pr(Ab) = \Pr(aB) = \frac{\alpha}{2}.$$

The random union of these gametes produces 16 possible genotypes, i.e., $AABB, ..., aabb$, nine of which are distinguishable at the genotypic level, assuming that maternal or paternal origin of the chromosome cannot be traced. For example, AB/Ab cannot be distinguished from Ab/AB. The focus of inference is the parameter α and to evaluate the efficiency of this and of another design in terms of expected information per observation. The data are genotypic counts scored in the progeny from the appropriate matings. Table 4.2 provides the distribution of genotypes in matings between coupling heterozygotes, as well as the expected amount of information per observation (for each genotype) calculated with formula (4.66).

Using (4.66), the expected amount of information per observation is obtained by summing elements in the third column of Table 4.2, yielding

$$I_1(\alpha) = 4 + \frac{8\left(\frac{1}{2} - \alpha\right)^2}{\alpha(1-\alpha)} + \frac{(4\alpha - 2)^2}{(4\alpha^2 - 4\alpha + 2)}.$$

For example, if the loci are in different chromosomes, $\alpha = 1/2$ and $I_1(\alpha) = 4$. As the loci become more closely linked, this measure of information increases and tends to infinity as $\alpha \to 0$. From Table 4.2, the log-likelihood

Genotype	$\theta_i(\alpha)$	$I_1^{[i]}(\alpha)$	Counts
$AABB$	$\frac{1}{4}(1-\alpha)^2$	1	n_1
$AaBb$	$\left(\frac{1}{2}-\alpha+\alpha^2\right)$	$(4\alpha-2)^2/\left(4\alpha^2-4\alpha+2\right)$	n_2
$AABb$	$\frac{1}{2}\alpha(1-\alpha)$	$2\left(\frac{1}{2}-\alpha\right)^2/\left(\alpha-\alpha^2\right)$	n_3
$AaBB$	$\frac{1}{2}\alpha(1-\alpha)$	$2\left(\frac{1}{2}-\alpha\right)^2/\left(\alpha-\alpha^2\right)$	n_4
$aabb$	$\frac{1}{4}(1-\alpha)^2$	1	n_5
$Aabb$	$\frac{1}{2}\alpha(1-\alpha)$	$2\left(\frac{1}{2}-\alpha\right)^2/\left(\alpha-\alpha^2\right)$	n_6
$aaBb$	$\frac{1}{2}\alpha(1-\alpha)$	$2\left(\frac{1}{2}-\alpha\right)^2/\left(\alpha-\alpha^2\right)$	n_7
$AAbb$	$\frac{1}{4}\alpha^2$	1	n_8
$aaBB$	$\frac{1}{4}\alpha^2$	1	n_9

TABLE 4.2. Probability distribution of genotypes in progeny from crosses between coupling heterozygotes and contribution of each genotype to expected information per observation (third column).

of the parameter α can be deduced to be

$$l(\alpha) = (n_1+n_5)\ln\left[\frac{(1-\alpha)^2}{4}\right] + (n_3+n_4+n_6+n_7)\ln\left[\frac{\alpha(1-\alpha)}{2}\right] + (n_8+n_9)\ln\left(\frac{\alpha^2}{4}\right) + n_2\left\{\left[\frac{(1-\alpha)^2}{2}+\frac{\alpha^2}{2}\right]\right\}.$$

Given data, the log-likelihood can be maximized numerically. Suppose that, out of $n = 65$ descendants from matings involving coupling heterozygote parents, the genotypic count recovered is such that $n_1 + n_5 = 30$, $n_3 + n_4 + n_6 + n_7 = 8$, $n_8 + n_9 = 9$, and $n_2 = 18$. The ML estimate of the recombination rate is, in this case, $\widehat{\alpha} = .2205$. Using second derivatives, an approximation to the asymptotic standard error based on observed information is $\sqrt{\widehat{Var}(\widehat{\alpha})} = .0412$. The expected information per observation is estimated as $I_1(\widehat{\alpha}) = 8.58743$. Thus, with $n = 65$, (4.65) gives $I_{65}(\widehat{\alpha}) = (65)(8.58743) = 558.1830$, as an estimate of the expected information about the recombination rate parameter. The corresponding estimated asymptotic standard error is $\sqrt{558.1830^{-1}} = .0423$. There is good agreement between the standard errors based on observed and expected information. ■

Example 4.9 *Estimating the recombination rate between two loci from a backcross*

An alternative genetic design for estimating α consists of crossing the coupling heterozygotes AB/ab to ab/ab (i.e., one of the parental genotypes). This is called a backcross. The nonrecombinant types appearing in the

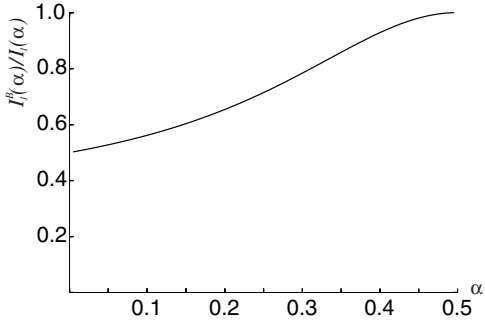

FIGURE 4.1. Plot of $I_1^B(\alpha)/I_1(\alpha)$ as a function of recombination α.

progeny with a total probability of $1-\alpha$, are AB/ab and ab/ab. The recombinant types (with probability α) are Ab/ab and aB/ab. Using (4.66), the expected information per observation obtained from this design is

$$I_1^B(\alpha) = \sum_{i=1}^{T} \frac{1}{\theta_i(\alpha)} \left(\frac{d\theta_i(\alpha)}{d\alpha}\right)^2$$
$$= 2\frac{1}{\frac{(1-\alpha)}{2}}\left(-\frac{1}{2}\right)^2 + 2\frac{1}{\frac{\alpha}{2}}\left(\frac{1}{2}\right)^2 = \frac{1}{\alpha(1-\alpha)},$$

where the superscript B denotes the backcross design. If the loci are in different chromosomes $I_1^B(\alpha) = 4$ and, when α goes to zero, then $I_1^B(\alpha) \to \infty$, as in the preceding design. A plot of $I_1^B(\alpha)/I_1(\alpha)$ as a function of α is given in Figure 4.1. It is seen that a design involving matings between coupling heterozygotes is always more efficient, although the designs differ little for values of α near $1/2$. ■

4.6 Analysis of Linear Logistic Models

Response variables that are categorical in expression often arise in genetic analysis. For example, Wright (1934) analyzed the inheritance of the number of digits in guinea pigs and proposed relating the expression of this variable to an underlying normal process. It would be at this level where gene substitutions operate. This model, known as the threshold model, was elaborated further in quantitative genetics by Dempster and Lerner (1950), Falconer (1965), and Gianola (1982), among others. Here, the special case of binary response variables is considered. Attention is given to the development of ML machinery appropriate for some situations of interest in quantitative genetic applications.

The point of departure of a likelihood analysis is the probability mass function of the data, given the parameters, which is developed below. Sup-

pose there is an underlying or latent unobservable variable, l, often called liability, and that the categories of response (two such categories assumed here) result from the value of l relative to a fixed threshold t. (The same letter l, is used for liability and for log-likelihood, although the latter is indexed by the parameters of the model. The common use of l should therefore not lead to ambiguities). Let the dichotomy be "survival" versus "death", say. If $l < t$ then the individual survives and the binary variable takes the value $Y = 1$. If $l \geq t$ the individual dies and $Y = 0$. Denote the liability associated with datum i as l_i, and suppose that the underlying variate is related to an unknown parameter vector $\boldsymbol{\beta}$ (of order $p \times 1$) in terms of the linear model

$$l_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, 2, ..., N, \tag{4.67}$$

where $\mathbf{x}_i'$ is the ith row of the known, nonstochastic $N \times p$ matrix of explanatory variables $\mathbf{X}$ and e_i is a random residual with p.d.f. $p(e_i)$. Assume that the residuals are independent and identically distributed. The probability of survival of individual i (which is the p.m.f. of the random variable Y_i) is

$$\begin{aligned}\Pr(Y_i = 1|\boldsymbol{\beta}) &= \Pr(l_i < t|\boldsymbol{\beta}) = \Pr(l_i - \mathbf{x}_i'\boldsymbol{\beta} < t - \mathbf{x}_i'\boldsymbol{\beta}|\boldsymbol{\beta}) \\ &= \Pr(e_i < t - \mathbf{x}_i'\boldsymbol{\beta}|\boldsymbol{\beta}) = \int_{-\infty}^{t-\mathbf{x}_i'\boldsymbol{\beta}} p(e_i)\, de_i \\ &= 1 - \int_{t-\mathbf{x}_i'\boldsymbol{\beta}}^{\infty} p(e_i)\, de_i = 1 - \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{t}} p(e_i)\, de_i. \end{aligned} \tag{4.68}$$

The last equality above requires that e_i is symmetrically distributed around 0. The liabilities cannot be observed, and a convenient origin is to set the value of the threshold to 0. Hence, the scale is one of deviations from the threshold. This constraint makes the likelihood model identifiable and the Hessian becomes negative definite.

It is interesting to note that, although in the underlying scale l_i changes with $\mathbf{x}_i$ at a constant rate, this is not so at the level of the probabilities. This is verified by noting that

$$\frac{\partial l_i}{\partial \mathbf{x}_i} = \boldsymbol{\beta},$$

whereas from expression (4.68)

$$\begin{aligned}\frac{\partial \Pr(y = 1|\boldsymbol{\beta})}{\partial \mathbf{x}_i} &= \frac{\partial}{\partial \mathbf{x}_i}\left[1 - \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} p(e_i)\, de_i\right] \\ &= -\frac{\partial}{\partial \mathbf{x}_i'\boldsymbol{\beta}}\left[\int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} p(e_i)\, de_i\right]\frac{\partial \mathbf{x}_i'\boldsymbol{\beta}}{\partial \mathbf{x}_i} \\ &= -p(\mathbf{x}_i'\boldsymbol{\beta})\,\boldsymbol{\beta}. \end{aligned} \tag{4.69}$$

The change is not constant and depends on the value of the explanatory vector $\mathbf{x}_i$.

4.6.1 The Logistic Distribution

In the analysis of binary responses (Cox and Snell, 1989), two distributions are often assigned to the residuals. A natural candidate is the normal distribution, as in linear models. However, because the underlying variable cannot be observed, the unit of measurement is set to be equal to the standard deviation, so a $N(0,1)$ residual distribution is adopted. This leads to the probit model, and parameter estimates must be interpreted as deviations from the threshold in units of standard deviation. Alternatively, a logistic distribution can be adopted, because it leads to somewhat simpler algebraic expressions. The parameters are also in standard deviation units. It should be understood, however, that if mechanistic considerations dictate a normal distribution, this should be preferred.

Consider a random variable Z having, as density function,

$$p(z) = \frac{\exp(z)}{[1+\exp(z)]^2}, \quad -\infty < z < \infty. \tag{4.70}$$

Then Z has a logistic distribution with $E(Z) = 0$ and $Var(Z) = \pi^2/3$. Hence, for a constant k,

$$\begin{aligned}\Pr(Z < k) &= \int_{-\infty}^{k} p(z)\,dz = \int_{-\infty}^{k} \frac{\exp(z)}{[1+\exp(z)]^2}dz \\ &= \frac{\exp(k)}{[1+\exp(k)]}.\end{aligned} \tag{4.71}$$

If the residual distribution in (4.67) is logistic, the probability of survival in (4.68) is

$$\begin{aligned}\Pr(Y=1|\boldsymbol{\beta}) &= 1 - \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} p(e_i)\,de_i = 1 - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})]} \\ &= [1+\exp(\mathbf{x}_i'\boldsymbol{\beta})]^{-1} = p_i(\boldsymbol{\beta}),\end{aligned} \tag{4.72}$$

and the probability of death is

$$\Pr(Y=0|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})]} = 1 - p_i(\boldsymbol{\beta}). \tag{4.73}$$

4.6.2 Likelihood Function under Bernoulli Sampling

The data consist of binary responses on N subjects and inferential interest is on $\boldsymbol{\beta}$, the location vector of the underlying distribution. Each of the $(0,1)$

outcomes can be viewed as a Bernoulli trial with probability

$$\Pr\left(Y_i = y_i|\boldsymbol{\beta}\right) = \left[\frac{1}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right]^{y_i}\left[\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right]^{1-y_i}, \tag{4.74}$$

with $Y_i = 1$ if the individual survives, and $Y_i = 0$ otherwise. If, given $\boldsymbol{\beta}$, the N responses are mutually independent, the probability of observing the data $\mathbf{y}$ is

$$p\left(\mathbf{y}|\boldsymbol{\beta}\right) = \prod_{i=1}^{N}\left[\frac{1}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right]^{y_i}\left[\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right]^{1-y_i}. \tag{4.75}$$

This is the likelihood function when viewed as a function of $\boldsymbol{\beta}$. The resulting log-likelihood can be written as

$$l\left(\boldsymbol{\beta}|\mathbf{y}\right) = \sum_{i=1}^{N}\left\{\left(1-y_i\right)\mathbf{x}_i'\boldsymbol{\beta} - \ln\left[1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]\right\}. \tag{4.76}$$

The score vector is:

$$\begin{aligned}
\mathbf{l}'\left(\boldsymbol{\beta}|\mathbf{y}\right) &= \sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\beta}}\left\{\left(1-y_i\right)\mathbf{x}_i'\boldsymbol{\beta} - \ln\left[1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]\right\} \\
&= \sum_{i=1}^{N}\left(1-y_i\right)\mathbf{x}_i - \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\mathbf{x}_i \\
&= \sum_{i=1}^{N}\left\{1-y_i-\left[1-p_i\left(\boldsymbol{\beta}\right)\right]\right\}\mathbf{x}_i \\
&= -\sum_{i=1}^{N}\left[y_i - p_i\left(\boldsymbol{\beta}\right)\right]\mathbf{x}_i.
\end{aligned} \tag{4.77}$$

Now let the $N \times 1$ vector of probabilities of survival for the N individuals be $\mathbf{p}\left(\boldsymbol{\beta}\right)$, and observe that

$$\begin{aligned}
\sum_{i=1}^{N}\left[y_i - p_i\left(\boldsymbol{\beta}\right)\right]\mathbf{x}_i &= \left\{\mathbf{x}_1\left[y_1 - p_1\left(\boldsymbol{\beta}\right)\right], ..., \mathbf{x}_N\left[y_N - p_N\left(\boldsymbol{\beta}\right)\right]\right\} \\
&= \mathbf{X}'\left[\mathbf{y} - \mathbf{p}\left(\boldsymbol{\beta}\right)\right].
\end{aligned}$$

The vector $\mathbf{y} - \mathbf{p}\left(\boldsymbol{\beta}\right)$ consists of deviations of the observations from their expectations, that is, residuals in the discrete scale. Using this representation in (4.77) it can be seen that the first-order condition for a maximum is satisfied if

$$\mathbf{X}'\mathbf{p}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{X}'\mathbf{y} \tag{4.78}$$

where $\mathbf{p}\left(\widehat{\boldsymbol{\beta}}\right)$ is the vector of probabilities of survival for the N individuals evaluated at the ML estimator $\widehat{\boldsymbol{\beta}}$, if this exists. The estimating equations (4.78) are not explicit in $\widehat{\boldsymbol{\beta}}$ and must be solved iteratively. The Newton–Raphson algorithm requires second derivatives, and an additional differentiation of the log-likelihood with respect to the parameters gives

$$\begin{aligned} l''(\boldsymbol{\beta}|\mathbf{y}) &= \frac{\partial^2 l(\boldsymbol{\beta}|\mathbf{y})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'} = \frac{\partial}{\partial\boldsymbol{\beta}'}\left\{-\sum_{i=1}^{N}\left[y_i - p_i(\boldsymbol{\beta})\right]\mathbf{x}_i\right\} \\ &= \sum_{i=1}^{N}\mathbf{x}_i\frac{\partial}{\partial\boldsymbol{\beta}'}p_i(\boldsymbol{\beta}). \end{aligned} \tag{4.79}$$

Now,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}'}p_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial\boldsymbol{\beta}'}\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{-1} \\ &= -\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{-2}\exp(\mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i' \\ &= -p_i(\boldsymbol{\beta})\left[1-p_i(\boldsymbol{\beta})\right]\mathbf{x}_i'. \end{aligned}$$

Using this in (4.79)

$$l''(\boldsymbol{\beta}|\mathbf{y}) = -\sum_{i=1}^{N}\mathbf{x}_i p_i(\boldsymbol{\beta})\left[1-p_i(\boldsymbol{\beta})\right]\mathbf{x}_i' = -\mathbf{X}'\mathbf{D}(\boldsymbol{\beta})\,\mathbf{X}, \tag{4.80}$$

where $\mathbf{D}(\boldsymbol{\beta}) = \{p_i(\boldsymbol{\beta})\left[1-p_i(\boldsymbol{\beta})\right]\}$ is an $N \times N$ diagonal matrix. Because the second derivatives do not depend on the observations, the expected information is equal to the observed information in this case. Hence, the Newton–Raphson and the scoring algorithms are identical. From (4.1) the iteration can be represented as

$$\left[\mathbf{X}'\mathbf{D}\left(\boldsymbol{\beta}^{[t]}\right)\mathbf{X}\right]\boldsymbol{\beta}^{[t+1]} = \left[\mathbf{X}'\mathbf{D}\left(\boldsymbol{\beta}^{[t]}\right)\mathbf{X}\right]\boldsymbol{\beta}^{[t]} + \mathbf{X}'\mathbf{v}\left(\boldsymbol{\beta}^{[t]}\right), \tag{4.81}$$

where the vector $\mathbf{v}\left(\boldsymbol{\beta}^{[t]}\right) = \mathbf{p}\left(\boldsymbol{\beta}^{[t]}\right) - \mathbf{y}$. Now let

$$\mathbf{y}^*\left(\boldsymbol{\beta}^{[t]}\right) = \mathbf{X}\boldsymbol{\beta}^{[t]} + \mathbf{D}^{-1}\left(\boldsymbol{\beta}^{[t]}\right)\mathbf{v}\left(\boldsymbol{\beta}^{[t]}\right)$$

be a pseudo-data vector evaluated at iteration $[t]$. Then the system (4.81) can be written as

$$\left[\mathbf{X}'\mathbf{D}\left(\boldsymbol{\beta}^{[t]}\right)\mathbf{X}\right]\boldsymbol{\beta}^{[t+1]} = \mathbf{X}'\mathbf{D}\left(\boldsymbol{\beta}^{[t]}\right)\mathbf{y}^*\left(\boldsymbol{\beta}^{[t]}\right). \tag{4.82}$$

This is an iterative reweighted least-squares system where the matrix of weights

$$\mathbf{D}\left(\boldsymbol{\beta}^{[t]}\right) = \left\{p_i\left(\boldsymbol{\beta}^{[t]}\right)\left[1-p_i\left(\boldsymbol{\beta}^{[t]}\right)\right]\right\}$$

Calf	Birth weight	Score	Calf	Birth weight	Score
1	40	1	7	47	0
2	40	1	8	47	0
3	40	0	9	47	1
4	43	0	10	50	0
5	43	1	11	50	0
6	43	0	12	50	0

TABLE 4.3. Hypothetical data on birth weight and calving scores taken on 12 calves.

is the reciprocal of the variance of the logit

$$\ln \frac{p_i(\boldsymbol{\beta})}{1-p_i(\boldsymbol{\beta})}$$

evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}$; this was shown in Example 4.5. The Newton–Raphson algorithm is iterated until the change in successive rounds is negligible. If convergence is to a global maximum $\widehat{\boldsymbol{\beta}}$, then this is the ML estimate. The asymptotic variance covariance matrix is estimated as

$$\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)=\left[\mathbf{X}'\mathbf{D}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{X}\right]^{-1}. \tag{4.83}$$

Example 4.10 *Birth weight and calving difficulty*

Suppose that each of 12 genetically unrelated cows of the same breed gives birth to a calf. These are weighed at birth, and a score is assigned to indicate if there were birth difficulties. The scoring system is: 1 if calving is normal and 0 otherwise. The objective of the analysis is to assess if there is a relationship between birth weight and birth difficulty. A logit model is used where the underlying variate is expressed as

$$l_i=\alpha+\beta x_i+e_i, \quad i=1,2,...,12,$$

and where x_i is the birth weight in kilograms. The hypothetical data are in Table 4.3.

Using (4.76) the log-likelihood is

$$\begin{aligned} l(\alpha,\beta|\mathbf{y}) \propto\ & (\alpha+\beta\,40)-3\ln\left[1+\exp(\alpha+\beta\,40)\right] \\ & +2(\alpha+\beta\,43)-3\ln\left[1+\exp(\alpha+\beta\,43)\right] \\ & +2(\alpha+\beta\,47)-3\ln\left[1+\exp(\alpha+\beta\,47)\right] \\ & +3(\alpha+\beta\,50)-3\ln\left[1+\exp(\alpha+\beta\,50)\right]. \end{aligned}$$

The Newton–Raphson algorithm gives $\widehat{\alpha}=-12.81$ and $\widehat{\beta}=0.305$. The estimated asymptotic variance–covariance matrix is

$$Var\begin{bmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{bmatrix}=\begin{bmatrix}82.476 & -1.881\\ -1.881 & .043\end{bmatrix}.$$

The asymptotic standard error of $\widehat{\beta}$ is $\sqrt{.043} = .2074$. A confidence region of size 95% is approximately $(-.101, .711)$. The interval includes the value 0, so this data cannot refute the hypothesis that birth weight does not affect the probability of a difficult calving. Because α is a nuisance parameter, and it is not obvious how a marginal likelihood can be constructed for β, the profile likelihood $L\left(\beta, \widehat{\widehat{\alpha}}(\beta)\right)$ is calculated where $\widehat{\widehat{\alpha}}(\beta)$ is the partial ML estimator at a fixed value of β. Numerically, this is done by making a grid of values of β, finding the partial ML estimator of α corresponding to each value of β, and then computing the value of the resulting log-likelihood. Values of $\widehat{\widehat{\alpha}}(\beta)$ and of the profile log-likelihood at selected values of β are shown in Table 4.4. The value $\beta = .305$, the ML estimate, is the maximizer of the profile likelihood, as discussed in Section 4.4.3. ■

4.6.3 Mixed Effects Linear Logistic Model

Assume that the underlying variable is modelled now as:

$$l_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a} + e_i, \quad i = 1, 2, ..., N, \tag{4.84}$$

where $\mathbf{a} \sim [\mathbf{0}, \mathbf{G}(\boldsymbol{\phi})]$ is a vector of random effects having some distribution (typically, multivariate normality is assumed in quantitative genetics) with covariance matrix $\mathbf{G}(\boldsymbol{\phi})$. In turn, this covariance matrix may depend on unknown parameters $\boldsymbol{\phi}$, which may be variance and covariance components, for example, for traits that are subject to maternal genetic influences (Willham, 1963). The vector $\mathbf{z}_i'$ is a row incidence vector that plays the same role as $\mathbf{x}_i'$. The residual e_i has a logistic distribution, as before. The probability of survival for the ith individual, given $\boldsymbol{\beta}$ and $\mathbf{a}$, after setting the threshold to 0, is now

$$\begin{aligned}\Pr(Y_i = 1|\boldsymbol{\beta}, \mathbf{a}) = \Pr(l_i < t|\boldsymbol{\beta}, \mathbf{a}) &= 1 - \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}} p(e_i)\, de_i \\ &= \left[1 + \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}\right)\right]^{-1} = p_i(\boldsymbol{\beta}, \mathbf{a}),\end{aligned} \tag{4.85}$$

and the conditional probability of death is $\Pr(Y_i = 0|\boldsymbol{\beta}, \mathbf{a}) = 1 - p_i(\boldsymbol{\beta}, \mathbf{a})$. Under Bernoulli sampling the conditional probability of observing the data

β	$\widehat{\widehat{\alpha}}(\beta)$	$L\left(\beta, \widehat{\widehat{\alpha}}(\beta)\right)$
.100	−3.78275	−6.82957
.200	−8.21124	−6.38209
.300	−12.5950	−6.24025
.305	−12.8135	−6.23996
.400	−16.9390	−6.33550
.500	−21.2498	−6.60396

TABLE 4.4. Profile likelihood for β.

obtained is then:

$$p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{a}\right)=\prod_{i=1}^{N}\left[p_i\left(\boldsymbol{\beta},\mathbf{a}\right)\right]^{y_i}\left[1-p_i\left(\boldsymbol{\beta},\mathbf{a}\right)\right]^{1-y_i}.$$

In order to form the likelihood function, the marginal probability distribution of the observations is needed. Thus, the joint distribution $[\mathbf{y},\mathbf{a}|\boldsymbol{\beta},\boldsymbol{\phi}]$ must be integrated over $\mathbf{a}$ to obtain

$$p\left(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\phi}\right)=\int\prod_{i=1}^{N}\left[p_i\left(\boldsymbol{\beta},\mathbf{a}\right)\right]^{y_i}\left[1-p_i\left(\boldsymbol{\beta},\mathbf{a}\right)\right]^{1-y_i}p\left(\mathbf{a}|\boldsymbol{\phi}\right)d\mathbf{a}, \tag{4.86}$$

where $p\left(\mathbf{a}|\boldsymbol{\phi}\right)$ is the density of the joint distribution of the random effects. This integral cannot be expressed in closed form, and must be evaluated by numerical procedures, such as Gaussian quadrature. As shown below, this is feasible only when the random effects are independent (which is seldom the case in genetic applications), because then the problem reduces to one of evaluating unidimensional integrals. An alternative is to use Monte Carlo integration procedures, such as the Gibbs sampler. This will be discussed in subsequent chapters.

Let the model for liability be

$$l_{ij}=\mathbf{x}_{ij}'\boldsymbol{\beta}+a_i+e_{ij},\quad i=1,2,...,S,\ j=1,2,...,n_i.$$

Thus, there are S random effects and n_i observations are associated with random effect a_i. Suppose the random effects are i.i.d. with distribution $a_i\sim N\left(0,\boldsymbol{\phi}\right)$. Then the joint density of the vector of random effects is

$$p\left(\mathbf{a}|\boldsymbol{\phi}\right)=\prod_{i=1}^{S}p\left(a_i|\boldsymbol{\phi}\right).$$

Using this, (4.86) can be written as

$$p\left(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\phi}\right)=\prod_{i=1}^{S}\int\prod_{j=1}^{n_i}\left[p_{ij}\left(\boldsymbol{\beta},a_i\right)\right]^{y_{ij}}\left[1-p_{ij}\left(\boldsymbol{\beta},a_i\right)\right]^{1-y_{ij}}p\left(a_i|\boldsymbol{\phi}\right)da_i \tag{4.87}$$

indicating that in (4.87), S single dimension integrals need to be evaluated, instead of a multivariate integral of order S, as pointed out in connection with (4.86). However, the integral is not explicit. This illustrates that there are computational challenges in connection with ML estimation. The same is true of Bayesian methods, which are introduced in Chapter 5.

This page intentionally left blank

5

An Introduction to Bayesian Inference

5.1 Introduction

The potential appeal of Bayesian techniques for statistical analysis of genetic data will be motivated using examples from the field of animal breeding. Here, two types of data are encountered most often. First, there are observations obtained from animal production and disease recording programs; for example, birth and weaning weights in beef cattle breeding schemes and udder disease data in dairy cattle breeding. These are called "field" records, which are collected directly on the farms where animals are located. Second, there are data from genetic selection experiments conducted under fairly controlled conditions. For example, there may be lines of mice selected for increased ovulation rate, and lines in which there is random selection, serving as "controls". Field records are usually available in massive amounts, whereas experimental information is often limited.

Suppose milk yield field records have been taken on a sample of cows, and that one wishes to infer the amount of additive genetic variance in the population, perhaps using a mixed effects linear model. Often, use may be made of a large part of an entire national data base. In this situation, the corresponding full, marginal, or profile likelihood functions are expected to be sharp, leading to fairly precise estimates of additive genetic variance. Naturally, this is highly dependent on the model adopted for analysis. That is, if an elaborate multivariate structure with a large number of nuisance parameters is fitted, it is not necessarily true that all profile or marginal likelihoods will be sharp. However, in many instances, the amount

of information is so large that genetic parameters are well-estimated, with asymptotic theory working handsomely. Here, it is reasonable to expect that inferences drawn from alternative statistical methods will seldom lead to qualitatively different conclusions.

Consider now inferences drawn using data from designed experiments, where estimates tend to be far from precise. A common outcome is that experimental results are ambiguous, and probably consistent with several alternative explanations of the state of nature. This is because the scarce amount of resources available dictates a small effective population size of the experimental lines. In this situation, there can be much uncertainty about the parameters to be inferred remaining after analysis. Here, one would need to adopt methods capable of conveying accurately the limited precision of the estimates obtained. Arguably, inference assuming samples of an infinite size (asymptotic theory) should not be expected to be satisfactory.

How would many quantitative geneticists address the following question:

> "How much genetic change has taken place in the course of selection?"

Suppose that the assumption of joint normality of additive genetic effects and of phenotypic values is tenable. First, they would probably attempt to estimate components of variance (or covariance) in the base population using either a full or a marginal likelihood; in the latter case, this leads to REML estimates of the parameters. Second, conditionally on these estimates, they would proceed to obtain best linear unbiased predictions (BLUP) of the additive genetic effects (treated as random) (Henderson, 1973). Theoretically, BLUP is the linear combination of the observations that minimizes prediction error variance in the class of linear functions whose expected value is equal to the expected value of the predictand (the genetic effects). BLUP can be derived only if the dispersion parameters (variance and covariance components) are known, at least to proportionality. However, when the latter parameters are estimated from the data at hand (by REML, say), the resulting empirical BLUP is no longer linear or best, and remains unbiased only under certain conditions, assuming random sampling and absence of selection or of assortative mating (Kackar and Harville, 1981). Quantitative geneticists often ignore this problem, and proceed to predict genetic means (these being random over conceptual repeated sampling) for different generations or cohorts using the empirical BLUP. From the estimated means, measures of genetic change can be obtained. It must be noted that if dispersion parameters are known and random sampling holds, the estimated means have a jointly Gaussian distribution, with known mean vector and variance–covariance matrix. However, one could argue that if genetic variances were known, there would not be any need for conducting the experiment! Further, if the REML estimates are

used in lieu of the true values, the BLUP analysis ignores their error of estimation. Here, one could invoke asymptotic theory and hope that the data are informative enough, such that reality can be approximated well using limit arguments. The resulting "BLUP" of a generation mean has an unknown distribution, so how does one construct tests of the hypothesis "selection is effective" in such a situation? Also, complications caused by nonrandom sampling mechanisms are encountered, as parents of a subsequent generation are not chosen at random. Unless selection is ignorable, in some sense, inferences are liable to be distorted. In short, this is an example of a situation where the animal breeding paradigms for parameter estimation (maximum likelihood, asymptotic theory) and for prediction of random variables (BLUP), used together, are incapable of providing a completely satisfactory answer to one of the most important questions that can be posed in applied quantitative genetics. Should one use the conditional distribution of the random effects, given the data, ignoring selection and the error of estimation of the parameters for inferring genetic change?

If one wishes to know what to expect, at least in the frequentist sense of hypothetical repeated sampling, the only answer would seem to reside in simulating all conceivable selection schemes and designs, for all possible values of the parameters. Clearly, this is not feasible. Simulations would need to be sensibly restricted to experimental settings and to parameter values that are likely to reflect reality. This implies that at least something must be known about the state of nature, before experimentation. However, there is always uncertainty, ranging from mild to large.

An alternative is to adopt the Bayesian point of view. Under this setting, all unknown quantities in a statistical system are treated as random variables, reflecting (typically) subjective uncertainty measured by a probability distribution. The unknowns may include parameters (e.g., heritability or the inbreeding coefficient), random effects (e.g., the additive genetic value of an artificial insemination bull), data that are yet to be observed (e.g., the mean of the offspring of a pair of parents that will be measured under certain conditions), the sampling distribution adopted for the data generating process (e.g., given the parameters, the observations may have either a Gaussian or a multivariate-t distribution), or the entire model itself, engulfing all assumptions made. Here, there may be a number of competing probability models, of varying dimension. In the Bayesian approach, the idea is to combine what is known about the statistical ensemble before the data are observed (this knowledge is represented in terms of a prior probability distribution) with the information coming from the data, to obtain a posterior distribution, from which inferences are made using the standard probability calculus techniques presented in Chapters 1 and 2.

The inferences to be drawn depend on the question posed. Sometimes, one may seek a marginal posterior distribution, whereas in other instances, joint or conditional posterior distributions of subsets of variables may be targets in the analysis. Since any unknown quantity, irrespective of whether

it is a model, a parameter, or a future data point, is treated symmetrically, the answer is always found in the same manner, that is, by arriving at the corresponding posterior distribution via probability theory. The results of a Bayesian analysis can be presented by displaying the entire posterior distribution (or density function), or just some posterior summaries, such as the mean, median, variance, or percentiles. The results are interpreted probabilistically. For example, if one wishes to infer the mean (μ) of a distribution, one would say "the posterior probability that μ is between a and b is so much". This illustrates how different the Bayesian construct is from the frequentist-likelihood paradigms.

An overview of the basic elements of the Bayesian approach to inference will be presented in this chapter. The treatment begins with a description of Bayes theorem and of its consequences. Subsequently, joint, marginal, and conditional posterior distributions are introduced, including a presentation of the Bayesian manner of handling nuisance parameters. For the sake of clarity, the Bayesian probability models considered to illustrate developments emphasize linear specifications for Gaussian observations; linear models are well-known by quantitative geneticists. The presentation is introductory, and a much deeper coverage can be found, for example, in Zellner (1971), Box and Tiao (1973), Lee (1989), Bernardo and Smith (1994), O'Hagan (1994), Gelman et al. (1995), and Leonard and Hsu (1999). Additional topics in Bayesian analysis are discussed in Chapters 6, 7, and 8.

5.2 Bayes Theorem: Discrete Case

Suppose a scientist has M disjoint hypotheses $(H_1, H_2, \ldots, H_M)$ about some mechanism, these being mutually exclusive and, at least temporarily, exhaustive. The latter is an important consideration because at any point in time, one cannot formulate all possible hypotheses. Rather, the set $H_1, H_2, \ldots, H_M$ constitutes the collection of all hypotheses that can be formulated by this scientist now, in the light of existing knowledge (Malécot, 1947). Additional, competing, hypotheses would surely emerge, as more knowledge is acquired. Obviously, the "true" hypothesis cannot be observed, but there may be some inclination by the scientist toward accepting one in the set as being more likely than the others. In other words, there may be more certainty that one such hypothesis is true, relative to the alternatives. In Bayesian analysis, this uncertainty is expressed in terms of probabilities. In such a context, it is reasonable to speak of a random variable H taking one of M mutually exclusive and exhaustive states.

Let $p(H_i)$ be the prior probability assigned by this scientist to the event "hypothesis H_i is true", with

$$0 \le p(H_i) \le 1, \quad i = 1, 2, ..., M,$$
$$p(H_i \cap H_j) = 0, \quad i \ne j,$$
$$\sum_{i=1}^{M} p(H_i) = 1.$$

Then $p(H_i)$ $(i = 1, 2, \ldots, M)$ gives the prior probability distribution of the competing hypotheses. The prior distribution may be elicited either on subjective grounds, on mechanistic considerations, on evidence available so far, or using a combination of these three approaches to assess beliefs.

The Bayesian approach provides a description of how existing knowledge is modified by experience. Now let there be N observable effects $E_1, E_2, \ldots, E_N$. Given that hypothesis H_i holds, one expects to observe effects with conditional probabilities

$$0 \le p(E_j|H_i) \le 1, \quad i = 1, 2, \ldots, M, \ j = 1, 2, \ldots, N,$$
$$p(E_j \cap E_{j'}|H_i) = 0, \quad j \ne j',$$
$$\sum_{j=1}^{N} p(E_j|H_i) = 1.$$

Thus, $p(E_j|H_i)$ gives the conditional probability of the effects observed under hypothesis H_i, with these effects being disjoint, that is, E_j and $E_{j'}$ cannot be observed simultaneously. For example, under a one-locus Mendelian model (the hypothesis), an individual cannot be homozygote and heterozygote at the same time. This conditional distribution represents the probabilities of effects to be observed if experimentation proceeded under the conditions imposed by the hypothesis. Again, under Mendelian inheritance, random mating, no migration or mutation and a large population, one would expect the probabilities of observing AA, Aa, and aa individuals to be those resulting from the Hardy–Weinberg equilibrium.

Then let E be a random variable taking one of the states E_j $(j = 1, 2, \ldots, N)$, and let H and E have the joint distribution

$$p(H = H_i, E = E_j) = p(E_j|H_i)p(H_i).$$

The conditional probability that hypothesis H_i holds, given that effects E_j are observed, is then

$$p(H_i|E_j) = \frac{p(H = H_i, E = E_j)}{p(E_j)} = \frac{p(E_j|H_i)p(H_i)}{p(E_j)}, \tag{5.1}$$

where $p(E_j)$ is the marginal or total probability of observing effect E_j, that is, the probability of observing E_j over all possible hypotheses

$$p(E_j) = \sum_{i=1}^{M} p(E_j|H_i)p(H_i) = E_H\left[p(E_j|H_i)\right], \tag{5.2}$$

where $E_H(\cdot)$ indicates an expectation taken with respect to the prior distribution of the hypotheses. Using (5.2) in (5.1)

$$\begin{aligned} p(H_i|E_j) &= \frac{p(E_j|H_i)p(H_i)}{E_H\left[p(E_j|H_i)\right]} &(5.3)\\ &\propto p(E_j|H_i)p(H_i). &(5.4)\end{aligned}$$

This standard result of conditional probability is also known as Bayes theorem when applied to the specific problem of inferring "causes" from "effects", and it is also called "inverse probability". It states that the probability of a cause or hypothesis H_i, given evidence E_j, is proportional to the product of the prior probability assigned to the hypothesis, $p(H_i)$, times the conditional probability of observing the effect E_j under hypothesis H_i. The distribution in (5.3) or (5.4) is called the posterior probability distribution, with the denominator in (5.3) acting as normalizing constant.

Expression (5.4) illustrates a concept called "Bayesian learning". This is the process by which a prior opinion (with the associated uncertainty stated by the prior distribution) is modified by evidence E (generated with uncertainty under a sampling model characterized by a distribution with probabilities $p(E_j|H_i)$), to become a posterior opinion, this having a distribution with probabilities $p(H_i|E_j)$. Suppose now that additional evidence $E'_{j'}$ accrues. Using Bayes theorem in (5.3) or (5.4):

$$\begin{aligned} p(H_i|E'_{j'}, E_j) &= \frac{p(E'_{j'}, E_j|H_i)p(H_i)}{E_H[p(E'_{j'}, E_j|H_i)]}\\ &\propto p(E'_{j'}, E_j|H_i)p(H_i)\\ &\propto p(E'_{j'}|E_j, H_i)p(E_j|H_i)p(H_i)\\ &\propto p(E'_{j'}|E_j, H_i)p(H_i|E_j). &(5.5)\end{aligned}$$

The preceding indicates that the posterior distribution after evidence E_j conveys the prior opinion before $E'_{j'}$ is observed. It also describes how opinions are revised sequentially or, equivalently, how knowledge is modified by evidence. If, given H_i, $E'_{j'}$ is conditionally independent of E_j, then

$$p(E'_{j'}|E_j, H_i)p(E_j|H_i) = p(E'_{j'}|H_i)p(E_j|H_i).$$

More generally, for S pieces of evidence, assuming conditional independence,

$$p(H_i|E^S_{j^S}, E^{S-1}_{j^{S-1}}, ..., E^2_{j^2}, E^1_{j^1}) \propto \prod_{k=1}^{S} p(E^k_{j^k}|H_i)p(H_i), \tag{5.6}$$

where $E_{j^k}^k$ denotes the evidence in datum k, with $j = 1, 2, ..., N$ indicating the different values that E can take at any step k of the process of accumulating information. Letting $\mathbf{E}$ be the entire evidence, (5.6) can be written as

$$\begin{aligned} p(H_i|\mathbf{E}) &\propto \exp\left\{\sum_{k=1}^{S} \log\left[p(E_{j^k}^k|H_i)\right] + \log p(H_i)\right\} \\ &\propto \exp\left[S\overline{L} + \log p(H_i)\right] \\ &\propto \exp\left\{S\overline{L}\left[1 + \frac{\log p(H_i)}{S\ \overline{L}}\right]\right\}, \end{aligned} \tag{5.7}$$

where

$$\overline{L} = \frac{1}{S}\sum_{k=1}^{S} \log\left[p(E_{j^k}^k|H_i)\right]$$

is the average log-probability of observing $E_{j^k}^k$ under hypothesis H_i. Now, letting $S \to \infty$, and provided that $p(H_i) > 0$, it can be seen that the exponent in expression (5.7) tends toward $S\overline{L}$, indicating that the contribution of the prior to the posterior is of order $1/S$. This implies that the evidence tends to overwhelm the prior as more and more information accumulates so, for large S,

$$p(H_i|\mathbf{E}) \propto \prod_{k=1}^{S} p(E_{j^k}^k|H_i). \tag{5.8}$$

Following O'Hagan (1994), note from (5.1) that evidence E will increase the probability of a hypothesis H only if

$$p(E|H) > p(E),$$

where subscripts are ignored, for simplicity. Now, from (5.2) and denoting as $\overline{H}$ the event "H not true", with $p(\overline{H})$ being the corresponding probability, one can write

$$\begin{aligned} p(E) &= p(E|H)p(H) + p(E|\overline{H})p(\overline{H}) \\ &= p(E|H)\left[1 - p(\overline{H})\right] + p(E|\overline{H})p(\overline{H}). \end{aligned}$$

Rearranging

$$p(E|H) - p(E) = \left[p(E|H) - p(E|\overline{H})\right] p(\overline{H}).$$

This indicates that evidence E will increase the probability of a hypothesis if and only if $p(E|H) > p(E|\overline{H})$, that is, if the chances of observing E are larger under H than under any of the competing hypothesis. If this is the case, it is said that E confers a higher likelihood to H than to $\overline{H}$; thus, $p(E|H)$ is called the likelihood of H. This is exactly the concept

of likelihood function discussed in Chapter 3, that is, the probability (or density) of the observations viewed as a function of the parameters (the hypotheses play the role of the parameter values in this discussion). Thus, the maximum likelihood estimator is the function of the data conferring the highest likelihood to a particular value of the parameter.

The controversy in statistics about the use of Bayes theorem in science centers on that the prior distribution is often based on subjective, if not arbitrary (or convenient), elicitation. In response to this criticism, Savage (1972) wrote:

> "It has been countered, I believe, that if experience systematically leads people with opinions originally different to hold a common opinion, then that common opinion, and it only, is the proper subject of scientific probability theory. There are two inaccuracies in this argument. In the first place, the conclusion of the personalistic view is not that evidence brings holders of different opinions to the same opinions, but rather to similar opinions. In the second place, it is typically true of any observational program, however extensive but prescribed in advance, that there exist pairs of opinions, neither of which can be called extreme in any precisely defined sense, but which cannot be expected, either by their holders or any other person, to be brought into close agreement after the observational program."

Furthermore, Box (1980) argued as follows:

> "In the past, the need for probabilities expressing prior beliefs has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief...I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters."

The probability distribution of the hypotheses cannot be construed as a frequency distribution, that is, as a random process generated as if hypotheses were drawn at random, with replacement, from an urn. If this were the case, one could refute or corroborate the prior distribution by drawing a huge number of independent samples from the said urn. Technically, however, there would always be ambiguity, unless the number of samples is infinite! Instead, the prior probabilities in Bayesian inference must be interpreted as relative degrees of belief about the state of nature, before any experimentation or observation takes place.

The debate about the alternative methods of inference has a long history with eloquent arguments from both camps; we do not feel much can be added here. A balanced comparative overview of methods of inference

is presented by Barnett (1999). A partisan, in-depth presentation of the Bayesian approach is given in Bernardo and Smith (1994) and O'Hagan (1994). A philosophical discussion of the concept(s) of probability can be found in Popper (1972, 1982), who holds a strong, anti-inductive and anti-subjective position. Indeed, Popper (1982) is concerned with the influence that subjective probability has had on quantum mechanics. Popper cites Heisenberg (1958), who writes:

> "The conception of objective reality ... has thus evaporated ... into the transparent clarity of a mathematics that represents no longer the behavior of particles but rather our knowledge of this behavior".

For a detailed and focused philosophical critique of the Bayesian approach to inference, the reader can consult the books of Howson and Urbach (1989) and Earman (1992).

When a prior distribution is "universally agreed upon", or when it is based on mechanistic considerations, then Bayes theorem is accepted as a basis for inference without reservation. Two examples of the latter situation follow.

Example 5.1 *Incidence of a rare disease*

Consider a rare disease whose frequency in the population is 1 in $5,000$ (*i.e.*, 0.0002). A test for detecting the disease is available and it has a false positive rate of 0.05 and a false negative rate of 0.01. A person is taken at random from the population and the test gives a positive result. What is the probability that the person is diseased? Let θ represent a random variable taking the value 1 if the person is diseased and 0 otherwise. Let Y be a random variable that takes the value 1 if the test is positive, and 0 if the test is negative. The data are here represented by the single value $Y = 1$. The prior probability (before the test result is available) that a randomly chosen individual is diseased is

$$\Pr(\theta = 1) = 0.0002.$$

Based on the false positive and false negative rates, we can write

$$\begin{aligned} \Pr(Y = 1|\theta = 0) &= 0.05, \quad \Pr(Y = 0|\theta = 0) = 0.95, \\ \Pr(Y = 0|\theta = 1) &= 0.01, \quad \Pr(Y = 1|\theta = 1) = 0.99. \end{aligned}$$

Applying Bayes theorem, the posterior probability that the individual is diseased (after having observed $Y = 1$) is given by

$$\begin{aligned} &\Pr(\theta = 1|Y = 1) \\ &= \frac{\Pr(\theta = 1)\Pr(Y = 1|\theta = 1)}{\Pr(\theta = 1)\Pr(Y = 1|\theta = 1) + \Pr(\theta = 0)\Pr(Y = 1|\theta = 0)} \\ &= \frac{(0.0002)(0.99)}{(0.0002)(0.99) + (0.9998)(0.05)} \approx 0.0039. \end{aligned}$$

Thus, a positive test, has raised the probability that the individual has the disease from 0.0002 (the a priori probability before the test result is available) to 0.0039 (the posterior probability after observing $Y = 1$). ■

Example 5.2 *Inheritance of hemophilia*

The following is adapted from Gelman et al. (1995). Hemophilia is a genetic disease in humans. The locus responsible for its expression is located in the sex chromosomes (these are denoted as XX in women, and XY in men). The condition is observed in women only in double recessive individuals (aa), and in men that are carriers of the a allele in the X-chromosome. Suppose there is a nonhemophiliac woman whose father and mother are not affected by the disease, but her brother is known to be hemophiliac. This implies that her nonhemophiliac mother must be heterozygote, a carrier of a. What is the probability that the propositus woman is also a carrier of the gene? Let θ be a random variable taking one of two mutually exclusive and exhaustive values (playing the role of the hypotheses in the preceding section). Either $\theta = 1$ if the woman is a carrier, or $\theta = 0$ otherwise. Since it is known that the mother of the woman must be a carrier (this constitutes part of the system within which probabilities are assigned), the prior distribution of θ is

$$\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}.$$

In the absence of additional information, it is not possible to make a very sharp probability assignment. Suppose now that the woman has two sons, none of which is affected. Let Y_i be a binary random variable taking the value 0 if son i is not affected, or 1 if he has the disease; thus, the values of Y_1 and Y_2 constitute the evidence E. Given that $\theta = 1$, the probability of the observed data is

$$\begin{aligned} & \Pr(Y_1 = 0, Y_2 = 0 | \theta = 1) \\ = \; & \Pr(Y_1 = 0 | \theta = 1) \Pr(Y_2 = 0 | \theta = 1) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}. \end{aligned}$$

This follows because:
a) the observations are assumed to be independent, conditionally on θ and b) if the woman is a carrier ($\theta = 1$), there is a 50% probability that she will not transmit the allele.
On the other hand, if she is not a carrier ($\theta = 0$):

$$\begin{aligned} & \Pr(Y_1 = 0, Y_2 = 0 | \theta = 0) \\ = \; & \Pr(Y_1 = 0 | \theta = 0) \Pr(Y_2 = 0 | \theta = 0) = 1 \times 1 = 1, \end{aligned}$$

this being so because it is impossible for a son to have the disease unless the mother is a carrier (ignoring mutation). Hence, the data confer four times

more likelihood to the hypothesis that the mother is not a carrier. Using the information that none of the sons is diseased, the posterior distribution of θ is then

$$\begin{aligned}\Pr(\theta = 1|Y_1 = 0, Y_2 = 0) &= \frac{\Pr(\theta = 1)\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)}{\Pr(Y_1 = 0, Y_2 = 0)} \\ &= \frac{\Pr(\theta = 1)\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)}{\sum_{i=0}^{1} \Pr(\theta = i)\Pr(Y_1 = 0, Y_2 = 0|\theta = i)} \\ &= \frac{\frac{1}{2}\frac{1}{4}}{\frac{1}{2}1 + \frac{1}{2}\frac{1}{4}} = \frac{1}{5}\end{aligned}$$

and

$$\Pr(\theta = 0|Y_1 = 0, Y_2 = 0) = 1 - \frac{1}{5} = \frac{4}{5}.$$

A sharper probability assignment can be made now, and the combination of prior information with the evidence suggests that the mother is probably not a carrier. The latter possibility cannot be ruled out, however, as there is a 20% probability that the mother is heterozygote. The posterior odds in favor of the hypothesis that the mother is not a carrier is given by the ratio

$$\begin{aligned}\frac{\Pr(\theta = 0|Y_1 = 0, Y_2 = 0)}{\Pr(\theta = 1|Y_1 = 0, Y_2 = 0)} &= \frac{\Pr(Y_1 = 0, Y_2 = 0|\theta = 0)}{\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)} \frac{\Pr(\theta = 0)}{\Pr(\theta = 1)} \\ &= \frac{1}{\frac{1}{4}} \frac{\frac{1}{2}}{\frac{1}{2}} = 4,\end{aligned}$$

where the ratio

$$\frac{\Pr(\theta = 0)}{\Pr(\theta = 1)} = 1$$

is called the prior odds in favor of the hypothesis. Further,

$$B_{01} = \frac{\Pr(Y_1 = 0, Y_2 = 0|\theta = 0)}{\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)} = 4$$

is called the Bayes factor, that is, the factor by which the prior odds about the hypotheses are modified by the evidence and converted into posterior odds (a more thorough discussion of Bayes factors will be presented in Chapter 8). In this example, the odds in favor of the hypothesis that $\theta = 0$ relative to $\theta = 1$ increase by a factor of 4 after observing two sons that are not affected by the disease. Suppose that the woman suspected of being a carrier has n children. The posterior distribution of θ can be represented as

$$\Pr(\theta = i|\mathbf{y}) = \frac{\Pr(\theta = i)\Pr(\mathbf{y}|\theta = i)}{\Pr(\theta = i)\Pr(\mathbf{y}|\theta = i) + \Pr(\theta \neq i)\Pr(\mathbf{y}|\theta \neq i)}, \quad i = 0, 1,$$

where $\mathbf{y} = [Y_1, Y_2, ..., Y_n]'$. Partition the data as $\mathbf{y} = [\mathbf{y}'_A, \mathbf{y}'_B]'$, with $\mathbf{y}_A$ being the records on presence or absence of the disease for the first m progeny, and with $\mathbf{y}_B$ containing data on the last $n-m$ children. The posterior distribution is

$$\Pr(\theta = i|\mathbf{y}) = \frac{\Pr(\theta = i)\, p(\mathbf{y}_A|\theta = i)\, p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}{\sum_{i=0}^{1} \Pr(\theta = i)\, p(\mathbf{y}_A|\theta = i)\, p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}.$$

Dividing the numerator and denominator by the marginal probability of observing $\mathbf{y}_A$, that is, by $p(\mathbf{y}_A)$ gives

$$\Pr(\theta = i|\mathbf{y}) = \frac{\frac{\Pr(\theta = i)\, p(\mathbf{y}_A|\theta = i)}{p(\mathbf{y}_A)} p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}{\sum_{i=0}^{1} \frac{\Pr(\theta = i)\, p(\mathbf{y}_A|\theta = i)}{p(\mathbf{y}_A)} p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}.$$

Note, however, that

$$\frac{\Pr(\theta = i)\, p(\mathbf{y}_A|\theta = i)}{p(\mathbf{y}_A)} = \Pr(\theta = i|\mathbf{y}_A)$$

is the posterior probability after observing $\mathbf{y}_A$, which acts as a prior before observing $\mathbf{y}_B$. Then, it follows that

$$\Pr(\theta = i|\mathbf{y}) = \frac{\Pr(\theta = i|\mathbf{y}_A)\, p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}{\sum_{i=0}^{1} \Pr(\theta = i|\mathbf{y}_A)\, p(\mathbf{y}_B|\mathbf{y}_A, \theta = i)}$$

illustrating the "memory" property of Bayes theorem. If the observations are conditionally independent, as assumed in this example, then

$$p(\mathbf{y}_B|\mathbf{y}_A, \theta = i) = p(\mathbf{y}_B|\theta = i).$$

Suppose now that the woman has a third, unaffected, son. The prior distribution now assigns probabilities of $\frac{4}{5}$ and $\frac{1}{5}$ to the events "not being a carrier" and "carrying the allele", respectively. The posterior probability of the woman being a carrier, after observing a third, unaffected child, is

$$\begin{aligned} &\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, Y_3 = 0) \\ &= \frac{\frac{1}{5}\Pr(Y_3 = 0|\theta = 1)}{\frac{1}{5}\Pr(Y_3 = 0|\theta = 1) + \frac{4}{5}\Pr(Y_3 = 0|\theta = 0)} \\ &= \frac{\frac{1}{5}\frac{1}{2}}{\frac{1}{5}\frac{1}{2} + \frac{4}{5}1} = \frac{1}{9}. \end{aligned}$$

The same result is obtained starting from the prior before observing any children

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, Y_3 = 0) = \frac{\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3}{\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3 + \frac{1}{2}\cdot(1)^3}$$
$$= \frac{1}{9}.$$

■

5.3 Bayes Theorem: Continuous Case

In a somewhat narrower setting, consider now the situation where the role of the evidence $\mathbf{E}$ is played by a vector of observations $\mathbf{y}$, with the hypothesis H replaced by a vector of unknowns $\boldsymbol{\theta}$. The latter will be generally referred to as the "parameter" vector, although $\boldsymbol{\theta}$ can include random effects (in the usual frequentist sense), missing data, censored observations, etc. It will be assumed that $\boldsymbol{\theta}$ and $\mathbf{y}$ are continuous valued, and that a certain probability model M posits the joint distribution $[\boldsymbol{\theta}, \mathbf{y}|M]$. For example, M may postulate that this distribution is jointly Gaussian, whereas model M' supposes a multivariate-t distribution. Alternatively, M and M' could represent two alternative explanatory structures in a regression model. At this point it will be assumed that there is complete certainty about model M holding, although this may not be so, in which case one encounters the important problem of Bayesian model selection. We will return to this later on but now, with the understanding that developments are conditional on model M, the dependency on the model will be abandoned in the notation.

The joint density of $\boldsymbol{\theta}$ and $\mathbf{y}$ can be written as

$$h(\boldsymbol{\theta}, \mathbf{y}) = g(\boldsymbol{\theta})\, f(\mathbf{y}|\boldsymbol{\theta}) = m(\mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}), \tag{5.9}$$

where

- $g(\boldsymbol{\theta})$ is the marginal density of $\boldsymbol{\theta}$:

$$g(\boldsymbol{\theta}) = \int h(\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{y} = \int p(\boldsymbol{\theta}|\mathbf{y})\, m(\mathbf{y})\, d\mathbf{y} = E_{\mathbf{y}}\left[p(\boldsymbol{\theta}|\mathbf{y})\right],$$

  The corresponding distribution describes the plausibility of values that $\boldsymbol{\theta}$ takes in a parameter space $\boldsymbol{\Theta}$, unconditionally on the observations $\mathbf{y}$. This is the density of the prior distribution of $\boldsymbol{\theta}$, which provides a summary of nonsample information about the parameters. Values of $\boldsymbol{\theta}$ outside of $\boldsymbol{\Theta}$ have null density. For example, a reasonable Bayesian model would assign null density to values of a genetic correlation outside of the $[-1, 1]$ boundaries, or to negative values of

a variance component. In the preceding, as well as in all subsequent developments, it will be assumed that the required integrals exist, unless stated otherwise.

- $f(\mathbf{y}|\boldsymbol{\theta})$ is the density of values that $\mathbf{y}$ takes at a given, albeit unknown, value of $\boldsymbol{\theta}$. It represents the likelihood that evidence $\mathbf{y}$ confers to $\boldsymbol{\theta}$; the part of $f(\mathbf{y}|\boldsymbol{\theta})$ that varies with $\boldsymbol{\theta}$ is called the likelihood function or, simply, the likelihood of $\boldsymbol{\theta}$. The set of possible values that $\mathbf{y}$ can take is given by $\Re_{\mathbf{y}}$, the sampling space of $\mathbf{y}$. The integration above is implicitly over this sampling space. For example, if the data vector contains some truncated random variables, the sampling space would be given by the boundaries within which these variables are allowed to vary.

- $m(\mathbf{y})$ is the density of the marginal distribution of the observations. In the Bayesian context, this does not depend on $\boldsymbol{\theta}$, since

$$m(\mathbf{y}) = \int h(\boldsymbol{\theta}, \mathbf{y})\, d\boldsymbol{\theta} = \int f(\mathbf{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}\left[f(\mathbf{y}|\boldsymbol{\theta})\right],$$

  where the integration is over the sample space of $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$. This implies that $m(\mathbf{y})$ is the average, taken over the prior distribution of $\boldsymbol{\theta}$, of all possible likelihood values that the evidence $\mathbf{y}$ would confer to $\boldsymbol{\theta}$.

- $p(\boldsymbol{\theta}|\mathbf{y})$ is the density of the posterior distribution of $\boldsymbol{\theta}$, $[\boldsymbol{\theta}|\mathbf{y}]$, providing a summary of the information about $\boldsymbol{\theta}$ contained in both $\mathbf{y}$ and in the prior distribution $[\boldsymbol{\theta}]$. From (5.9), the posterior density can be written as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{g(\boldsymbol{\theta})\, f(\mathbf{y}|\boldsymbol{\theta})}{m(\mathbf{y})} \propto g(\boldsymbol{\theta})\, f(\mathbf{y}|\boldsymbol{\theta}) \tag{5.10}$$

  as one is interested in variation with respect to $\boldsymbol{\theta}$ only. Further, let $L(\boldsymbol{\theta}|\mathbf{y})$ be the part of $f(\mathbf{y}|\boldsymbol{\theta})$ varying with $\boldsymbol{\theta}$, or likelihood function. Thus, an alternative representation of the posterior density is

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto g(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{y}), \tag{5.11}$$

  or

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{g(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{y})}{\int g(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}}. \tag{5.12}$$

  A special case of Bayesian analysis is encountered when the prior density $g(\boldsymbol{\theta})$ is uniform over the entire parameter space; in other words, $g(\boldsymbol{\theta})$ is proportional to a constant. This is called a "flat prior". The constant cancels in the numerator and denominator of (5.12), and the posterior becomes

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})}{\int L(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}} \propto L(\boldsymbol{\theta}|\mathbf{y}).$$

Hence, the posterior is proportional solely to the likelihood, and it exists only if the integral of the likelihood over $\boldsymbol{\theta}$ is finite. Otherwise, the posterior density is said to be improper. Note that if the required integral converges, the integration constant of the posterior is the reciprocal of the integrated likelihood. Also, if the posterior distribution exists, the mode of the posterior density is identical to the ML estimator. This suggests that this estimator is a feature of any (existing) posterior distribution constructed from an initial position where the observer is indifferent to any of the possible values of $\boldsymbol{\theta}$, when this is represented by the uniform distribution.

Example 5.3 *Inferring additive genetic effects*
Let an observation for a quantitative trait be y. Suppose that a reasonable model for representing a phenotypic value is

$$y = \mu + a + e,$$

where μ is a known constant, a is the additive genetic effect of the individual, and e is an environmental deviation. Let $a \sim N(0, v_a)$ and $e \sim N(0, v_e)$ be independently distributed, where v_a and v_e are the additive genetic and environmental variances, respectively, both being assumed known here. Thus, y must also be normal. It follows that the conditional distribution of y given a is the normal process

$$[y|\mu, a, v_a, v_e] \sim N\left(\mu + a, v_e\right).$$

Suppose the problem is inferring a from y. Here, $\theta = a$. The prior density of a is normal, with the parameters given above. The posterior distribution of interest is then:

$$[a|y, \mu, v_a, v_e]$$

which is identical to the conditional distribution of a given y in a frequentist sense. If a and y are jointly normal, as is the case here, it follows directly that the posterior distribution must be normal as well, with mean

$$\begin{aligned} E\left(a|y, \mu, v_a, v_e\right) &= E\left(a\right) + Cov(a, y)\, Var^{-1}\left(y\right)\left(y - \mu\right) \\ &= \frac{v_a}{v_a + v_e}\left(y - \mu\right) = h^2\left(y - \mu\right), \end{aligned}$$

where h^2 is the heritability of the trait. The posterior variance is

$$Var\left(a|y, \mu, v_a, v_e\right) = v_a - \frac{v_a^2}{v_a + v_e} = v_a\left(1 - h^2\right)$$

so there will always be some uncertainty about the genetic value, given the phenotypic value, unless the genotype has complete penetrance, that is,

when there is no environmental variance. In view of the mean and variance of the posterior distribution, the corresponding density is

$$p\left(a|y,\mu,v_a,v_e\right)=\frac{1}{\sqrt{2\pi v_a\left(1-h^2\right)}}\exp\left\{-\frac{\left[a-h^2\left(y-\mu\right)\right]^2}{2v_a\left(1-h^2\right)}\right\}.$$

Note that the prior opinion has been modified by the evidence provided by observation of the phenotypic value. For example, in the prior distribution the modal value of a is 0, whereas it is $h^2\left(y-\mu\right)$, a posteriori. Further, in the prior distribution, the probability that $a>0$ is 1/2. In the posterior distribution, this value is

$$\begin{aligned}\Pr\left(a>0|y,\mu,v_a,v_e\right)&=1-\Pr\left(a\leq 0|y,\mu,v_a,v_e\right)\\&=1-\Phi\left[\frac{-h^2\left(y-\mu\right)}{\sqrt{v_a\left(1-h^2\right)}}\right]\\&=\Phi\left[\frac{h^2\left(y-\mu\right)}{\sqrt{v_a\left(1-h^2\right)}}\right].\end{aligned}$$

The marginal density of the observation is

$$p\left(y|\mu,v_a,v_e\right)=\frac{1}{\sqrt{2\pi\left(v_a+v_e\right)}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\left(v_a+v_e\right)}\right].$$

This follows because the model states that the phenotypic value is a linear combination of two normally distributed random variables. The expression is identical to the marginal density of the observations in a frequentist setting, because our Bayesian model does not postulate uncertainty about μ, v_a and v_e. Otherwise, the marginal density given above would need to be deconditioned over the prior distribution of μ, v_a, and v_e. ■

Example 5.4 *Inferring additive genetic effects from repeated measures* Let the setting be as in the preceding example. Suppose now that n independent measurements are taken on the individual, and that μ is an unknown quantity, with prior distribution $\mu\sim N\left(\mu_0,v_0\right)$, where the hyperparameters (parameters of the prior distribution) are known. The model for the ith measure is then

$$y_i=\mu+a+e_i. \tag{5.13}$$

If μ and a are assumed to be independent, a priori, the joint posterior density is

$$\begin{aligned} & p\left(\mu, a|y_1, y_2, ..., y_n, \mu_0, v_0, v_a, v_e\right) \\ & \propto \prod_{i=1}^{n} p(y_i|\mu, a, v_e) p\left(a|v_a\right) p\left(\mu|\mu_0, v_0\right) \\ \propto & \prod_{i=1}^{n} \exp\left[-\frac{\left(y_i-\mu-a\right)^2}{2v_e}\right] \exp\left[-\frac{a^2}{2v_a}\right] \exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2v_0}\right] \\ \propto & \exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\mu-a\right)^2}{2v_e}-\frac{a^2}{2v_a}\right] \exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2v_0}\right]. \end{aligned} \tag{5.14}$$

Now

$$\begin{aligned} \frac{\sum_{i=1}^{n}\left(y_i-\mu-a\right)^2}{v_e}+\frac{a^2}{v_a} &= \frac{\sum_{i=1}^{n}\left[a+\mu-y_i\right]^2}{v_e}+\frac{a^2}{v_a} \\ &= \frac{\sum_{i=1}^{n}\left[a-\left(\overline{y}-\mu\right)-\left(y_i-\overline{y}\right)\right]^2}{v_e}+\frac{a^2}{v_a} \\ &= \frac{n\left[a-\left(\overline{y}-\mu\right)\right]^2+\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{v_e}+\frac{a^2}{v_a}, \end{aligned}$$

where $\overline{y}$ is the average of the n records; recall that $\sum_{i=1}^{n}\left(y_i-\overline{y}\right)=0$. Using this in the joint posterior (5.14):

$$\begin{aligned} & p\left(\mu, a|\mathbf{y}, \mu_0, v_0, v_a, v_e\right) \\ \propto & \exp\left[-\frac{1}{2}\left\{\frac{n\left[a-\left(\overline{y}-\mu\right)\right]^2}{v_e}+\frac{a^2}{v_a}\right\}\right] \exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2v_0}\right] \end{aligned} \tag{5.15}$$

as expressions not involving μ or a get absorbed in the integration constant. Now, there is an identity for combining quadratic forms (Box and Tiao, 1973) stating that

$$M(z-m)^2+B(z-b)^2=(M+B)(z-c)^2+\frac{MB}{M+B}\left(m-b\right)^2 \tag{5.16}$$

with

$$c=\left(M+B\right)^{-1}\left(Mm+Bb\right). \tag{5.17}$$

Now put

$$\begin{aligned} m &= \overline{y} - \mu, \\ b &= 0, \\ M &= \frac{n}{v_e}, \\ B &= \frac{1}{v_a}, \\ z &= a. \end{aligned}$$

Hence, by analogy,

$$\frac{n\left[a-(\overline{y}-\mu)\right]^2}{v_e}+\frac{a^2}{v_a}=\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2 + \frac{\frac{n}{v_e}\frac{1}{v_a}}{\frac{n}{v_e}+\frac{1}{v_a}}(\overline{y}-\mu)^2.$$

Employing this decomposition, the joint posterior density in (5.15) is expressible as

$$\begin{aligned} &p\left(\mu, a|\mathbf{y}, \mu_0, v_0, v_a, v_e\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2\right] \\ &\quad \times \exp\left[-\frac{n(\overline{y}-\mu)^2}{2\left(nv_a+v_e\right)}\right]\exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2v_0}\right]. \end{aligned} \tag{5.18}$$

From this, one can proceed to derive a series of distributions of interest, as given below.

Conditional (given μ) posterior density of the additive genetic effect. This is obtained by retaining the part of the joint posterior that varies with a:

$$\begin{aligned} &p\left(a|\mu, \mathbf{y}, \mu_0, v_0, v_a, v_e\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2\right] \end{aligned} \tag{5.19}$$

This is clearly the density of a normal distribution with mean

$$\begin{aligned} E\left(a|\mu, \mathbf{y}, \mu_0, v_0, v_a, v_e\right) &= \left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right] \\ &= \frac{v_a}{v_a+\frac{v_e}{n}}(\overline{y}-\mu) \\ &= \frac{n}{n+\frac{1-h^2}{h^2}}(\overline{y}-\mu) \end{aligned}$$

and variance

$$Var\left(a|\mu,\mathbf{y},\mu_0,v_0,v_a,v_e\right)=\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}=v_e\left(n+\frac{1-h^2}{h^2}\right)^{-1}.$$

Note that this posterior distribution depends on the data only through $\overline{y}$.

Conditional (given a) posterior density of μ. This is arrived at in a similar manner, that is, by retaining in (5.14) only the terms varying with μ. One obtains

$$p\left(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e\right)\propto\exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\mu-a\right)^2}{2v_e}\right]\exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2v_0}\right].$$

Here, using some of the previous results,

$$\begin{aligned}&\frac{\sum_{i=1}^{n}\left(y_i-\mu-a\right)^2}{v_e}+\frac{\left(\mu-\mu_0\right)^2}{v_0}\\&=\frac{n\left[\mu-\left(\overline{y}-a\right)\right]^2+\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{v_e}+\frac{\left(\mu-\mu_0\right)^2}{v_0}.\end{aligned}$$

There are now two quadratic forms on μ that can be combined, employing (5.16) and (5.17), as

$$\begin{aligned}&\frac{n\left[\mu-\left(\overline{y}-a\right)\right]^2}{v_e}+\frac{\left(\mu-\mu_0\right)^2}{v_0}\\&=\left(\frac{n}{v_e}+\frac{1}{v_0}\right)\left\{\mu-\left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-a)+\frac{\mu_0}{v_0}\right]\right\}^2\\&\quad+\frac{\frac{n}{v_e}\cdot\frac{1}{v_0}}{\frac{n}{v_e}+\frac{1}{v_0}}\left(\overline{y}-\mu_0-a\right)^2.\end{aligned}\tag{5.20}$$

Using this in the conditional density, and retaining only the terms that vary with μ

$$\begin{aligned}&p\left(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e\right)\\&\propto\exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_0}\right)\left\{\mu-\left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-a)+\frac{\mu_0}{v_0}\right]\right\}^2\right].\end{aligned}\tag{5.21}$$

It follows that the preceding density is that of a normal process with parameters

$$E(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e) = \left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-a)+\frac{\mu_0}{v_0}\right]$$
$$= \mu_0 + \frac{n}{n+\frac{v_e}{v_0}}(\overline{y}-\mu_0-a)$$

and

$$Var(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e) = \left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1} = v_e\left(n+\frac{v_e}{v_0}\right)^{-1}.$$

Marginal posterior density of μ. This is found by integrating the joint density (5.18) over a:

$$p(\mu|\mathbf{y},\mu_0,v_0,v_a,v_e)$$
$$= \int p(\mu,a|\mathbf{y},\mu_0,v_0,v_a,v_e)\,da$$
$$\propto \exp\left\{-\frac{1}{2}\left[\frac{n}{(nv_a+v_e)}(\overline{y}-\mu)^2+\frac{(\mu-\mu_0)^2}{v_0}\right]\right\}$$
$$\times\int \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2\right]da.$$

The integral is over a normal kernel, and evaluates to

$$\sqrt{2\pi\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}}.$$

Noting that the expression does not involve μ, it follows that

$$p(\mu|\mathbf{y},\mu_0,v_0,v_a,v_e) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{(nv_a+v_e)}(\overline{y}-\mu)^2+\frac{(\mu-\mu_0)^2}{v_0}\right]\right\}.$$

Using (5.16) and (5.17), the two quadratic forms on μ can be combined as

$$\frac{n}{(nv_a+v_e)}(\mu-\overline{y})^2+\frac{(\mu-\mu_0)^2}{v_0}$$
$$= \frac{1}{V_\mu}(\mu-\overline{\mu})^2+\frac{1}{\left(v_0+v_a+\frac{v_e}{n}\right)}(\overline{y}-\mu_0)^2$$

where

$$\overline{\mu} = \mu_0+\frac{v_0}{\left(v_0+v_a+\frac{v_e}{n}\right)}(\overline{y}-\mu_0),$$
$$V_\mu = v_0\left(1-\frac{v_0}{v_0+v_a+\frac{v_e}{n}}\right).$$

Thus

$$p\left(\mu|\mathbf{y},\mu_0,v_0,v_a,v_e\right) \propto \exp\left[-\frac{1}{2}V_\mu^{-1}\left(\mu-\overline{\mu}\right)^2\right]$$
$$\times \exp\left[-\frac{1}{2}\left(v_0+v_a+\frac{v_e}{n}\right)^{-1}\left(\overline{y}-\mu_0\right)^2\right].$$

The second term does not depend on μ, so

$$p\left(\mu|\mathbf{y},\mu_0,v_0,v_a,v_e\right) \propto \exp\left[-\frac{1}{2}V_\mu^{-1}\left(\mu-\overline{\mu}\right)^2\right]. \tag{5.22}$$

Thus the marginal posterior distribution is $\mu \sim N\left(\overline{\mu}, V_\mu\right)$. It is seen, again, that this posterior distribution depends on the data through $\overline{y}$.

Marginal posterior density of a. Note that

$$p\left(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e\right) = \frac{p\left(\mu,a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)}{p\left(a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)},$$

so

$$p\left(a|\mathbf{y},\mu_0,v_0,v_a,v_e\right) = \frac{p\left(\mu,a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)}{p\left(\mu|a,\mathbf{y},\mu_0,v_0,v_a,v_e\right)} \tag{5.23}$$

with the densities in the numerator and denominator given in (5.15) and (5.21), respectively. Alternatively, an instructive representation can be obtained by rewriting the joint density in (5.15). Employing (5.20), this can be put as

$$p\left(\mu,a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)$$
$$\propto \exp\left[-\frac{1}{2}\left\{\frac{\frac{n}{v_e}\cdot\frac{1}{v_0}}{\frac{n}{v_e}+\frac{1}{v_0}}\left(\overline{y}-\mu_0-a\right)^2+\frac{a^2}{v_a}\right\}\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_0}\right)\left\{\mu-\left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-a)+\frac{\mu_0}{v_0}\right]\right\}^2\right].$$

Integrating over μ, to obtain the marginal posterior density of a, yields

$$p\left(a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)$$
$$\propto \exp\left[-\frac{1}{2}\left\{\frac{\frac{n}{v_e}\cdot\frac{1}{v_0}}{\frac{n}{v_e}+\frac{1}{v_0}}\left(\overline{y}-\mu_0-a\right)^2+\frac{a^2}{v_a}\right\}\right]$$
$$\times \int \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_0}\right)\left\{\mu-\left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-a)+\frac{\mu_0}{v_0}\right]\right\}^2 d\mu\right]. \tag{5.24}$$

The integral involves a normal kernel and evaluates to

$$\sqrt{2\pi\left(\frac{n}{v_e}+\frac{1}{v_0}\right)^{-1}}$$

which, by not being a function of a, gets absorbed in the integration constant. Further, using (5.16) and (5.17),

$$\begin{aligned}&\frac{\frac{n}{v_e}\cdot\frac{1}{v_0}}{\frac{n}{v_e}+\frac{1}{v_0}}\left(\overline{y}-\mu_0-a\right)^2+\frac{a^2}{v_a}\\ &=\frac{n}{nv_0+v_e}\left[a-\left(\overline{y}-\mu_0\right)\right]^2+\frac{a^2}{v_a}\\ &=\frac{1}{V_a}(a-\overline{a})^2+\frac{1}{\left(v_0+v_a+\frac{v_e}{n}\right)}\left(\overline{y}-\mu_0\right)^2,\end{aligned}\tag{5.25}$$

where

$$\overline{a}=\frac{v_a}{v_0+v_a+\frac{v_e}{n}}\left(\overline{y}-\mu_0\right),$$

$$V_a=v_a\left(1-\frac{v_a}{v_0+v_a+\frac{v_e}{n}}\right).$$

Employing (5.25) in (5.24), and retaining only the term in a,

$$p\left(a|\mathbf{y},\mu_0,v_0,v_a,v_e\right)\propto\exp\left[-\frac{1}{2}V_a^{-1}(a-\overline{a})^2\right].\tag{5.26}$$

In conclusion, the marginal posterior density of the additive genetic effect is normal with mean $\overline{a}$ and variance V_a; it depends on the data only through $\overline{y}$.

Marginal distribution of the data. Finding the marginal density of the observations, i.e., the denominator of Bayes theorem, is also of interest. As we shall see later, the corresponding distribution plays an important role in model assessment. First, observe that the joint density of $\mathbf{y}$, μ, and a, given the hyperparameters, is

$$p(\mathbf{y},\mu,a|\mu_0,v_0,v_a,v_e)=\prod_{i=1}^{n}p(y_i|\mu,a,v_e)p\left(a|v_a\right)p\left(\mu|\mu_0,v_0\right)\tag{5.27}$$

with all kernels in a normal form. Using results developed previously, the forms inside of the exponents can be combined as

$$\frac{\sum_{i=1}^{n}\left(y_i-\mu-a\right)^2}{v_e}+\frac{a^2}{v_a}+\frac{\left(\mu-\mu_0\right)^2}{v_0}$$

$$= \frac{n\left[a-(\overline{y}-\mu)\right]^2 + \sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{a^2}{v_a} + \frac{(\mu-\mu_0)^2}{v_0}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{(\mu-\mu_0)^2}{v_0} + \frac{n\left[a-(\overline{y}-\mu)\right]^2}{v_e} + \frac{a^2}{v_a}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{(\mu-\mu_0)^2}{v_0}$$

$$+ \left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2$$

$$+ \frac{\frac{n}{v_e}\frac{1}{v_a}}{\frac{n}{v_e}+\frac{1}{v_a}}(\overline{y}-\mu)^2.$$

Using the preceding, the marginal density of the observations is

$$p(\mathbf{y}|\mu_0, v_0, v_a, v_e)$$

$$= \int\int \prod_{i=1}^{n} p(y_i|\mu, a, v_e)p\left(a|v_a\right)p\left(\mu|\mu_0, v_0\right)da\ d\mu$$

$$\propto \exp\left[-\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{2v_e}\right]\int \exp\left[-\frac{n(\overline{y}-\mu)^2}{2\left(nv_a+v_e\right)}\right]\exp\left[-\frac{(\mu-\mu_0)^2}{2v_0}\right]d\mu$$

$$\times \int \exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\overline{y}-\mu)\right]\right\}^2\right]da. \tag{5.28}$$

The last integral evaluates to

$$\sqrt{2\pi\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}}$$

and since it does not involve $\mathbf{y}$, it gets absorbed in the integration constant of (5.28). Thus

$$p(\mathbf{y}|\mu_0, v_0, v_a, v_e) \propto \exp\left[-\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{2v_e}\right]$$

$$\times \int \exp\left[-\frac{n(\overline{y}-\mu)^2}{2\left(nv_a+v_e\right)}\right]\exp\left[-\frac{(\mu-\mu_0)^2}{2v_0}\right]d\mu. \tag{5.29}$$

It was seen before that

$$\frac{n(\mu-\overline{y})^2}{(nv_a+v_e)}+\frac{(\mu-\mu_0)^2}{v_0}=\frac{n(\overline{y}-\mu)^2}{(nv_a+v_e)}+\frac{(\mu-\mu_0)^2}{v_0}$$
$$=\frac{(\mu-\overline{\mu})^2}{V_\mu}+\frac{(\overline{y}-\mu_0)^2}{\left(v_0+v_a+\frac{v_e}{n}\right)}.$$

Making use of this in (5.29), one gets, after integrating over μ,

$$p(\mathbf{y}|\mu_0,v_0,v_a,v_e)\propto\exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e}+\frac{(\overline{y}-\mu_0)^2}{\left(v_0+v_a+\frac{v_e}{n}\right)}\right]\right\}. \tag{5.30}$$

This is the density of an n-dimensional distribution, that depends on the data through $\overline{y}$, and through the sum of squared deviations of the observations from the mean. It will be shown that this is the kernel of the density of the n-variate normal distribution

$$\mathbf{y}|\mu_0,v_0,v_a,v_e\sim N(\mathbf{1}\mu_0,\mathbf{V}) \tag{5.31}$$

where $\mathbf{1}$ denotes a vector of $1'$s of order n, and

$$\mathbf{V}=(v_0+v_a)\,\mathbf{J}+v_e\mathbf{I}$$

is the variance–covariance matrix of the process, where $\mathbf{J}$ is an $n\times n$ matrix of $1'$s. Now, the inverse of $\mathbf{V}$ is given (Searle et al., 1992) by

$$\mathbf{V}^{-1}=\frac{1}{v_e}\left[\mathbf{I}-\frac{v_0+v_a}{v_e+n\,(v_0+v_a)}\mathbf{J}\right].$$

Hence, the kernel of the density of the multivariate normal distribution $[\mathbf{y}|\mu_0,v_0,v_a,v_e]$ can be put as

$$p\,(\mathbf{y}|\mu_0,v_0,v_a,v_e)\propto\exp\left[-\left\{\frac{1}{2}\,(\mathbf{y}-\mathbf{1}\mu_0)'\,\frac{1}{v_e}\left[\mathbf{I}-\frac{v_0+v_a}{v_e+n\,(v_0+v_a)}\mathbf{J}\right]\right\}\right.$$
$$\left.\times\,(\mathbf{y}-\mathbf{1}\mu_0)\right].$$

Now,

$$(\mathbf{y}-\mathbf{1}\mu_0)'\,\frac{1}{v_e}\left[\mathbf{I}-\frac{v_0+v_a}{v_e+n\,(v_0+v_a)}\mathbf{J}\right](\mathbf{y}-\mathbf{1}\mu_0)$$
$$=[\mathbf{y}-\mathbf{1}\overline{y}-\mathbf{1}\,(\mu_0-\overline{y})]'\,\frac{1}{v_e}\left[\mathbf{I}-\frac{v_0+v_a}{v_e+n\,(v_0+v_a)}\mathbf{J}\right]$$
$$\times\,[\mathbf{y}-\mathbf{1}\overline{y}-\mathbf{1}\,(\mu_0-\overline{y})]$$

$$= \frac{1}{v_e}\left[\sum_{i=1}^{n}(y_i-\overline{y})^2 + n(\overline{y}-\mu_0)^2 - \frac{(v_0+v_a)\,n^2(\overline{y}-\mu_0)^2}{v_e+n(v_0+v_a)}\right]$$

$$= \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{n}{v_e}(\overline{y}-\mu_0)^2\left[1-\frac{n(v_0+v_a)}{v_e+n(v_0+v_a)}\right]$$

$$= \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{(\overline{y}-\mu_0)^2}{(v_0+v_a)+\frac{v_e}{n}}.$$

Thus

$$p(\mathbf{y}|\mu_0,v_0,v_a,v_e) \propto \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{v_e} + \frac{(\overline{y}-\mu_0)^2}{\left(v_0+v_a+\frac{v_e}{n}\right)}\right]\right\}$$

and this is precisely (5.30). Hence, the marginal distribution of the observations is the normal process given in (5.31). This distribution is often referred to as the prior predictive distribution of the observation, that is, the stochastic process describing the probabilities with which data occur, before any observation is made. Its density depends entirely on the parameters of the prior distributions, commonly called "hyperparameters". This result could have been anticipated by viewing (5.13) as a random effects model, where the two independently distributed random effects have distributions $\mu \sim N(\mu_0, v_0)$ and $a \sim N(0, v_a)$. Since, in vector notation,

$$\mathbf{y} = \mathbf{1}(\mu + a) + \mathbf{e}$$

is a linear combination of normal variates, where $\mathbf{e} \sim N(0, \mathbf{I}v_e)$, it follows immediately that $\mathbf{y}$ (given μ_0) must be normal, with mean vector

$$E(\mathbf{y}\,|\mu_0) = \mathbf{1}\mu_0$$

and variance–covariance matrix

$$\begin{aligned} E(\mathbf{y}\,|v_0,v_a,v_e) &= \mathbf{1}\mathbf{1}'(v_0+v_a) + \mathbf{I}v_e \\ &= \mathbf{J}(v_0+v_a) + \mathbf{I}v_e. \end{aligned}$$

■

5.4 Posterior Distributions

Consider Bayes theorem in any of the forms given in (5.10) to (5.12), and partition the vector of all quantities subject to uncertainty as $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$,

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent distinct, unknown features of the probability model. For example, in a linear model, $\boldsymbol{\theta}_1$ may be the location vector and $\boldsymbol{\theta}_2$ the dispersion components, i.e., the variance and covariance parameters. The joint posterior density of all unknowns is

$$\begin{aligned} p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) &= \frac{L\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) g\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right)}{\int\int L\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) g\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) d\boldsymbol{\theta}_1 \, d\boldsymbol{\theta}_2} \\ &\propto L\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) g\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right), \end{aligned} \tag{5.32}$$

where $L\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right)$ is the likelihood function (joint density or distribution of the observations, viewed as a function of the unknowns), and $g\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right)$ is the joint prior density. The latter is typically a multivariate density function, and elicitation may be facilitated by writing

$$g\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = g\left(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2\right) g\left(\boldsymbol{\theta}_2\right) = g\left(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1\right) g\left(\boldsymbol{\theta}_1\right).$$

Here $g\left(\boldsymbol{\theta}_i\right)$ is the marginal prior density of $\boldsymbol{\theta}_i$, and $g\left(\boldsymbol{\theta}_j | \boldsymbol{\theta}_i\right)$ for $i \neq j$ is the conditional prior density of $\boldsymbol{\theta}_j$ given $\boldsymbol{\theta}_i$. Hence, one can assign prior probabilities to $\boldsymbol{\theta}_1$, and then to $\boldsymbol{\theta}_2$ at each of the values of $\boldsymbol{\theta}_1$, or to $\boldsymbol{\theta}_2$, and then to $\boldsymbol{\theta}_1$, given $\boldsymbol{\theta}_2$. Irrespective of the form and order of elicitation, one must end up with the same joint process; otherwise, there would be incoherence in the probabilistic ensemble. Often, it happens (sometimes for mathematical convenience) that the two sets of parameters are assigned independent prior distributions. However, it is uncommon that parameters remain mutually independent, a posteriori. For this to occur, the likelihood must factorize into independent pieces, as well, with each of the portions corresponding to each of the sets of parameters. It will be seen later that parameters that are independent a priori can become dependent a posteriori, even when a single data point is observed.

The marginal posterior densities of each parameter (or set of parameters) are, by definition

$$p\left(\boldsymbol{\theta}_1 | \mathbf{y}\right) = \int p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) d\boldsymbol{\theta}_2 \tag{5.33}$$

and

$$p\left(\boldsymbol{\theta}_2 | \mathbf{y}\right) = \int p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) d\boldsymbol{\theta}_1. \tag{5.34}$$

It may be necessary to carry out the marginalization to further levels in a Bayesian analysis. For example, suppose that $\boldsymbol{\theta}_1 = \left[\boldsymbol{\theta}_{1A}', \boldsymbol{\theta}_{1B}'\right]'$ where $\boldsymbol{\theta}_{1A}$ is a vector of additive genetic effects, say, and $\boldsymbol{\theta}_{1B}$ includes some other location parameters. Then, if one wishes to assign posterior probabilities (inference) to additive genetic effects, the marginal posterior density to be used is

$$\begin{aligned} p\left(\boldsymbol{\theta}_{1A} | \mathbf{y}\right) &= \int\int p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}\right) d\boldsymbol{\theta}_{1B} \, d\boldsymbol{\theta}_2 \\ &= \int p\left(\boldsymbol{\theta}_1 | \mathbf{y}\right) d\boldsymbol{\theta}_{1B}. \end{aligned} \tag{5.35}$$

In any of the situations described above, the parameters that are integrated out of the joint posterior density are referred to as nuisance parameters. Technically, these are components of the statistical model that need to be considered, because the probabilistic structure adopted requires it, but that are not of primary inferential interest. Suppose that $\boldsymbol{\theta}_1$ is of primary interest, in which case the distribution to be used for inference is the process $[\boldsymbol{\theta}_1|\mathbf{y}]$. The corresponding marginal density can be rewritten as

$$\begin{aligned} p(\boldsymbol{\theta}_1|\mathbf{y}) &= \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\, d\boldsymbol{\theta}_2 \\ &= \int p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\, p(\boldsymbol{\theta}_2|\mathbf{y})\, d\boldsymbol{\theta}_2 \qquad (5.36) \\ &= E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\right], \qquad (5.37) \end{aligned}$$

where $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ is the density of the conditional posterior distribution of $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$. Representations (5.36) and (5.37) indicate that the marginal density of the primary parameter $\boldsymbol{\theta}_1$ is the weighted average of an infinite number of conditional densities $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$, where the weighting or mixing function is the marginal posterior density of the nuisance parameter, $p(\boldsymbol{\theta}_2|\mathbf{y})$. The conditional posterior distribution $[\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}]$ describes the uncertainty of inferences about $\boldsymbol{\theta}_1$ that can be drawn when the nuisance parameter $\boldsymbol{\theta}_2$ is known, whereas the marginal distribution $[\boldsymbol{\theta}_2|\mathbf{y}]$ gives the relative plausibility of different values of the nuisance parameter, in the light of any prior information and of other assumptions built into the model, and of the evidence provided by the data (Box and Tiao, 1973).

The conditional posterior distributions can be identified (at least conceptually) from the joint posterior distribution, with the latter following directly from the assumptions, once a prior and a data generating process are postulated. Note that the conditional posterior density of a parameter can be expressed as

$$p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}) = \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})}{p(\boldsymbol{\theta}_2|\mathbf{y})}.$$

Since, in this distribution, one is interested in variation with respect to $\boldsymbol{\theta}_1$ only, the denominator enters merely as part of the integration constant. Thus, one can write

$$\begin{aligned} p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}) &\propto p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) \\ &\propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\, p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &\propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\, p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \\ &\propto L(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\, p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2). \qquad (5.38) \end{aligned}$$

Above, $L(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ is the likelihood function with $\boldsymbol{\theta}_2$ treated as a known constant, rather than as a feature subject to uncertainty. The preceding development implies that a conditional posterior distribution can (often) be

identified by inspection of the joint posterior density and by retaining only the parts that vary with the parameter(s) of interest, treating the remaining parts as known. This method can be useful for identifying conditional posterior distributions in the context of MCMC methods, as discussed in a subsequent chapter.

Example 5.5 *Posterior dependence between parameters*
Suppose that a sample of size n is obtained by drawing independently from the same normal distribution $N\left(\mu, \sigma^2\right)$, where the mean and variance are both unknown. Assume that the parameters are taken as following independent distributions, with prior densities,

$$p\left(\mu|a,b\right) = \frac{1}{b-a}, \quad p\left(\sigma^2|c,d\right) = \frac{1}{d-c},$$

where a, b, c, d are bounds on parameter values that have been elicited somehow. The joint posterior density is

$$p\left(\mu,\sigma^2|\mathbf{y},a,b,c,d\right) \propto \prod_{i=1}^{n}\left(\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right] \frac{1}{\left(b-a\right)\left(d-c\right)}$$

$$\propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right]$$

$$\propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 + n\left(\mu-\overline{y}\right)^2}{2\sigma^2}\right] \tag{5.39}$$

this being nonnull for $a < \mu < b$ and $c < \sigma^2 < d$. The marginal posterior density of σ^2 is obtained by integrating over μ:

$$p\left(\sigma^2|\mathbf{y},a,b,c,d\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right]$$

$$\times \int_a^b \frac{\sqrt{2\pi\sigma^2/n}}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\tfrac{n(\mu-\overline{y})^2}{2\sigma^2}\right] d\mu$$

$$\propto \left(\sigma^2\right)^{-\frac{n-1}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right] \left[\Phi\left(\tfrac{b-\overline{y}}{\sigma/\sqrt{n}}\right) - \Phi\left(\tfrac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right]. \tag{5.40}$$

The difference between the integrals is equal to 1 if μ is allowed to take values anywhere in the real line (i.e., if $a = -\infty$ and $b = \infty$). The marginal

posterior density of μ is found by integrating (5.39) with respect to σ^2:

$$p(\mu|\mathbf{y}, a, b, c, d) \propto \int_c^d \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right] d\sigma^2$$
$$\propto \int_c^d \left(\sigma^2\right)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_\mu}{\sigma^2}\right] d\sigma^2,$$

where

$$S_\mu = \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2 .$$

The integral above cannot be written in closed form. However, if one takes the positive part of the real line as parameter space for σ^2, that is, $c = 0$ and $d = \infty$, use of properties of the inverse gamma (or scaled inverse chi-square) distribution seen in Chapter 1 yields

$$p(\mu|\mathbf{y}, a, b) \propto \int_0^\infty \left(\sigma^2\right)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_\mu}{\sigma^2}\right] d\sigma^2 = \frac{\Gamma\left(\frac{n-2}{2}\right)}{S_\mu^{\frac{n-2}{2}}}$$
$$\propto S_\mu^{-\frac{n-2}{2}} \propto \left[\sum_{i=1}^n (y_i-\overline{y})^2 + n(\mu-\overline{y})^2\right]^{-\frac{n-2}{2}} .$$

In this density, only variation with respect to μ is of concern. Hence, one can factor out the sum of squared deviations of the observations from the sample mean, to obtain

$$p(\mu|\mathbf{y}, a, b) \propto \left[1 + \frac{n(\mu-\overline{y})^2}{\sum_{i=1}^n (y_i-\overline{y})^2}\right]^{-\frac{n-3+1}{2}}$$
$$\propto \left[1 + \frac{(\mu-\overline{y})^2}{(n-3)\frac{\widehat{s}^2}{n}}\right]^{-\frac{n-3+1}{2}}, \qquad (5.41)$$

where

$$\widehat{s}^2 = \frac{1}{n-3}\sum_{i=1}^n (y_i-\overline{y})^2 .$$

If μ were allowed to take values anywhere in the real line, (5.41) gives the kernel of a t-distribution with $n-3$ degrees of freedom, mean $\overline{y}$, and variance equal to

$$\frac{\widehat{s}^2(n-3)}{n(n-5)} = \frac{\sum_{i=1}^n (y_i-\overline{y})^2}{n(n-5)}.$$

This distribution is proper if $n > 3$, and the variance is finite if $n > 5$. However, since in the example μ takes density only in the $[a, b]$ interval, it turns out that the marginal posterior distribution is a truncated t-process with density

$$p(\mu|\mathbf{y}, a, b) = \frac{\left[1 + \frac{(\mu-\overline{y})^2}{(n-3)\frac{\widehat{s}^2}{n}}\right]^{-\frac{n-3+1}{2}}}{\int_a^b \left[1 + \frac{(\mu-\overline{y})^2}{(n-3)\frac{\widehat{s}^2}{n}}\right]^{-\frac{n-3+1}{2}} d\mu}. \tag{5.42}$$

Finally, note that the product of (5.42) and (5.40) does not yield (5.39), even if $a = -\infty$, $b = \infty$, $c = 0$, and $d = \infty$. Hence μ and σ^2 are not independent in the posterior distribution, even if they are so, a priori. ■

Example 5.6 *Conditional posterior distribution*
Revisit Example 5.5, and consider finding the two induced conditional posterior distributions. From representation (5.39) of the joint posterior density, one can deduce the density of $\left[\mu|\mathbf{y}, a, b, c, d, \sigma^2\right]$, treating σ^2 as a constant. This yields

$$p\left(\mu|\mathbf{y}, a, b, c, d, \sigma^2\right) \propto \exp\left[-\frac{n(\mu-\overline{y})^2}{2\sigma^2}\right] I(a < \mu < b)$$

which is the density of a normal distribution truncated between a and b. The mean and variance of the untruncated distribution are $\overline{y}$ and σ^2/n, respectively. Similarly, the density of the conditional distribution $\left[\sigma^2|\mathbf{y}, a, b, c, d, \mu\right]$ is arrived at by regarding μ as a constant in the joint density, to obtain

$$p\left(\sigma^2|\mathbf{y}, a, b, c, d, \mu\right) \propto \left(\sigma^2\right)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_\mu}{\sigma^2}\right] I\left(c < \sigma^2 < d\right).$$

This is in an inverse gamma form (truncated between c and d), and the parameters of the distribution in the absence of truncation are $(n-2)/2$ and S_μ. ■

Example 5.7 *Posterior distributions in a linear regression model with t distributed errors*
Suppose a response y is related to a predictor variable x according to the linear relationship

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n, \tag{5.43}$$

where the residuals are i.i.d. as

$$t\left(0, \sigma^2, \nu\right),$$

where σ^2 is the scale parameter of the t-distribution and ν are the degrees of freedom, with the latter assumed known. Recall from Chapter 1 that an equivalent representation of (5.43) is given by

$$y_i = \beta_0 + \beta_1 x_i + \frac{e_i}{\sqrt{w_i}}, \tag{5.44}$$

where $e_i \sim N\left(0, \sigma^2\right)$ and $w_i \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ are independently distributed random variables, $i = 1, 2, ..., n$. If model (5.44) is deconditioned with respect to the gamma random variable, then (5.43) results. Let the parameter vector include the unknown parameters of the model, that is, the regression coefficients and σ^2, plus all the gamma weights w_i. This is known as "data augmentation" (Tanner and Wong, 1987) and is discussed in Chapter 11. The augmented parameter vector is

$$\boldsymbol{\theta} = \left[\beta_0, \beta_1, \sigma^2, w_1, w_2, ..., w_n\right]'$$

and adopt as joint prior density

$$p\left(\beta_0, \beta_1, \sigma^2, w_1, w_2, ..., w_n|\nu\right) = p\left(\beta_0\right) p\left(\beta_1\right) p\left(\sigma^2\right) \prod_{i=1}^{n} p\left(w_i|\nu\right). \tag{5.45}$$

The joint posterior density is

$$\begin{aligned} p\left(\beta_0, \beta_1, \sigma^2, w_1, w_2, ..., w_n|\mathbf{y}, \nu\right) \propto & \prod_{i=1}^{n} \left[p\left(y_i|\beta_0, \beta_1, \sigma^2, w_i\right) p\left(w_i|\nu\right)\right] \\ & \times p\left(\beta_0\right) p\left(\beta_1\right) p\left(\sigma^2\right). \end{aligned} \tag{5.46}$$

Explicitly, this takes the form

$$\begin{aligned} & p\left(\beta_0, \beta_1, \sigma^2, w_1, w_2, ..., w_n|\mathbf{y}, \nu\right) \\ \propto & \prod_{i=1}^{n} \left(\frac{\sigma^2}{w_i}\right)^{-\frac{1}{2}} w_i^{\frac{\nu}{2}-1} \exp\left\{-\frac{w_i}{2}\left[\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2 + \nu\sigma^2}{\sigma^2}\right]\right\} \\ & \times p\left(\beta_0\right) p\left(\beta_1\right) p\left(\sigma^2\right). \end{aligned} \tag{5.47}$$

Conditional posterior distribution of w_i given all other parameters. Note in (5.47) that, given all other parameters, the w's are mutually independent. The kernel of the conditional posterior density of w_i is found by collecting terms that depend on this random variable. Thus, for $i = 1, 2, ..., n$,

$$p\left(w_i|\beta_0, \beta_1, \sigma^2, \mathbf{y}, \nu\right) \propto w_i^{\frac{\nu+1}{2}-1} \exp\left(-\frac{w_i S_i}{2}\right),$$

where

$$S_i = \frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2 + \nu\sigma^2}{\sigma^2}.$$

This implies that the conditional posterior distribution of w_i is the gamma process

$$w_i|\beta_0,\beta_1,\sigma^2,\mathbf{y},\nu \sim Ga\left(\frac{\nu+1}{2},\frac{\nu+\left(y_i-\beta_0-\beta_1x_i\right)^2/\sigma^2}{2}\right). \tag{5.48}$$

The observations enter only through data point i.

Conditional posterior distribution of β_0 and β_1 given all other parameters. From the joint density (5.47):

$$\begin{aligned} &p\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu\right) \\ &\propto \prod_{i=1}^{n}\exp\left[-\frac{w_i}{2\sigma^2}\left(y_i-\beta_0-\beta_1x_i\right)^2\right]p\left(\beta_0\right)p\left(\beta_1\right). \end{aligned} \tag{5.49}$$

It is not possible to be more specific about the form of the distribution unless the priors are stated explicitly. For example, suppose that, a priori, $\beta_0 \sim N\left(\alpha_0,\sigma^2_{\beta_0}\right)$ and $\beta_1 \sim N\left(\alpha_1,\sigma^2_{\beta_1}\right)$. Then,

$$\begin{aligned} &p\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}w_i\left(y_i-\beta_0-\beta_1x_i\right)^2\right.\right. \\ &\left.\left.+\frac{\sigma^2}{\sigma^2_{\beta_0}}\left(\beta_0-\alpha_0\right)^2+\frac{\sigma^2}{\sigma^2_{\beta_1}}\left(\beta_1-\alpha_1\right)^2\right]\right\}. \end{aligned} \tag{5.50}$$

Now

$$\sum_{i=1}^{n}w_i\left(y_i-\beta_0-\beta_1x_i\right)^2=\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right), \tag{5.51}$$

where

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$$

and

$$\begin{aligned} \mathbf{W} &= Diag\left(w_i\right), \quad \mathbf{X}=\begin{bmatrix}\mathbf{1} & \mathbf{x}\end{bmatrix}_{n\times 2}, \\ \mathbf{1} &= \{1\}, \quad \mathbf{x}=[x_1,x_2,...,x_n]'. \end{aligned}$$

Define the following function of the data (and of the w's):

$$\widehat{\boldsymbol{\beta}}=\begin{bmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{bmatrix}=\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{5.52}$$

Then

$$\begin{aligned}
&(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\,\mathbf{W}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&= \left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\
&= \left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right]' \mathbf{W}\left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right] \\
&= \left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right) \qquad (5.53)
\end{aligned}$$

because the cross-product term vanishes, as a consequence of the definition of $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}
2\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\mathbf{X}'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right) &= 2\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\mathbf{X}'\mathbf{W}\mathbf{y}-\mathbf{X}'\mathbf{W}\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\
&= 2\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)(\mathbf{X}'\mathbf{W}\mathbf{y}-\mathbf{X}'\mathbf{W}\mathbf{y}) = \mathbf{0}.
\end{aligned}$$

Employing (5.53) in (5.51), the posterior density in (5.50) becomes

$$\begin{aligned}
&p\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)\right.\right. \\
&\left.\left.+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)+\frac{\sigma^2(\beta_0-\alpha_0)^2}{\sigma^2_{\beta_0}}+\frac{\sigma^2(\beta_1-\alpha_1)^2}{\sigma^2_{\beta_1}}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right.\right. \\
&\left.\left.+\frac{\sigma^2(\beta_0-\alpha_0)^2}{\sigma^2_{\beta_0}}+\frac{\sigma^2(\beta_1-\alpha_1)^2}{\sigma^2_{\beta_1}}\right]\right\}, \qquad (5.54)
\end{aligned}$$

upon retaining only the terms that vary with $\boldsymbol{\beta}$. Defining

$$\boldsymbol{\alpha}=\begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}, \quad \boldsymbol{\Lambda}=\begin{bmatrix}\frac{\sigma^2}{\sigma^2_{\beta_0}} & 0\\ 0 & \frac{\sigma^2}{\sigma^2_{\beta_1}}\end{bmatrix}=\begin{bmatrix}\lambda_0 & 0\\ 0 & \lambda_1\end{bmatrix}$$

the conditional posterior that concerns us is expressible as

$$\begin{aligned}
&p\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\
&\propto \quad \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right.\right. \\
&\left.\left.+(\boldsymbol{\beta}-\boldsymbol{\alpha})'\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\alpha})\right]\right\}. \qquad (5.55)
\end{aligned}$$

The two quadratic forms on $\boldsymbol{\beta}$ can be combined using an extension of formulas (5.16) and (5.17), as given in Box and Tiao (1973),

$$\begin{aligned} & (\mathbf{z}-\mathbf{m})'\mathbf{M}(\mathbf{z}-\mathbf{m})+(\mathbf{z}-\mathbf{b})'\mathbf{B}(\mathbf{z}-\mathbf{b}) \\ = \; & (\mathbf{z}-\mathbf{c})'(\mathbf{M}+\mathbf{B})(\mathbf{z}-\mathbf{c}) \\ & +(\mathbf{m}-\mathbf{b})'\,\mathbf{M}\,(\mathbf{M}+\mathbf{B})^{-1}\,\mathbf{B}\,(\mathbf{m}-\mathbf{b}) \end{aligned} \tag{5.56}$$

with

$$\mathbf{c}=(\mathbf{M}+\mathbf{B})^{-1}\,(\mathbf{M}\mathbf{m}+\mathbf{B}\mathbf{b})\,. \tag{5.57}$$

Employing this in the context of (5.55):

$$\begin{aligned} & \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)+(\boldsymbol{\beta}-\boldsymbol{\alpha})'\,\boldsymbol{\Lambda}\,(\boldsymbol{\beta}-\boldsymbol{\alpha}) \\ & =\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'(\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right) \\ & +(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha})'\mathbf{X}'\mathbf{W}\mathbf{X}\,(\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\,\boldsymbol{\Lambda}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\alpha}) \end{aligned} \tag{5.58}$$

with

$$\begin{aligned} \overline{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{\Lambda}\boldsymbol{\alpha}\right) && (5.59)\\ &= (\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\,(\mathbf{X}'\mathbf{W}\mathbf{y}+\boldsymbol{\Lambda}\boldsymbol{\alpha})\,. && (5.60) \end{aligned}$$

Using (5.58) in (5.55) and retaining only the part that varies with $\boldsymbol{\beta}$:

$$\begin{aligned} & p\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\ & \propto \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'(\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)\right]. \end{aligned} \tag{5.61}$$

Thus, the posterior distribution of the regression coefficients, given all other parameters, is normal, with mean vector as in (5.59) or (5.60) and variance–covariance matrix

$$Var\left(\beta_0,\beta_1|\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right)=(\mathbf{X}'\mathbf{W}\mathbf{X}+\boldsymbol{\Lambda})^{-1}\,\sigma^2. \tag{5.62}$$

Note that the mean of this posterior distribution is a matrix-weighted average of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}$, where the weights are $\mathbf{X}'\mathbf{W}\mathbf{X}$ and $\boldsymbol{\Lambda}$, respectively.

Conditional posterior distribution of β_1 given β_0 and all other parameters. If β_0 is known, it can be treated as an offset in the model, that is, one can write a "new" response variable

$$r_i=y_i-\beta_0=\beta_1 x_i+\epsilon_i.$$

Put $\mathbf{r}=\{r_i\}$. Using this in the joint posterior density, and treating β_0 as a constant, yields

$$\begin{aligned} & p\left(\beta_1|\beta_0,\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\ \propto \; & \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}w_i\,(r_i-\beta_1 x_i)^2+\lambda_1\,(\beta_1-\alpha_1)^2\right]\right\}. \end{aligned}$$

Using similar algebra as in (5.51)to (5.60), the conditional density can be represented in the Gaussian form

$$\begin{aligned} & p\left(\beta_1|\beta_0,\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\ & \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} w_i\left(r_i-\beta_1 x_i\right)^2+\lambda_1\left(\beta_1-\alpha_1\right)^2\right]\right\} \\ & \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta_1-\overline{\beta}_{1.0}\right)'\left(\mathbf{x}'\mathbf{W}\mathbf{x}+\lambda_1\right)\left(\beta_1-\overline{\beta}_{1.0}\right)\right], \end{aligned} \tag{5.63}$$

where

$$\begin{aligned} \overline{\beta}_{1.0} &= \left(\mathbf{x}'\mathbf{W}\mathbf{x}+\lambda_1\right)^{-1}\left(\mathbf{x}'\mathbf{W}\mathbf{r}+\lambda_1\alpha_1\right) \\ &= \frac{\sum_{i=1}^{n} w_i x_i\left(y_i-\beta_0\right)+\lambda_1\alpha_1}{\sum_{i=1}^{n} w_i x_i^2+\lambda_1} \end{aligned}$$

and the variance of the process is

$$\begin{aligned} Var\left(\beta_1|\beta_0,\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) &= \left(\mathbf{x}'\mathbf{W}\mathbf{x}+\lambda_1\right)^{-1}\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} w_i x_i^2+\lambda_1}. \end{aligned}$$

Conditional posterior distribution of β_0 given β_1 and all other parameters. The development is similar, except that the offset is now $\beta_1 x_i$, to form the "new" response

$$t_i = y_i - \beta_1 x_i = \beta_0 + \epsilon_i$$

with $\mathbf{t} = \{y_i - \beta_1 x_i\}$. The resulting density is

$$\begin{aligned} & p\left(\beta_0|\beta_1,\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right) \\ & \propto \exp\left[-\frac{1}{2\sigma^2}\left(\beta_0-\overline{\beta}_{0.1}\right)'\left(\mathbf{1}'\mathbf{W}\mathbf{1}+\lambda_0\right)\left(\beta_0-\overline{\beta}_{0.1}\right)\right], \end{aligned} \tag{5.64}$$

where

$$\begin{aligned} \overline{\beta}_{0.1} &= \left(\mathbf{1}'\mathbf{W}\mathbf{1}+\lambda_0\right)^{-1}\left(\mathbf{1}'\mathbf{W}\mathbf{t}+\lambda_0\alpha_0\right) \\ &= \frac{\sum_{i=1}^{n} w_i\left(y_i-\beta_1 x_i\right)+\lambda_0\alpha_0}{\sum_{i=1}^{n} w_i+\lambda_0} \end{aligned}$$

and

$$Var\left(\beta_0|\beta_1,\sigma^2,\mathbf{w},\mathbf{y},\nu,\alpha_0,\sigma^2_{\beta_0},\alpha_1,\sigma^2_{\beta_1}\right)=\frac{\sigma^2}{\sum\limits_{i=1}^{n}w_i+\lambda_0}.$$

Conditional posterior distribution of σ^2 given all other parameters. Retaining terms in σ^2 in the joint density of all parameters given in (5.47), one obtains

$$\begin{aligned}&p\left(\sigma^2|\beta_0,\beta_1,w_1,w_2,...,w_n,\mathbf{y},\nu\right)\\&\propto\left(\sigma^2\right)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}w_i\left(y_i-\beta_0-\beta_1x_i\right)^2\right]p\left(\sigma^2\right).\end{aligned}\tag{5.65}$$

It is not possible to go further unless an explicit statement is made about the prior distribution of σ^2. For example, suppose that the prior distribution is lognormal, that is, the logarithm of σ^2 follows a Gaussian distribution with mean 0 and variance ω. Then, as seen in Chapter 2, the prior density of σ^2 takes the form

$$p\left(\sigma^2|\omega\right)\propto\left(\sigma^2\right)^{-1}\exp\left[-\frac{\left(\log\sigma^2\right)^2}{2\omega}\right].$$

The conditional posterior density of σ^2 would be

$$\begin{aligned}&p\left(\sigma^2|\beta_0,\beta_1,\mathbf{w},\mathbf{y},\nu\right)\\&\propto\left(\sigma^2\right)^{-\frac{n+2}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}w_i\left(y_i-\beta_0-\beta_1x_i\right)^2+\frac{\sigma^2\left(\log\sigma^2\right)^2}{\omega}\right]\right\},\end{aligned}$$

which is not in a recognizable form. In this situation, further analytical treatment is not feasible. On the other hand, suppose that elicitation yields a scaled inverse chi-square distribution with parameters Q and R as a reasonable prior. The corresponding density (see Chapter 1) is

$$p\left(\sigma^2|Q,R\right)\propto\left(\sigma^2\right)^{-\left(\frac{Q}{2}+1\right)}\exp\left[-\frac{QR}{2\sigma^2}\right].\tag{5.66}$$

Upon using (5.66) in (5.65), the conditional posterior density of σ^2 turns out to be

$$p\left(\sigma^2|\beta_0,\beta_1,w_1,w_2,...,w_n,\mathbf{y},\nu\right)\propto\left(\sigma^2\right)^{-\left(\frac{n+Q}{2}+1\right)}\exp\left\{-\frac{Q^*R^*}{2\sigma^2}\right\}.\tag{5.67}$$

This is the density of a scaled inverse chi-square process with parameters $Q^*=n+Q$ and

$$R^*=\frac{ns^2+QR}{n+Q},$$

with

$$s^2 = \frac{\sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{n}.$$

Note that R^* is a weighted average between R ("a prior value of the variance") and s^2 ("a variance provided by the data, recalling that the w's are treated as observed in this conditional distribution). This is another example of the data-prior compromise that arises in Bayesian analysis.

Conditional posterior distribution of β_0 and β_1 given w. Suppose the prior density of σ^2 is scaled inverse chi-squared, as in (5.66). Then, using this in (5.47) and integrating over σ^2 while keeping $\mathbf{w}$ fixed, gives

$$\begin{aligned} p\left(\beta_0, \beta_1 | \mathbf{w}, \mathbf{y}, Q, R, \nu\right) &\propto p\left(\beta_0\right) p\left(\beta_1\right) \int_0^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}} \\ &\times \prod_{i=1}^{n} \exp\left[-\frac{w_i}{2} \frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\sigma^2}\right] \\ &\times \left(\sigma^2\right)^{-\left(\frac{Q}{2}+1\right)} \exp\left[-\frac{QR}{2\sigma^2}\right] d\sigma^2 \\ &\propto p\left(\beta_0\right) p\left(\beta_1\right) \int_0^{\infty} \left(\sigma^2\right)^{-\left(\frac{n+Q}{2}+1\right)} \\ &\times \exp\left\{-\frac{\sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2 + QR}{2\sigma^2}\right\} d\sigma^2 \\ &\propto p\left(\beta_0\right) p\left(\beta_1\right) \int_0^{\infty} \left(\sigma^2\right)^{-\left(\frac{n+Q}{2}+1\right)} \\ &\times \exp\left\{-\frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{W} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + QR}{2\sigma^2}\right\} d\sigma^2 \end{aligned} \tag{5.68}$$

after making use of (5.51). The integrand is in a scaled inverse gamma form, so the expression becomes

$$\begin{aligned} &p\left(\beta_0, \beta_1 | \mathbf{w}, \mathbf{y}, Q, R, \nu\right) \\ &\propto p\left(\beta_0\right) p\left(\beta_1\right) \left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{W} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + QR\right]^{-\left(\frac{n+Q}{2}\right)}. \end{aligned}$$

Recall the decomposition in (5.53), that is,

$$\begin{aligned} &\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{W} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \\ = \;& \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \mathbf{W} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{W}\mathbf{X} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right). \end{aligned}$$

Using this in the preceding density, rearranging and keeping only terms in $\boldsymbol{\beta}$ gives

$$p\left(\beta_0,\beta_1|\mathbf{w},\mathbf{y},Q,R,\nu\right)$$

$$\propto p\left(\beta_0\right)p\left(\beta_1\right)\left[1+\frac{\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)+QR}\right]^{-\left(\frac{n-2+Q+2}{2}\right)}$$

$$\propto p\left(\beta_0\right)p\left(\beta_1\right)\left[1+\frac{\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{W}\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{\left(n-2+Q\right)k^2}\right]^{-\left(\frac{n-2+Q+2}{2}\right)}, \qquad (5.69)$$

where

$$k^2=\frac{\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)+QR}{\left(n-2+Q\right)}.$$

The expression in brackets in (5.69) is the kernel of the density of a bivariate t distribution having mean vector $\widehat{\boldsymbol{\beta}}$, scale parameter matrix

$$\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}k^2$$

and $n-2+Q$ degrees of freedom. However, the form of the distribution $[\beta_0,\beta_1|\mathbf{w},\mathbf{y},Q,R,\nu]$ will depend on the form of the priors adopted for the regression coefficients. For example, if these priors are uniform, the posterior distribution is a truncated bivariate-t defined inside the corresponding boundaries.

Joint posterior density of β_0, β_1, *and* σ^2, *unconditionally to* w. This is obtained by integrating over $\mathbf{w}$. Note that the joint density (5.47) factorizes into n independent parts that can be integrated separately, to obtain

$$p\left(\beta_0,\beta_1,\sigma^2|\mathbf{y},\nu\right)\propto\prod_{i=1}^{n}\int_0^\infty\left(\frac{\sigma^2}{w_i}\right)^{-\frac{1}{2}}$$

$$\times w_i^{\frac{\nu}{2}-1}\exp\left\{-\frac{w_i}{2}\left[\frac{\left(y_i-\beta_0-\beta_1x_i\right)^2+\nu\sigma^2}{\sigma^2}\right]\right\}dw_i p\left(\beta_0\right)p\left(\beta_1\right)p\left(\sigma^2\right)$$

$$\propto p\left(\beta_0\right)p\left(\beta_1\right)p\left(\sigma^2\right)\left(\sigma^2\right)^{-\frac{n}{2}}\prod_{i=1}^{n}\int_0^\infty w_i^{\frac{\nu+1}{2}-1}\exp\left\{-\frac{w_iS_i}{2}\right\}dw_i$$

recalling that

$$S_i=\frac{\left(y_i-\beta_0-\beta_1x_i\right)^2+\nu\sigma^2}{\sigma^2}.$$

The integrand in the preceding expression is the kernel of a gamma density with parameters $(\nu+1)/2$ and $S_i/2$, so that one obtains

$$\begin{aligned} p\left(\beta_0,\beta_1,\sigma^2|\mathbf{y},\nu\right) &\propto p\left(\beta_0\right)p\left(\beta_1\right)p\left(\sigma^2\right)\left(\sigma^2\right)^{-\frac{n}{2}} \\ &\times\prod_{i=1}^{n}\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{S_i}{2}\right)^{-\frac{\nu+1}{2}} \\ &\propto p\left(\beta_0\right)p\left(\beta_1\right)p\left(\sigma^2\right)\left(\sigma^2\right)^{-\frac{n}{2}} \\ &\times\prod_{i=1}^{n}\left[\frac{\left(y_i-\beta_0-\beta_1x_i\right)^2+\nu\sigma^2}{\sigma^2}\right]^{-\frac{\nu+1}{2}} \\ &\propto p\left(\beta_0\right)p\left(\beta_1\right)p\left(\sigma^2\right) \\ &\times\prod_{i=1}^{n}\left(\sigma^2\right)^{-\frac{1}{2}}\left[1+\frac{\left(y_i-\beta_0-\beta_1x_i\right)^2}{\nu\sigma^2}\right]^{-\frac{\nu+1}{2}}. \end{aligned} \tag{5.70}$$

This is precisely the joint posterior distribution of all parameters for the linear regression model in (5.43), assuming known degrees of freedom. Additional marginalization is not possible by analytical means, but one can estimate lower-dimensional posterior distributions of interest using Monte Carlo methods, as discussed later in the book. ■

5.5 Bayesian Updating

The concept of Bayesian learning in a discrete setting was discussed in a previous section; see (5.5). This will be revisited for continuous-valued parameters and observations. Suppose that data accrue sequentially as $\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_K$, and that the problem is to infer a parameter vector $\boldsymbol{\theta}$. The posterior density of $\boldsymbol{\theta}$ is

$$\begin{aligned} p\left(\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_K\right) &= \frac{g\left(\boldsymbol{\theta}\right)p\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_K|\boldsymbol{\theta}\right)}{m\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_K\right)} \\ &= \frac{g\left(\boldsymbol{\theta}\right)p\left(\mathbf{y}_1|\boldsymbol{\theta}\right)p\left(\mathbf{y}_2|\mathbf{y}_1,\boldsymbol{\theta}\right)...p\left(\mathbf{y}_K|\mathbf{y}_1,\ldots,\mathbf{y}_{K-1},\boldsymbol{\theta}\right)}{m\left(\mathbf{y}_1\right)m\left(\mathbf{y}_2|\mathbf{y}_1\right)\ldots m\left(\mathbf{y}_K|\mathbf{y}_1,\ldots,\mathbf{y}_{K-1}\right)} \\ &\propto g\left(\boldsymbol{\theta}\right)p\left(\mathbf{y}_1|\boldsymbol{\theta}\right)p\left(\mathbf{y}_2|\mathbf{y}_1,\boldsymbol{\theta}\right)\ldots p\left(\mathbf{y}_K|\mathbf{y}_1,\ldots,\mathbf{y}_{K-1},\boldsymbol{\theta}\right). \end{aligned} \tag{5.71}$$

The posterior distribution of any function $h(\boldsymbol{\theta})$ is arrived at by making a one-to-one transformation of the parameter vector such that one of the new variables is h (or set of variables, if h is vector-valued), and then integrating over the remaining "dummy" variables. From (5.71) it follows that

$$\begin{aligned} p\left(\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_K\right) &\propto p\left(\boldsymbol{\theta}|\mathbf{y}_1\right)p\left(\mathbf{y}_2|\mathbf{y}_1,\boldsymbol{\theta}\right)\ldots p\left(\mathbf{y}_K|\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{K-1},\boldsymbol{\theta}\right) \\ &\propto p\left(\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2\right)\ldots p\left(\mathbf{y}_K|\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{K-1},\boldsymbol{\theta}\right) \\ &\propto p\left(\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{K-1}\right)p\left(\mathbf{y}_K|\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{K-1},\boldsymbol{\theta}\right). \end{aligned} \tag{5.72}$$

Note that the posterior distribution at stage i of learning acts as a prior for stage $i+1$, as already noted in the discrete case. This implies that a single Bayesian analysis carried out at the end of the process will lead to the same inferences about $\boldsymbol{\theta}$ as one carried out sequentially. If data accruing at different stages are conditionally independent, (5.71) becomes

$$\begin{aligned} p\left(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\right) &\propto g\left(\boldsymbol{\theta}\right) p\left(\mathbf{y}_1|\boldsymbol{\theta}\right) p\left(\mathbf{y}_2|\boldsymbol{\theta}\right) \ldots p\left(\mathbf{y}_K|\boldsymbol{\theta}\right) \\ &\propto g\left(\boldsymbol{\theta}\right) \prod_{i=1}^{K} p\left(\mathbf{y}_i|\boldsymbol{\theta}\right). \end{aligned} \tag{5.73}$$

Example 5.8 *Progeny test of dairy bulls*
Suppose that S unrelated dairy bulls are mated to unrelated cows, leading to n_i daughters per bull. Suppose, for simplicity, that the milk production of such daughters is measured under the same environmental conditions. A linear model for the production of daughter j of bull i could be

$$y_{ij} = s_i + e_{ij},$$

where, in the dairy cattle breeding lexicon, s_i is known as the "transmitting ability" of bull i, and $e_{ij} \sim N\left(0, \sigma_e^2\right)$ is a residual, assumed to be independently distributed of any s_i, and of any other residual. Let the average production of daughters of bull i be $\overline{y}_i$. The model for such an average is

$$\overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = s_i + \overline{e}_i,$$

where $\overline{e}_i \sim N\left(0, \sigma_e^2/n_i\right)$ is the average of individual residuals. We seek to infer s_i, given information on the average production of the daughters. Suppose that, a priori, each of the transmitting abilities is assigned the distribution $s_i \sim N\left(\mu, \sigma_s^2\right)$, and that these are independent among bulls; assume that μ, σ_s^2, and σ_e^2 are known. The posterior distribution of all transmitting abilities, based on averages is

$$\begin{aligned} &p\left(s_1, s_2, \ldots, s_S|\overline{y}, \overline{y}_2, \ldots, \overline{y}_S, \mu, \sigma_s^2, \sigma_e^2\right) \\ &\propto \prod_{i=1}^{S} \exp\left[-\frac{n_i}{2\sigma_e^2}\left(\overline{y}_i - s_i\right)^2\right] \prod_{i=1}^{S} \exp\left[-\frac{1}{2\sigma_s^2}\left(s_i - \mu\right)^2\right] \\ &= \prod_{i=1}^{S} \exp\left\{-\frac{1}{2\sigma_e^2}\left[n_i\left(s_i - \overline{y}_i\right)^2 + \frac{\sigma_e^2}{\sigma_s^2}\left(s_i - \mu\right)^2\right]\right\}. \end{aligned} \tag{5.74}$$

Hence, the transmitting abilities of all bulls are also independent a posteriori; this would not be so if bulls were genetically related, in which case a multivariate prior would need to be elicited (a situation to be encountered

in the following example). Now, the two quadratic forms in the transmitting ability in (5.74) can be combined employing (5.16) as

$$n_i \left(s_i - \overline{y}_i\right)^2 + \frac{\sigma_e^2}{\sigma_s^2}\left(s_i - \mu\right)^2 = \left(n_i + \frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i - \overline{s}_i\right)^2 + \frac{n_i \frac{\sigma_e^2}{\sigma_s^2}}{n_i + \frac{\sigma_e^2}{\sigma_s^2}}\left(\overline{y}_i - \mu\right)^2, \tag{5.75}$$

where:

$$\begin{aligned}\overline{s}_i &= \left(n_i + \frac{\sigma_e^2}{\sigma_s^2}\right)^{-1}\left(n_i \overline{y}_i + \frac{\sigma_e^2}{\sigma_s^2}\mu\right) \\ &= \frac{n_i \overline{y}_i}{n_i + \frac{\sigma_e^2}{\sigma_s^2}} + \left(1 - \frac{n_i}{n_i + \frac{\sigma_e^2}{\sigma_s^2}}\right)\mu \\ &= \mu + \frac{n_i}{n_i + \frac{\sigma_e^2}{\sigma_s^2}}\left(\overline{y}_i - \mu\right). \end{aligned} \tag{5.76}$$

Using (5.75) in (5.74) and retaining only the portion that varies with s_i gives as posterior density of the transmitting ability of bull i:

$$p\left(s_i | \overline{y}_i, \mu, \sigma_s^2, \sigma_e^2\right) \propto \exp\left[-\frac{\left(n_i + \frac{\sigma_e^2}{\sigma_s^2}\right)}{2\sigma_e^2}\left(s_i - \overline{s}_i\right)^2\right]. \tag{5.77}$$

Hence, the posterior distribution is normal, with mean as in (5.76) and variance

$$Var\left(s_i | \overline{y}_i, \mu, \sigma_s^2, \sigma_e^2\right) = \frac{\sigma_e^2}{n_i + \frac{\sigma_e^2}{\sigma_s^2}}. \tag{5.78}$$

Suppose now that, for each sire, the data arrive sequentially in two "crops" of daughters of sizes n_{i1} and n_{i2}, respectively. We proceed to verify that if the posterior distribution after crop 1 is used as prior for crop 2, one obtains the posterior distribution with density as in (5.77). From the preceding developments, it follows immediately that the posterior density after crop 1 is

$$p\left(s_i | \overline{y}_{i1}, \mu, \sigma_s^2, \sigma_e^2\right) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(n_{i1} + \frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i - \overline{s}_{i1}\right)^2\right\}, \tag{5.79}$$

where $\overline{y}_{i1}$ is the average production of daughters in crop 1, and $\overline{s}_{i1}$ is the mean of the posterior distribution after crop 1. Using this as prior for crop 2, the posterior distribution, after observing that the average production of the n_{i2} cows is $\overline{y}_{i2}$, is

$$\begin{aligned} &p\left(s_i | \overline{y}_{i1}, \overline{y}_{i2}, \mu, \sigma_s^2, \sigma_e^2\right) \\ &\propto \exp\left[-\frac{n_{i2}\left(\overline{y}_{i2} - s_i\right)^2}{2\sigma_e^2} - \frac{\left(n_{i1} + \frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i - \overline{s}_{i1}\right)^2}{2\sigma_e^2}\right]. \end{aligned} \tag{5.80}$$

Combining the two quadratic forms as before

$$n_{i2}\left(s_i-\overline{y}_{i2}\right)^2+\left(n_{i1}+\frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i-\overline{s}_{i1}\right)^2$$
$$=\left(n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i-\overline{\overline{s}}_i\right)^2+\frac{n_{i2}\left(n_{i1}+\frac{\sigma_e^2}{\sigma_s^2}\right)}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}\left(\overline{y}_{i2}-\overline{s}_{i1}\right)^2$$

where

$$\overline{\overline{s}}_i=\frac{n_{i2}\overline{y}_{i2}+\left(n_{i1}+\frac{\sigma_e^2}{\sigma_s^2}\right)\overline{s}_{i1}}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}$$
$$=\frac{n_{i2}\overline{y}_{i2}+\left(n_{i1}+\frac{\sigma_e^2}{\sigma_s^2}\right)\left(\mu+\frac{n_{i1}}{n_{i1}+\frac{\sigma_e^2}{\sigma_s^2}}\left(\overline{y}_{i1}-\mu\right)\right)}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}$$
$$=\frac{n_{i1}\overline{y}_{i1}+n_{i2}\overline{y}_{i2}+\frac{\sigma_e^2}{\sigma_s^2}\mu}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}$$
$$=\frac{n_{i1}\overline{y}_{i1}+n_{i2}\overline{y}_{i2}}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}+\left[1-\frac{n_{i1}+n_{i2}}{n_{i1}+n_{i2}+\frac{\sigma_e^2}{\sigma_s^2}}\right]\mu$$
$$=\mu+\frac{n_i}{n_i+\frac{\sigma_e^2}{\sigma_s^2}}\left(\overline{y}_i-\mu\right)=\overline{\overline{s}}_i,$$

which is identical to (5.76). Use of this in (5.80), after retaining only the part that involves s_i, and noting that $n_{i1}+n_{i2}=n_i$, gives

$$p\left(s_i|\overline{y}_{i1},\overline{y}_{i2},\mu,\sigma_s^2,\sigma_e^2\right)\propto\exp\left[-\frac{\left(n_i+\frac{\sigma_e^2}{\sigma_s^2}\right)\left(s_i-\overline{\overline{s}}_i\right)^2}{2\sigma_e^2}\right].$$

Thus, the posterior density is identical to (5.77), illustrating that Bayes theorem has "memory", and that inferences can be updated sequentially. As a side point, note that there can be two alternative scenarios in this hypothetical scheme. In scenario A, say, inferences are done at the end, and all one needs for constructing the posterior (given the prior information) is n_i and $\overline{y}_i$, without knowledge of the averages of each of the two crops of daughters. In the sequential updating setting (scenario B), one needs $\overline{y}_{i1}$, $\overline{y}_{i2}$, n_{i1}, and n_{i2}. Hence, scenario B requires knowing the progeny group sizes and the averages at each crop, i.e., more information about the data collection process is needed. At any rate, the Bayesian analysis leads to the same inferences. This is related to what are called "stopping rules" in

sequential experimental design (O'Hagan, 1994). Suppose an experiment is designed such that it terminates after collecting n observations. Then, the Bayesian analysis infers the parameters using the n observations, and inferences are the same irrespective of whether the experiment has been designed sequentially or not. On the other hand, classical methods give different inferences for the same data collected either in a sequential or nonsequential manner. For example, if a hypothesis is tested, the sequential experiment gives a lower degree of significance than the nonsequential one (O'Hagan, 1994). ■

Example 5.9 *Updating additive genetic effects*

The setting is similar to that of the preceding example. Suppose that at stage 1 (2), measurements $\mathbf{y}_1$ ($\mathbf{y}_2$) are taken on n_1 (n_2) different individuals (so that an individual measured at any stage is not recorded at the other stage), and that the objective is to infer their additive genetic effects $\mathbf{a}_1$ ($\mathbf{a}_2$). Suppose the following linear model holds

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_1\mu_1 \\ \mathbf{1}_2\mu_2 \end{bmatrix} + \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \tag{5.81}$$

where μ_1 and μ_2 are known location parameters common to records collected in stages 1 and 2, respectively, and

$$\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \Big| \sigma_e^2 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix} \sigma_e^2 \right)$$

is a vector of independently distributed residual effects, where σ_e^2 is the (known) residual variance; $\mathbf{I}_i$ is an identity matrix of order $n_i \times n_i$, $(i = 1, 2)$. The conditional distribution of the observations, given the additive genetic effects, is

$$\begin{aligned} &p\left(\mathbf{y}_1, \mathbf{y}_2 | \mu_1, \mu_2, \mathbf{a}_1, \mathbf{a}_2, \sigma_e^2\right) \\ &\propto \exp\left[-\frac{\sum_{i=1}^{2} \left(\mathbf{y}_i - \mathbf{1}_i\mu_i - \mathbf{a}_i\right)' \left(\mathbf{y}_i - \mathbf{1}_i\mu_i - \mathbf{a}_i\right)}{2\sigma_e^2} \right]. \end{aligned} \tag{5.82}$$

In the classical infinitesimal model of inheritance, the additive genetic effects are assumed to follow the multivariate normal distribution (acting as a prior in the Bayesian sense):

$$\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} \Big| \sigma_a^2 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \sigma_a^2 \right) \tag{5.83}$$

where σ_a^2 is the additive genetic variance in the population (also assumed known), and

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

is the matrix of additive genetic relationships between individuals, or twice the matrix of coefficients of coancestry. This matrix is assumed to have full rank, that is, clones or identical twins are not encountered. For example, $\mathbf{A}_{12}$ contains the additive genetic relationships between individuals measured at stages 1 and 2. Recall that

$$\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\sigma_a^2 = \mathbf{A}_{1.2}\sigma_a^2 \tag{5.84}$$

is the covariance matrix of the conditional distribution of $\mathbf{a}_1$ given $\mathbf{a}_2$, and that

$$\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)\sigma_a^2 = \mathbf{A}_{2.1}\sigma_a^2 \tag{5.85}$$

is the covariance matrix of the conditional distribution of $\mathbf{a}_2$ given $\mathbf{a}_1$. Further, $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}$ is the multivariate regression of additive genetic effects of individuals measured in stage 2 on additive genetic effects of individuals measured in stage 1. In fact, one can write

$$\begin{aligned}\mathbf{a}_2 &= E\left(\mathbf{a}_2|\mathbf{a}_1\right) + \boldsymbol{\epsilon} \\ &= E\left(\mathbf{a}_2\right) + \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\left[\mathbf{a}_1 - E\left(\mathbf{a}_1\right)\right] + \boldsymbol{\epsilon} \\ &= \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{a}_1 + \boldsymbol{\epsilon},\end{aligned} \tag{5.86}$$

where $\boldsymbol{\epsilon}$ is a residual distributed independently of $\mathbf{a}_1$, and having the distribution

$$\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \mathbf{A}_{2.1}\sigma_a^2\right). \tag{5.87}$$

Further, using properties of inverses of partitioned matrices (Searle, 1971), let

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix},$$

where

$$\begin{aligned}\mathbf{A}^{11} &= \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} = \mathbf{A}_{1.2}^{-1}, \\ \mathbf{A}^{12} &= -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}, \\ \mathbf{A}^{21} &= -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11},\end{aligned}$$

and

$$\mathbf{A}^{22} = \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{22}^{-1}.$$

Suppose that data collection at stage 2 has been completed, and that we proceed to infer the breeding values of all individuals. The posterior density of the additive genetic effects, in view of (5.82) and (5.83), can be written as

$$\begin{aligned}&p\left(\mathbf{a}_1, \mathbf{a}_2|\mathbf{y}_1, \mathbf{y}_2, \mu_1, \mu_2, \sigma_e^2, \sigma_a^2\right) \\ &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{a}-\mathbf{w}\right)'\left(\mathbf{a}-\mathbf{w}\right) + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right]\right\},\end{aligned} \tag{5.88}$$

where

$$\mathbf{w} = \begin{bmatrix} \mathbf{y}_1 - \mathbf{1}_1\mu_1 \\ \mathbf{y}_2 - \mathbf{1}_2\mu_2 \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}.$$

Combining the two quadratic forms in (5.88), by means of (5.56) to (5.57),

$$\begin{aligned} &(\mathbf{a}-\mathbf{w})'(\mathbf{a}-\mathbf{w}) + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \\ &= (\mathbf{a}-\widehat{\mathbf{a}})'\left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)(\mathbf{a}-\widehat{\mathbf{a}}) + \mathbf{w}'\left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\mathbf{w}, \end{aligned}$$

where

$$\widehat{\mathbf{a}} = \left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{w} = \left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\begin{bmatrix} \mathbf{y}_1 - \mathbf{1}_1\mu_1 \\ \mathbf{y}_2 - \mathbf{1}_2\mu_2 \end{bmatrix}. \tag{5.89}$$

Using this in (5.88) and retaining only the part that varies with $\mathbf{a}$ gives as posterior density of the additive genetic effects,

$$\begin{aligned} &p\left(\mathbf{a}_1, \mathbf{a}_2 | \mathbf{y}_1, \mathbf{y}_2, \mu_1, \mu_2, \sigma_e^2, \sigma_a^2\right) \\ &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\mathbf{a}-\widehat{\mathbf{a}})'\left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)(\mathbf{a}-\widehat{\mathbf{a}})\right]\right\}. \end{aligned} \tag{5.90}$$

Thus, the joint posterior of all additive genetic effects is normal with mean vector $\widehat{\mathbf{a}}$ and variance–covariance matrix

$$Var\left(\mathbf{a}_1, \mathbf{a}_2 | \mathbf{y}_1, \mathbf{y}_2, \mu_1, \mu_2, \sigma_e^2, \sigma_a^2\right) = \left(\mathbf{I}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2. \tag{5.91}$$

Using (5.89) to (5.91), the density of the "first-stage" distribution

$$\left[\mathbf{a}_1 | \mathbf{y}_1, \mu_1, \sigma_e^2, \sigma_a^2\right]$$

is immediately found to be

$$\begin{aligned} &p\left(\mathbf{a}_1 | \mathbf{y}_1, \mu_1, \sigma_e^2, \sigma_a^2\right) \\ &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\mathbf{a}_1-\widetilde{\mathbf{a}}_1)'\left(\mathbf{I}_1+\mathbf{A}_{11}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)(\mathbf{a}_1-\widetilde{\mathbf{a}}_1)\right]\right\}. \end{aligned} \tag{5.92}$$

The posterior mean at stage 1 is then

$$\widetilde{\mathbf{a}}_1 = \left(\mathbf{I}_1+\mathbf{A}_{11}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}(\mathbf{y}_1 - \mathbf{1}_1\mu_1) = \left(\mathbf{I}_1+\mathbf{A}_{11}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{w}_1 \tag{5.93}$$

and the posterior covariance is

$$Var\left(\mathbf{a}_1 | \mathbf{y}_1, \mu_1, \sigma_e^2, \sigma_a^2\right) = \left(\mathbf{I}_1+\mathbf{A}_{11}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2 = \widetilde{\mathbf{C}}_1. \tag{5.94}$$

What can be said about all additive genetic effects at stage 1? The joint posterior at stage 1 is

$$p\left(\mathbf{a}_1,\mathbf{a}_2|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)\propto\left[p\left(\mathbf{y}_1|\mu_1,\sigma_e^2\right)p\left(\mathbf{a}_1|\sigma_a^2\right)\right]p\left(\mathbf{a}_2|\mathbf{a}_1,\sigma_a^2\right).$$

Noting that the expression in brackets is the posterior after stage 1, one can write

$$p\left(\mathbf{a}_1,\mathbf{a}_2|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)\propto p\left(\mathbf{a}_2|\mathbf{a}_1,\sigma_a^2\right)p\left(\mathbf{a}_1|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)$$

and this is the density of a normal process because the two intervening densities are in normal forms. Hence, the marginal distribution of $\mathbf{a}_2$ at stage 1 is normal as well, with marginal density

$$p\left(\mathbf{a}_2|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)=\int p\left(\mathbf{a}_2|\mathbf{a}_1,\sigma_a^2\right)p\left(\mathbf{a}_1|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)d\mathbf{a}_1.$$

The representation above indicates that the mean of the posterior distribution of $\mathbf{a}_2$ at stage 1 can be found to be, making use of (5.86),

$$\begin{aligned}\widetilde{\mathbf{a}}_2=E\left(\mathbf{a}_2|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)&=E_{\mathbf{a}_1|\mathbf{y}_1}\left[E\left(\mathbf{a}_2|\mathbf{a}_1\right)\right]\\&=E_{\mathbf{a}_1|\mathbf{y}_1}\left(\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{a}_1\right)\\&=\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\widetilde{\mathbf{a}}_1\\&=\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\left(\mathbf{I}_1+\mathbf{A}_{11}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{w}_1.\end{aligned}\tag{5.95}$$

Note that this has the form of $E\left(\mathbf{a}_2|\mathbf{a}_1\right)$, but with $\mathbf{a}_1$ replaced by its posterior expectation. Likewise,

$$\begin{aligned}\widetilde{\mathbf{C}}_2&=Var\left(\mathbf{a}_2|\mathbf{y}_1,\mu_1,\sigma_e^2,\sigma_a^2\right)\\&=E_{\mathbf{a}_1|\mathbf{y}_1}\left[Var\left(\mathbf{a}_2|\mathbf{a}_1\right)\right]+Var_{\mathbf{a}_1|\mathbf{y}_1}\left[E\left(\mathbf{a}_2|\mathbf{a}_1\right)\right]\\&=E_{\mathbf{a}_1|\mathbf{y}_1}\left[\mathbf{A}_{22}\sigma_a^2-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\sigma_a^2\right]+Var_{\mathbf{a}_1|\mathbf{y}_1}\left(\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{a}_1\right)\\&=\mathbf{A}_{22}\sigma_a^2-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\sigma_a^2+\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\widetilde{\mathbf{C}}_1\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.\end{aligned}\tag{5.96}$$

The first term represents the variance of $\mathbf{a}_2$ before observing anything; the second term is the reduction in variance that would be obtained if $\mathbf{a}_1$ were known, and the third term is a penalty that results from having to infer $\mathbf{a}_1$ from $\mathbf{y}_1$. Thus, the prior distribution for $\mathbf{a}_2$ to be used at stage 2 is a normal process with mean vector (5.95) and covariance matrix (5.96). Finally, at stage 2, the posterior density of $\mathbf{a}_2$ is

$$\begin{aligned}p\left(\mathbf{a}_2|\mathbf{y}_1,\mathbf{y}_2,\mu_1,\mu_2,\sigma_e^2,\sigma_a^2\right)&\propto p\left(\mathbf{y}_2|\mu_2,\mathbf{a}_2,\sigma_e^2\right)p\left(\mathbf{a}_2|\widetilde{\mathbf{a}}_2,\widetilde{\mathbf{C}}_2\right)\\&\propto\exp\left[-\frac{\left(\mathbf{y}_2-\mathbf{1}_2\mu_2-\mathbf{a}_2\right)'\left(\mathbf{y}_2-\mathbf{1}_2\mu_2-\mathbf{a}_2\right)}{2\sigma_e^2}\right.\\&\quad\left.-\frac{\left(\mathbf{a}_2-\widetilde{\mathbf{a}}_2\right)'\widetilde{\mathbf{C}}_2^{-1}\sigma_e^2\left(\mathbf{a}_2-\widetilde{\mathbf{a}}_2\right)}{2\sigma_e^2}\right].\end{aligned}\tag{5.97}$$

We know that the density is in a normal form, so the quadratics on $\mathbf{a}_2$ can be combined in the usual manner, to arrive at the mean vector and covariance matrix of the distribution. Alternatively, noting that a normal distribution is unimodal (so the mean is identical to the mode), the posterior mean at stage 2 can be found by maximizing the logarithm of (5.97). Let

$$F(\mathbf{a}_2) = -\left[\frac{(\mathbf{y}_2 - \mathbf{1}_2\mu_2 - \mathbf{a}_2)'(\mathbf{y}_2 - \mathbf{1}_2\mu_2 - \mathbf{a}_2)}{2\sigma_e^2} + \frac{(\mathbf{a}_2 - \widetilde{\mathbf{a}}_2)'\widetilde{\mathbf{C}}_2^{-1}\sigma_e^2(\mathbf{a}_2 - \widetilde{\mathbf{a}}_2)}{2\sigma_e^2}\right]$$

so

$$\frac{\partial F(\mathbf{a}_2)}{\partial \mathbf{a}_2} = (-1)\frac{-2(\mathbf{y}_2 - \mathbf{1}_2\mu_2 - \mathbf{a}_2) + 2\widetilde{\mathbf{C}}_2^{-1}\sigma_e^2(\mathbf{a}_2 - \widetilde{\mathbf{a}}_2)}{2\sigma_e^2}.$$

Setting to $\mathbf{0}$ and solving for $\mathbf{a}_2$ yields

$$\widehat{\mathbf{a}}_2 = \left(\mathbf{I}_2 + \widetilde{\mathbf{C}}_2^{-1}\sigma_e^2\right)^{-1}\left(\mathbf{y}_2 - \mathbf{1}_2\mu_2 + \widetilde{\mathbf{C}}_2^{-1}\sigma_e^2\widetilde{\mathbf{a}}_2\right) \tag{5.98}$$

as mean of the posterior distribution of $\mathbf{a}_2$, after stages 1 and 2. This is a matrix weighted average of $\widetilde{\mathbf{a}}_2$ and of $\mathbf{y}_2 - \mathbf{1}_2\mu_2$. The variance–covariance matrix of the distribution is given by

$$Var\left(\mathbf{a}_2|\mathbf{y}_1, \mathbf{y}_2, \mu_1, \mu_2, \sigma_e^2, \sigma_a^2\right) = \left(\mathbf{I}_2 + \widetilde{\mathbf{C}}_2^{-1}\sigma_e^2\right)^{-1}\sigma_e^2. \tag{5.99}$$

It can be verified that this is equal to the inverse of minus the matrix of second derivatives of $F(\mathbf{a}_2)$ with respect to $\mathbf{a}_2$. One can also verify that (5.98) is identical to the $\mathbf{a}_2$-component of the solution to (5.89). This requires very tedious algebra, so it is not shown here. ■

5.6 Features of Posterior Distributions

The marginal posterior distribution gives an exact, complete description of the state of knowledge about an unknown, after having observed the data. This unknown can be a parameter, a hypothesis, a model, or a yet to be observed data point, and can be unidimensional or multidimensional, depending on the inferential objectives. In principle, for reporting purposes, one could present the posterior distribution, and then make a commented tour of it, highlighting zones of relatively high density or probability, pointing out the existence of any multimodality, and indicating areas where the true value of the parameter may be located. The posterior holds for samples of any size, so this gives a complete solution to the problem of finite sample size inference. In a Bayesian report, however, due to space considerations, all posterior distributions of interest cannot be presented. Instead,

these are typically replaced by selected posterior summaries. Clearly, when condensing all the information in the posterior distribution into a couple of posterior summaries, some information about the form of the posterior is lost. There are exceptions; for example, if the posterior distribution is normal, then a report of the mean and variance suffices for characterizing the posterior process in full. Here we present some of the most widely used posterior summaries, and discuss their justification from a decision-theoretic point of view.

5.6.1 Posterior Probabilities

A natural summary is provided by a set of probabilities that the true parameter falls in some regions of interest. If $\boldsymbol{\Theta}$ is the parameter space, the probability that $\boldsymbol{\theta}$ falls in some region $\Re$ of $\boldsymbol{\Theta}$ is

$$\Pr\left(\boldsymbol{\theta} \in \Re|\mathbf{y}\right) = \int_{\Re} p\left(\boldsymbol{\theta}|\mathbf{y}\right) d\boldsymbol{\theta}. \tag{5.100}$$

Often, interest centers on just some of the elements of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$, with the remaining elements ($\boldsymbol{\theta}_2$) acting as nuisance parameters. In this case, the required probability is

$$\begin{aligned}\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}\right) &= \int_{\Re_1}\int_{\boldsymbol{\Theta}_2} p\left(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta} \\ &= \int_{\Re_1} p\left(\boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta}_1,\end{aligned} \tag{5.101}$$

where $\boldsymbol{\Theta}_2$ is the parameter space of $\boldsymbol{\theta}_2$. An alternative expression is

$$\begin{aligned}\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}\right) &= \int_{\Re_1}\int_{\boldsymbol{\Theta}_2} p\left(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta} \\ &= \int_{\Re_1}\int_{\boldsymbol{\Theta}_2} p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}\right) p\left(\boldsymbol{\theta}_2|\mathbf{y}\right) d\boldsymbol{\theta}.\end{aligned} \tag{5.102}$$

Reversing the order of integration one can write

$$\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}\right) = \int_{\boldsymbol{\Theta}_2}\left[\int_{\Re_1} p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y}\right) d\boldsymbol{\theta}_1\right] p\left(\boldsymbol{\theta}_2|\mathbf{y}\right) d\boldsymbol{\theta}_2. \tag{5.103}$$

The term in brackets is the conditional probability that $\boldsymbol{\theta}_1 \in \Re_1$, given $\boldsymbol{\theta}_2$ and $\mathbf{y}$. Hence, the marginal probability can be expressed as

$$\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}\right) = E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\boldsymbol{\theta}_2, \mathbf{y}\right)\right]. \tag{5.104}$$

It follows that the posterior probability that $\boldsymbol{\theta}_1 \in \Re_1$ is the weighted average of the corresponding probabilities at each possible value of the nuisance parameter $\boldsymbol{\theta}_2$, with the weight function being the marginal posterior

density $p(\boldsymbol{\theta}_2|\mathbf{y})$. Expression (5.104) can be useful in connection with the estimation of probabilities by Monte Carlo methods. Briefly, suppose that m samples are drawn from the posterior distribution of the nuisance parameter $[\boldsymbol{\theta}_2|\mathbf{y}]$, and that these samples are $\boldsymbol{\theta}_2^{(1)}, \boldsymbol{\theta}_2^{(1)}, ..., \boldsymbol{\theta}_2^{(m)}$. A consistent estimator of the posterior probability that $\boldsymbol{\theta}_1 \in \Re_1$ is given by

$$\widehat{\Pr}(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} \Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\boldsymbol{\theta}_2^{(i)}, \mathbf{y}\right). \tag{5.105}$$

This sort of calculation can be useful when analytical integration over $\boldsymbol{\theta}_2$ is not feasible or very difficult. However, requirements include:
(a) it must be relatively easy to sample from $[\boldsymbol{\theta}_2|\mathbf{y}]$, and
(b) $\Pr\left(\boldsymbol{\theta}_1 \in \Re_1|\boldsymbol{\theta}_2^{(i)}, \mathbf{y}\right)$ must be available in closed form, so that this conditional probability can be evaluated at each draw $\boldsymbol{\theta}_2^{(i)}$.

An alternative consistent estimator of the integral (5.101), which is simpler to compute, is

$$\widehat{\Pr}(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} I\left(\boldsymbol{\theta}_1^{(i)} \in \Re_1\right),$$

where $I(\cdot)$ is the indicator function and $\boldsymbol{\theta}_1^{(i)}$ are Monte Carlo draws from $[\boldsymbol{\theta}_1|\mathbf{y}]$. Here one must be able to draw samples from $[\boldsymbol{\theta}_1|\mathbf{y}]$.

Example 5.10 *Posterior probability that the breeding value of an individual exceeds a certain threshold*

A sample of n unrelated individuals is drawn from a population of ovines raised to produce cashmere. The problem is to compute the probability that the breeding value (additive genetic effect) of a particular individual is larger or smaller than a certain quantity. This is similar to the setting in Example 5.3. Suppose that a tenable model (although very naive in practice) for describing cashmere fiber diameter measured in individual i is

$$y_i = a_i + e_i,$$

where a_i is the breeding value of i for fiber diameter and e_i is an environmental effect. It is known that a_i has the normal distribution $N\left(0, \sigma_a^2\right)$, where σ_a^2 is unknown, and that it is statistically independent of $e_i \sim N\left(0, \sigma_e^2\right)$, where σ_e^2 is also unknown. Assume that the "total" variance, $\sigma_a^2 + \sigma_e^2$, is known without error, and has been estimated from the variability between individual measurements in a large collection of individuals. This being the case, one can rescale the observations by dividing by $\sqrt{\sigma_a^2 + \sigma_e^2}$, to obtain

$$\begin{aligned} y_i^* &= a_i^* + e_i^* \\ &= \frac{a_i}{\sqrt{\sigma_a^2 + \sigma_e^2}} + \frac{e_i}{\sqrt{\sigma_a^2 + \sigma_e^2}}. \end{aligned}$$

Thus, $a_i^* \sim N\left(0, h^2\right)$ and $e_i^* \sim N\left(0, 1-h^2\right)$, where

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

is the unknown heritability of cashmere fiber diameter, a parameter ranging between 0 and 1. Hence, there is a single dispersion parameter, h^2. The joint posterior density of all unknowns is then

$$p\left(a_1^*, a_2^*, \ldots, a_n^*, h^2 | \mathbf{y}^*\right) \propto p\left(h^2\right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\left(1-h^2\right)}} \exp\left[-\frac{\left(y_i^* - a_i^*\right)^2}{2\left(1-h^2\right)}\right], \tag{5.106}$$

where $p\left(h^2\right)$ is the prior density of heritability; let the prior distribution of this parameter be uniform between 0 and 1. Combining the two quadratics in a_i^*:

$$\begin{aligned}\frac{\left(a_i^* - y_i^*\right)^2}{\left(1-h^2\right)} + \frac{a_i^{*2}}{h^2} &= \left(\frac{1}{1-h^2} + \frac{1}{h^2}\right)\left(a_i^* - \overline{a}_i^*\right)^2 + \frac{\left(\frac{1}{1-h^2} + \frac{1}{h^2}\right)^{-1}}{\left(1-h^2\right)h^2} y_i^{*2} \\ &= \frac{1}{\left(1-h^2\right)h^2}\left(a_i^* - \overline{a}_i^*\right)^2 + y_i^{*2}, \end{aligned} \tag{5.107}$$

where

$$\overline{a}_i^* = h^2 y_i^{*2}.$$

Using (5.107) in (5.106), the joint posterior is then

$$\begin{aligned} p\left(a_1^*, a_2^*, \ldots, a_n^*, h^2 | \mathbf{y}^*\right) \propto &\left[\frac{1}{\left(1-h^2\right)h^2}\right]^{\frac{n}{2}} \prod_{i=1}^{n} \exp\left[-\tfrac{1}{2}\frac{\left(a_i^* - \overline{a}_i^*\right)^2}{\left(1-h^2\right)h^2}\right] \\ &\times \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n} y_i^{*2}\right) p\left(h^2\right). \end{aligned} \tag{5.108}$$

This indicates that, given h^2, breeding values have independent normal distributions with means $\overline{a}_i^*$ and common variance $\left(1-h^2\right)h^2$. Integrating now over the n breeding values gives, as marginal posterior density of heritability,

$$\begin{aligned} p\left(h^2 | \mathbf{y}^*\right) &\propto \left[\left(1-h^2\right)h^2\right]^{-\frac{n}{2}} p\left(h^2\right) \prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp\left[-\frac{\left(a_i^* - \overline{a}_i^*\right)^2}{2\left(1-h^2\right)h^2}\right] da_i^* \\ &\propto \left[\left(1-h^2\right)h^2\right]^{-\frac{n}{2}} p\left(h^2\right)\left[\sqrt{\left(1-h^2\right)h^2}\right]^n \propto p\left(h^2\right). \end{aligned} \tag{5.109}$$

Hence, the Bayesian analysis yields the prior as posterior distribution of heritability. This is because the data do not contribute information about

the partition of the variance into an additive genetic and an environmental component, even if the total or phenotypic variance is known. The situation would be different if some of the individuals were genetically related, but this is beyond the scope of the example. It follows that the posterior distribution of the nuisance parameter (heritability in this case) is precisely the prior distribution of h^2. Now, to infer features of the posterior distribution of the breeding value of individual i while taking into account uncertainty about the nuisance parameter h^2, one must average over its posterior distribution. Recall that our objective is to calculate the posterior probability that the breeding value is larger than a certain value (0, say). Using (5.103), and taking into account that the posterior (or prior, in this case) density of heritability is uniform

$$\begin{aligned}\Pr\left(a_i^* > 0|\mathbf{y}^*\right) &= \int_0^1 \left[\int_0^\infty p\left(a_i^*|h^2,\mathbf{y}^*\right) da_i^*\right] p\left(h^2\right) dh^2 \\ &= \int_0^1 \left[1 - \Pr\left(a_i^* \leq 0|h^2,\mathbf{y}^*\right)\right] p\left(h^2\right) dh^2 \\ &= 1 - \int_0^1 \Pr\left[\frac{a_i^* - \overline{a}_i^*}{\sqrt{\left(1-h^2\right)h^2}} \leq \frac{-\overline{a}_i^*}{\sqrt{\left(1-h^2\right)h^2}}|h^2,\mathbf{y}^*\right] p\left(h^2\right) dh^2 \\ &= 1 - \int_0^1 \Phi\left(\frac{-\overline{a}_i^*}{\sqrt{\left(1-h^2\right)h^2}}\right) p\left(h^2\right) dh^2. \end{aligned} \tag{5.110}$$

The integral in (5.110) cannot be expressed in closed form, since the integrand is an integral itself. However, it can be evaluated by Monte Carlo methods, as follows:
(1) Draw m independent values of h^2 from a $U\left(0,1\right)$ distribution; let each draw be $h^{2(j)}$.
(2) Compute, for each draw, $\overline{a}_i^{*(j)} = h^{2(j)} y_i^{*2}$.
(3) For each draw, calculate

$$\Phi\left(\frac{-\overline{a}_i^{*(j)}}{\sqrt{\left[1-h^{2(j)}\right]h^{2(j)}}}\right).$$

(4) Form the consistent estimator of the desired probability

$$\widehat{\Pr}\left(a_i^* > 0|\mathbf{y}^*\right) = 1 - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(\frac{-\overline{a}_i^{*(j)}}{\sqrt{\left[1-h^{2(j)}\right]h^{2(j)}}}\right). \tag{5.111}$$

■

5.6.2 Posterior Quantiles

Consider the scalar posterior distribution $[\theta|\mathbf{y}]$ defined in $\theta_L \leq \theta \leq \theta_U$, where θ_L and θ_U are the minimum and maximum values, respectively, that θ can take. The α-quantile of the posterior distribution is the value q satisfying the equation

$$\Pr(\theta \leq q|\mathbf{y}) = \alpha.$$

A quantile commonly used to characterize location of the posterior distribution is the posterior median, m, such that

$$\Pr(\theta \leq q|\mathbf{y}) = \Pr(\theta > q|\mathbf{y}) = \frac{1}{2}$$

or, in the continuous case,

$$\int_{-\infty}^{m} p(\theta|\mathbf{y})\, d\theta = \int_{m}^{\infty} p(\theta|\mathbf{y})\, d\theta.$$

Quantiles arise in the construction of high-credibility sets, that is, sets of θ values that contain the true parametric value at high probability. For example, a credibility set of size $1-\alpha$ is given by all possible values between q_1 and q_2 such that

$$\Pr(q_1 \leq \theta \leq q_2|\mathbf{y}) = 1 - \alpha \tag{5.112}$$

$$\Pr(\theta \leq q_1|\mathbf{y}) = \frac{\alpha}{2},$$

$$\Pr(\theta > q_2|\mathbf{y}) = \frac{\alpha}{2}.$$

If $\alpha = 0.05$, say, then q_1 is the 2.5% quantile of the posterior distribution and q_2 is the 97.5% quantile. In principle, there can be many, perhaps infinite, credibility regions of size $1-\alpha$. However, if the probability statement (5.112) is such that all values within the interval are required to have higher posterior density than values outside it, then (q_1, q_2) is called a $1-\alpha$ highest posterior density interval, or HPD for short (Box and Tiao, 1973). If such a region exists, then an HPD region of size $1-\alpha$ defines the boundaries unambiguously.

Using the median of a posterior distribution as a "point estimate" of θ has a justification on decision-theoretic grounds. To illustrate, let the parameter vector be $\boldsymbol{\theta} = \left[\theta_1, \boldsymbol{\theta}_2'\right]'$, where θ_1 varies over the real line, and $\boldsymbol{\theta}_2 \in \Re_{\boldsymbol{\theta}_2}$ are nuisance parameters. Define $L\left(\widehat{\theta}_1, \theta_1\right)$ to be a loss function, with minimum value 0 when $L\left(\widehat{\theta}_1 = \theta_1, \theta_1\right)$. Here, $\widehat{\theta}_1$ is a function involving the data and possibly the hyperparameters. Now let the loss function have the form

$$L\left(\widehat{\theta}_1, \theta_1\right) = a\left|\widehat{\theta}_1 - \theta_1\right| \tag{5.113}$$

where a is a scalar constant. This implies that the loss is proportional to the absolute "error of estimation". The expected posterior loss is

$$E\left[L\left(\widehat{\theta}_1,\theta_1\right)|\mathbf{y}\right] = \int_{-\infty}^{\infty}\int_{\Re_{\boldsymbol{\theta}_2}} a\left|\widehat{\theta}_1-\theta_1\right| p\left(\theta_1,\boldsymbol{\theta}_2|\mathbf{y}\right) d\theta_1 d\boldsymbol{\theta}_2$$

$$= \int_{-\infty}^{\infty} a\left|\widehat{\theta}_1-\theta_1\right| \left[\int_{\Re_{\boldsymbol{\theta}_2}} p\left(\theta_1,\boldsymbol{\theta}_2|\mathbf{y}\right) d\boldsymbol{\theta}_2\right] d\theta_1$$

$$= \int_{-\infty}^{\infty} a\left|\widehat{\theta}_1-\theta_1\right| p\left(\theta_1|\mathbf{y}\right) d\theta_1 \tag{5.114}$$

since integration of the joint density over $\boldsymbol{\theta}_2$ gives the marginal posterior of θ_1. Now

$$\left|\widehat{\theta}_1-\theta_1\right| = \begin{cases} \widehat{\theta}_1-\theta_1, & \text{for } \theta_1 \le \widehat{\theta}_1, \\ \theta_1-\widehat{\theta}_1, & \text{for } \widehat{\theta}_1 < \theta_1. \end{cases}$$

Hence, following Zellner (1971),

$$E\left[L\left(\widehat{\theta}_1,\theta_1\right)|\mathbf{y}\right]$$

$$= a\left[\int_{-\infty}^{\widehat{\theta}_1}\left(\widehat{\theta}_1-\theta_1\right) p\left(\theta_1|\mathbf{y}\right) d\theta_1 + \int_{\widehat{\theta}_1}^{\infty}\left(\theta_1-\widehat{\theta}_1\right) p\left(\theta_1|\mathbf{y}\right) d\theta_1\right]$$

$$\propto \left\{\widehat{\theta}_1 \Pr\left(\theta_1 \le \widehat{\theta}_1|\mathbf{y}\right) - \int_{-\infty}^{\widehat{\theta}_1} \theta_1 p\left(\theta_1|\mathbf{y}\right) d\theta_1 \right.$$

$$\left. + \int_{\widehat{\theta}_1}^{\infty} \theta_1 p\left(\theta_1|\mathbf{y}\right) d\theta_1 - \widehat{\theta}_1\left[1-\Pr\left(\theta_1 \le \widehat{\theta}_1|\mathbf{y}\right)\right]\right\}. \tag{5.115}$$

We seek now the $\widehat{\theta}_1$ minimizing the expected posterior loss. Differentiating (5.115) with respect to $\widehat{\theta}_1$, recalling that

$$\frac{\partial\left[\int_{-\infty}^{\widehat{\theta}_1} \theta_1 p\left(\theta_1|\mathbf{y}\right) d\theta_1\right]}{\partial\widehat{\theta}_1} = \widehat{\theta}_1 p\left(\widehat{\theta}_1|\mathbf{y}\right)$$

and setting the derivative to 0, yields, after some terms cancel out

$$\frac{\partial E\left[L\left(\widehat{\theta}_1,\theta_1\right)|\mathbf{y}\right]}{\partial\widehat{\theta}_1} = 2\Pr\left(\theta_1 \le \widehat{\theta}_1|\mathbf{y}\right) - 1 = 0.$$

This gives the equation

$$\Pr\left(\theta_1 \leq \widehat{\theta}_1 | \mathbf{y}\right) = \frac{1}{2}$$

which is satisfied by taking $\widehat{\theta}_1$ to be the median of the posterior distribution. Hence, the median is optimum in the sense of minimizing the expected absolute "error of estimation". The median is functionally invariant under one-to-one transformation. This implies that if $\widehat{\theta}_1$ is the median of the posterior distribution of θ, then $g\left(\widehat{\theta}_1\right)$ is the median of the posterior distribution of $g(\theta)$. Hence, $g\left(\widehat{\theta}_1\right)$ minimizes the expected posterior loss

$$E\left\{L\left[g\left(\widehat{\theta}_1\right), g(\theta_1)\right] | \mathbf{y}\right\} = \int_{-\infty}^{\infty} a\left|g\left(\widehat{\theta}_1\right) - g(\theta_1)\right| p(\theta_1|\mathbf{y})\, d\theta_1.$$

5.6.3 Posterior Modes

The modal vector of the posterior distribution is defined as

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} &= Arg\max_{\boldsymbol{\theta}} \left[p(\boldsymbol{\theta}|y)\right] = Arg\max_{\boldsymbol{\theta}} \left[c\, L(\boldsymbol{\theta}|y)\, g(\boldsymbol{\theta})\right] \\ &= Arg\max_{\boldsymbol{\theta}} \left\{\log\left[L(\boldsymbol{\theta}|y)\right] + \log\left[g(\boldsymbol{\theta})\right]\right\}, \end{aligned} \tag{5.116}$$

that is, as the value of the parameter having highest density (or probability in discrete situations). As noted in Section 5.3, if the prior distribution of $\boldsymbol{\theta}$ is uniform, then the posterior mode is identical to the maximum likelihood estimator; this is an incidental, and not a fundamental issue in Bayesian analysis. Since the posterior mode is interpretable as "the most likely value of the parameter", it is a natural candidate as a purveyor of information about the location of the posterior distribution.

The mode has a decision-theoretic justification as a point estimator of a parameter, at least in the single parameter situation. Suppose that the loss function has the following form (O'Hagan, 1994):

$$L\left(\widehat{\theta}_1, \theta_1\right) \begin{cases} 0, & \text{if } \left|\widehat{\theta}_1 - \theta_1\right| \leq b, \\ 1, & \text{if } \left|\widehat{\theta}_1 - \theta_1\right| > b, \end{cases}$$

for some constant b. The expected posterior loss is then

$$\begin{aligned} E\left[L\left(\widehat{\theta}_1,\theta_1\right)|\mathbf{y}\right] &= 0\times\Pr\left[\left|\widehat{\theta}_1-\theta_1\right|\leq b|\mathbf{y}\right]+1\times\Pr\left[\left|\widehat{\theta}_1-\theta_1\right|>b|\mathbf{y}\right] \\ &= \Pr\left[\left|\widehat{\theta}_1-\theta_1\right|>b|\mathbf{y}\right] \\ &= \Pr\left[\widehat{\theta}_1-\theta_1>b|\mathbf{y}\right]+\Pr\left[\theta_1-\widehat{\theta}_1>b|\mathbf{y}\right] \\ &= \int_{-\infty}^{\widehat{\theta}_1-b} p\left(\theta_1|\mathbf{y}\right)d\theta_1+\int_{\widehat{\theta}_1+b}^{\infty} p\left(\theta_1|\mathbf{y}\right)d\theta_1. \end{aligned} \tag{5.117}$$

Now, take derivatives of (5.117) with respect to $\widehat{\theta}_1$ to locate a minimum

$$\frac{\partial E\left[L\left(\widehat{\theta}_1,\theta_1\right)|\mathbf{y}\right]}{\partial\widehat{\theta}_1}=p\left(\widehat{\theta}_1-b|\mathbf{y}\right)-p\left(\widehat{\theta}_1+b|\mathbf{y}\right). \tag{5.118}$$

Setting to zero gives, as first-order condition,

$$p\left(\widehat{\theta}_1-b|\mathbf{y}\right)=p\left(\widehat{\theta}_1+b|\mathbf{y}\right).$$

If $\widehat{\theta}_1$ is a mode of the posterior distribution, it must be true that $p\left(\widehat{\theta}_1|\mathbf{y}\right)\geq p\left(\widehat{\theta}_1-b|\mathbf{y}\right)$ and $p\left(\widehat{\theta}_1|\mathbf{y}\right)\geq p\left(\widehat{\theta}_1+b|\mathbf{y}\right)$. As $b\to 0$ the condition is satisfied only by the mode. In the limit, the "optimal" $\widehat{\theta}_1$ is given by the posterior mode.

In a multiparameter situation, the modal vector has as many elements as there are parameters in the model. Here it is important to make a distinction between the components of the modal vector of the joint distribution, and the modes of the marginal distribution of one or of a set of parameters of interest, after some nuisance parameters have been integrated out. Most often, the marginal models differ from the corresponding component of the joint modal vector. Exceptions occur, such as when the posterior is multivariate normal or multivariate-t. In these cases, if one maximizes the joint distribution, each of the components of the modal vector is identical to the mode of the marginal distribution of the corresponding parameter.

With the usual notation, the mode of a joint posterior distribution is defined by the statement

$$\begin{bmatrix} \widetilde{\widetilde{\boldsymbol{\theta}}}_1 \\ \widetilde{\widetilde{\boldsymbol{\theta}}}_2 \end{bmatrix} = \underset{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}{Arg\max}\left\{\log\left(c\right)+\log\left[L\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}\right)\right]+\log\left[g\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)\right]\right\}.$$

The notation $\widetilde{\boldsymbol{\theta}}_1,\widetilde{\boldsymbol{\theta}}_2$ will be employed here to denote the marginal modes. In computing joint modes, it is useful to note that

$$\begin{aligned} \log\left[p\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}\right)\right] &= \log\left[p\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right]+\log\left[p\left(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\mathbf{y}\right)\right] \\ &= \log\left[p\left(\boldsymbol{\theta}_2|\mathbf{y}\right)\right]+\log\left[p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\mathbf{y}\right)\right]. \end{aligned}$$

In order to locate a maximum, the following system of equations must be solved, simultaneously,

$$\frac{\log\left[p\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}\right)\right]}{\partial\boldsymbol{\theta}_1}=\frac{\partial\log\left[p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\mathbf{y}\right)\right]}{\partial\boldsymbol{\theta}_1}=\mathbf{0} \tag{5.119}$$

and

$$\frac{\log\left[p\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y}\right)\right]}{\partial\boldsymbol{\theta}_2}=\frac{\partial\log\left[p\left(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\mathbf{y}\right)\right]}{\partial\boldsymbol{\theta}_2}=\mathbf{0}. \tag{5.120}$$

Typically, this will define an iterative algorithm (Lindley and Smith, 1972). Note that, in some sense, the two equations do not incorporate measures of uncertainty about nuisance parameters provided by the marginal densities of the parameters not involved in the differentiation. For example, in (5.119) one constructs a system of equations as if $\boldsymbol{\theta}_2$ were known. This suggests that if a joint mode is used as a point estimator, there may be an implicit overstatement of precision in the analysis. At least in some variance component problems (e.g., Thompson, 1980; Harville, 1977) it has been found that a component of a joint mode can lead to estimators of variance components that are identically equal to zero when used with vague priors. With multiparameter problems, the joint mode does not always take into account fully the uncertainty about other parameters in the model that may be acting as nuisances.

Example 5.11 *Joint and marginal modes in comparisons between treatment and control populations*

Independent samples of sizes n_1, n_2, n_3 are drawn at random from the three normal populations $N\left(\mu_1,\sigma^2\right)$, $N\left(\mu_2,\sigma^2\right)$, and $N\left(\mu_3,\sigma^2\right)$, respectively, where the variance is common, but unknown. One of the populations or "treatments" is a control of some sort. An attribute is measured in each of the items sampled, and the objective is to infer the differences between means $\mu_1-\mu_2$ and, possibly, $\mu_1-\mu_3$, marginally or jointly with $\mu_1-\mu_2$. Here σ^2 legitimately represents a nuisance parameter in all cases, so it is important to take into account uncertainty about dispersion in the analysis. Also, note that there are three location parameters, but the desired inferences involve either the marginal distribution of a linear combination of two means $(\mu_1-\mu_2)$, or a bivariate posterior distribution (that of $\mu_1-\mu_2$ and $\mu_1-\mu_3$). In these cases, although marginalization proceeds to different degrees, the Bayesian approach has a single (and simple) solution: transform the variables as needed, and integrate nuisance parameters to find the target distribution. In many instances, this cannot be done analytically, but sampling methods are available, as discussed later in this book. Fortunately, there is an analytical solution in our setting, which is developed below, step-by-step. The likelihood function of the four parameters

entering into the model is

$$\begin{aligned} & p\left(\mathbf{y}|\mu_1,\mu_2,\mu_3,\sigma^2\right) \\ & \propto \prod_{i=1}^{3}\prod_{j=1}^{n_i}\left(\sigma^2\right)^{-\frac{n_i}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_{ij}-\mu_i\right)^2\right] \\ & \propto \left(\sigma^2\right)^{-\frac{n_1+n_2+n_3}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_i\sum_j\left(y_{ij}-\mu_i\right)^2\right] \\ & \propto \left(\sigma^2\right)^{-\frac{n_1+n_2+n_3}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_i\left[\sum_j\left(y_{ij}-\overline{y}_i\right)^2\right.\right. \\ & \qquad \left.\left.+\,n_i\left(\mu_i-\overline{y}_i\right)^2\right]\right\}, \end{aligned} \tag{5.121}$$

where y_{ij} is observation j in treatment i, and $\overline{y}_i$ is the average value of all observations in the treatment. The maximum likelihood estimators of the parameters can be readily found to be $\overline{y}_i$ for μ_i $(i=1,2,3)$, and

$$\widehat{\sigma}^2=\frac{\sum_i\sum_j\left(y_{ij}-\overline{y}_i\right)^2}{n_1+n_2+n_3}$$

for σ^2. The finite sample distributions of the maximum likelihood estimators are

$$\begin{aligned} \overline{y}_i &\sim N\left(\mu_i,\frac{\sigma^2}{n_i}\right), \quad i=1,2,3, \\ \widehat{\sigma}^2 &\sim \frac{\sigma^2}{n_1+n_2+n_3}\chi^2_{n_1+n_2+n_3-3}. \end{aligned}$$

Put $n=n_1+n_2+n_3$. Recall that the asymptotic distribution of $\widehat{\sigma}^2$ is normal with mean σ^2 and variance $2\sigma^4/n$, which differs from the finite sample distribution given above. On the other hand, the exact and asymptotic distributions of the means are identical. We now give a Bayesian structure to the model, and adopt a bounded uniform prior between $\mu_{\min}$ and $\mu_{\max}$ for each of the three location parameters, and a scaled inverted chi-square distribution with parameters ν (degree of belief) and τ^2 ("prior value of the variance"). The joint posterior density of all parameters, omitting hyperparameters in the notation, can be written as

$$\begin{aligned} & p\left(\mu_1,\mu_2,\mu_3,\sigma^2|\mathbf{y}\right) \\ & \propto p\left(\mathbf{y}|\mu_1,\mu_2,\mu_3,\sigma^2\right)p\left(\mu_1,\mu_2,\mu_3\right)p\left(\sigma^2|\nu,\tau\right) \\ & \propto \left(\sigma^2\right)^{-\frac{n+\nu+2}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[n\widehat{\sigma}^2+\nu\tau^2+\sum_i n_i\left(\mu_i-\overline{y}_i\right)^2\right]\right\} \end{aligned} \tag{5.122}$$

defined within the boundaries given above. The mode of the joint posterior distribution is arrived at by differentiating the logarithm of (5.122) with respect to all parameters. The maximizers can be verified to be

$$\widetilde{\widetilde{\mu}}_i = \overline{y}_i, \quad \text{for } i = 1, 2, 3 \text{ and } \widetilde{\widetilde{\sigma}}^2 = \frac{n\widehat{\sigma}^2 + \nu\tau^2}{n + \nu + 2}. \tag{5.123}$$

The marginal posterior density of σ^2 is obtained by integrating (5.122) over the $\mu's$. Note that the resulting expression can be written as

$$\begin{aligned} p\left(\sigma^2|\mathbf{y}\right) &\propto \left(\sigma^2\right)^{-\frac{n+\nu+2}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(n\widehat{\sigma}^2 + \nu\tau^2\right)\right] \\ &\quad \times \prod_{i=1}^{3} \int_{\mu_{\min}}^{\mu_{\max}} \exp\left[-\frac{n_i\left(\mu_i - \overline{y}_i\right)^2}{2\sigma^2}\right] d\mu_i \\ &\propto \left(\sigma^2\right)^{-\frac{(n_1-1+n_2-1+n_3-1)+\nu+2}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(n\widehat{\sigma}^2 + \nu\tau^2\right)\right] \\ &\quad \times \prod_{i=1}^{3}\left[\Phi\left(\frac{\mu_{\max} - \overline{y}_i}{\sigma^2/n_i}\right) - \Phi\left(\frac{\mu_{\min} - \overline{y}_i}{\sigma^2/n_i}\right)\right]. \end{aligned} \tag{5.124}$$

This density is not explicit in σ^2, and finding the maximizer requires tailoring an iterative algorithm. Suppose that the upper and lower boundaries of the prior distributions of the means approach minus and plus infinity, respectively. In the limit, this would give a very vague (in fact, improper) uniform prior distribution. In this case, the difference between the normal c.d.f.'s given above goes to 1 for each of the three populations, and the marginal posterior density of σ^2 tends to

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{(n_1-1+n_2-1+n_3-1)+\nu+2}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(n\widehat{\sigma}^2 + \nu\tau^2\right)\right]. \tag{5.125}$$

This is a scaled inverted chi-square density with degree of belief parameter

$$\nu^* = n + \nu - 3,$$

and scale parameter

$$\tau^{*2} = \frac{n\widehat{\sigma}^2 + \nu\tau^2}{\nu^*}.$$

The mode of this marginal distribution is

$$\widetilde{\sigma}^2 = \frac{\nu^*\tau^{*2}}{\nu^* + 2}. \tag{5.126}$$

Comparison of (5.126) with $\widetilde{\widetilde{\sigma}}^2$ in (5.123), assuming the same vague priors for the means as in the last situation, illustrates that the mode of a marginal

distribution can differ from the corresponding modal component of a joint distribution. As a side issue, note that the process of integrating out the three unknown means from the posterior distribution results in a "loss of three degrees of freedom", reflecting the straightforward, automatic bookkeeping of information that the probability calculus makes in the Bayesian context. These three degrees of freedom are not accounted for in the joint maximization leading to the modal component given in (5.123). Assign now a scale inverted chi-square process prior to σ^2 but, instead, take $\nu = 0$ as a "prior degree of belief" value. In this scenario, the prior density degenerates to σ^{-2}, which is improper, as the integral between 0 and ∞ is not finite. However, the marginal posterior density is still proper provided that at least two observations are collected in at least one of the three populations. The improper prior σ^{-2} appears often in Bayesian analysis. In general, we recommend exercising utmost caution when improper priors are assigned to parameters, because the posterior distribution may turn out to be improper, and this is not always straightforward to check. On the other hand, proper priors ensure that the posterior process will be a proper one. Next, we proceed to find the marginal posterior distribution of the three means, after integrating σ^2 out of the joint density of all parameters, given in (5.122). When this density is viewed as a function of σ^2, it can be readily seen that it is in an inverse gamma form, whose integration constant we know (see Chapter 1). Then the desired integral is given by the reciprocal of the integration constant

$$p\left(\mu_1, \mu_2, \mu_3 | \mathbf{y}\right) \propto \left[n\widehat{\sigma}^2 + \nu\tau^2 + \sum_i n_i \left(\mu_i - \overline{y}_i\right)^2 \right]^{-\frac{n+\nu}{2}}.$$

Eliminating terms that do not depend on the means, and rearranging the exponent, yields,

$$p\left(\mu_1, \mu_2, \mu_3 | \mathbf{y}\right) \propto \left[1 + \frac{\sum_i n_i \left(\mu_i - \overline{y}_i\right)^2}{n\widehat{\sigma}^2 + \nu\tau^2} \right]^{-\frac{n-3+\nu+3}{2}}. \tag{5.127}$$

Now write

$$\begin{aligned} &\sum_i n_i \left(\mu_i - \overline{y}_i\right)^2 \\ &= \left[\left(\mu_1 - \overline{y}_1\right), \left(\mu_2 - \overline{y}_2\right), \left(\mu_3 - \overline{y}_3\right)\right] \mathbf{N} \begin{bmatrix} \mu_1 - \overline{y}_1 \\ \mu_2 - \overline{y}_2 \\ \mu_3 - \overline{y}_3 \end{bmatrix} \\ &= \left(\boldsymbol{\mu} - \overline{\mathbf{y}}\right)' \mathbf{N} \left(\boldsymbol{\mu} - \overline{\mathbf{y}}\right) \end{aligned}$$

where

$$\mathbf{N} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix},$$

and let:

$$c^2 = \frac{n\widehat{\sigma}^2 + \nu\tau^2}{n - 3 + \nu}.$$

Then (5.127) can be put as

$$p\left(\mu_1, \mu_2, \mu_3 | \mathbf{y}\right) \propto \left[1 + \frac{\left(\boldsymbol{\mu} - \overline{\mathbf{y}}\right)' \mathbf{N} \left(\boldsymbol{\mu} - \overline{\mathbf{y}}\right)}{\left(n - 3 + \nu\right) c^2}\right]^{-\frac{n-3+\nu+3}{2}}. \tag{5.128}$$

This is the density of a trivariate-t distribution with mean vector $\overline{\mathbf{y}}$, degrees of freedom $(n - 3 + \nu)$, and variance–covariance matrix

$$Var\left(\mu_1, \mu_2, \mu_3 | \mathbf{y}\right) = c^2 \mathbf{N}^{-1} \frac{\left(n - 3 + \nu\right)}{\left(n - 5 + \nu\right)}. \tag{5.129}$$

This distribution is unimodal and symmetric, and the mode is given by $\overline{\mathbf{y}}$. Hence, the mode obtained after marginalizing with respect to σ^2 is identical to the $\boldsymbol{\mu}-$component of the mode of the joint posterior distribution $\left[\mu_1, \mu_2, \mu_3, \sigma^2 | \mathbf{y}\right]$. Further, since the process is multivariate-t, it follows that all the marginal distributions are univariate t with mean vector $\overline{y}_i$, $(n - 3 + \nu)$ degrees of freedom and variance:

$$Var\left(\mu_i | \mathbf{y}\right) = \frac{c^2 \left(n - 3 + \nu\right)}{n_i \left(n - 5 + \nu\right)}.$$

It is important, however, to note that the true means μ_i are not mutually independent, even though they are not correlated (diagonal $\mathbf{N}$) a posteriori; this is because the joint density (multivariate-t) cannot be written as a product of the resulting marginal densities, which are all univariate-t. It also follows that the distribution of any linear contrast $\mu_i - \mu_{i'}$ is univariate-t, with mean $\overline{y}_i - \overline{y}_{i'}$, degrees of freedom equal to $(n - 3 + \nu)$ and posterior variance

$$Var\left(\mu_i - \mu_{i'} | \mathbf{y}\right) = \frac{c^2 \left(n - 3 + \nu\right)}{\left(n - 5 + \nu\right)} \left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right).$$

Another consequence is that the joint distribution of, for example, pairs of differences between population means is also bivariate-t, with parameters derived directly from the joint distribution having density (5.128). ■

Example 5.12 *Digression on multiple comparisons*

Consider now the problem of carrying out multiple comparisons between means (Milliken and Johnson, 1992) in the setting of Example 5.11. From a frequentist point of view, the chance of finding a difference that appears

to be "significant" increases with the number of comparisons made, even if such differences are truly null. For example, these authors note that if an experiment involves 100 "independent" tests at a significance level of $\alpha = 0.05$, one should expect (over repeated sampling) $(0.05)\,100 = 5$ such tests to be significant, just by chance. In order to adjust for the number of comparisons made, different error rates must be computed. For example, the "experimentwise error" rate is the probability of making at least one error in an experiment in which there are no real differences between means. A battery of statistical tests has been developed for multiple comparisons, where the experimentwise error rate is fixed at some level, e.g., 5%, although it is far from clear when one test ought to be preferred over another. Further, some of these tests have not been extended to accommodate unequal sample sizes or the presence of covariates in the model. From a Bayesian point of view, all this is a nonissue. First, "hypothesis testing" is approached in a completely different manner in Bayesian statistics, as it will be seen later. Second, the joint posterior distribution keeps track of any existing dependencies between parameters. Third, the probability calculus applied to the joint posterior distribution adjusts probability volumes automatically, that is, uncertainty about whatever is regarded as a nuisance in a multiple comparison is accounted for via integration. Joint inferences are then done from a distribution of appropriate dimension, and probability statements are either marginal or joint, depending on the objective of the analysis. Suppose that one is interested in inferring all possible differences between means, and that if one of such differences does not exceed a certain threshold $t_{1-\alpha}$, e.g., the $1-\alpha$ quantile of the posterior distribution of the difference, then the "hypothesis" that the difference is null is accepted. With the setting as in the preceding example, all possible differences between pairs of means can be written in matrix notation as,

$$\begin{bmatrix} \mu_{12} \\ \mu_{13} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}.$$

The third comparison is redundant because it can be obtained from the difference between the second and third comparisons. Hence, it suffices to work with the full-rank subset of comparisons

$$\overleftrightarrow{\boldsymbol{\mu}} = \begin{bmatrix} \mu_{12} \\ \mu_{13} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \mathbf{L}\boldsymbol{\mu}.$$

Since the joint posterior distribution of μ_1, μ_2, and μ_3 is trivariate-t, it follows that the posterior distribution of $\overleftrightarrow{\boldsymbol{\mu}}$ is bivariate-t (by virtue of being a linear combination), and that the marginal distributions of μ_{12} and μ_{13} are univariate-t. All these distributions have parameters that can be deduced from results in the preceding example. For instance, the posterior

covariance between μ_{12} and μ_{13} is

$$\begin{aligned} &Cov\left(\mu_{12},\mu_{13}|\mathbf{y}\right) \\ &= Var\left(\mu_1|\mathbf{y}\right) - Cov\left(\mu_1,\mu_3|\mathbf{y}\right) - Cov\left(\mu_1,\mu_2|\mathbf{y}\right) + Cov\left(\mu_2,\mu_3|\mathbf{y}\right) \\ &= Var\left(\mu_1|\mathbf{y}\right) \end{aligned}$$

since all posterior covariances between means are null, as $\mathbf{N}$ is a diagonal matrix. The posterior probability that $\mu_{12} > t_{1-\alpha}$ is

$$\begin{aligned} p_{12} &= \int_{t_{1-\alpha}}^{\infty} p\left(\mu_{12}|\mathbf{y}\right) d\mu_{12} \\ &= \frac{\int_{t_{1-\alpha}}^{\infty} \left\{1 + \frac{\left[\mu_{12} - \left(\overline{y}_1 - \overline{y}_2\right)\right]^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\left(n-3+\nu\right)c^2}\right\}^{-\frac{n-3+\nu+1}{2}} d\mu_{12}}{\int_{-\infty}^{\infty} \left\{1 + \frac{\left[\mu_{12} - \left(\overline{y}_1 - \overline{y}_2\right)\right]^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\left(n-3+\nu\right)c^2}\right\}^{-\frac{n-3+\nu+1}{2}} d\mu_{12}} \end{aligned}$$

which can be computed readily from tabled values of the standardized t distribution or from some available computer routine. Using a similar procedure one can calculate p_{13}, the posterior probability that $\mu_{13} > t_{1-\alpha}$. Next, we consider the joint distribution of μ_{12} and μ_{13}. This is a bivariate-t process with density

$$p\left(\mu_{12},\mu_{13}|\mathbf{y}\right) \propto \left[1 + \frac{\left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)' \mathbf{LNL}' \left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)}{\left(n-3+\nu\right)c^2}\right]^{-\frac{n-3+\nu+2}{2}}.$$

The posterior probability that both μ_{12} and μ_{13} exceed the threshold $t_{1-\alpha}$ is given by

$$p_{12,13} = \frac{\int_{t_{1-\alpha}}^{\infty}\int_{t_{1-\alpha}}^{\infty} \left[1 + \frac{\left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)' \mathbf{LNL}' \left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)}{\left(n-3+\nu\right)c^2}\right]^{-\frac{n-3+\nu+2}{2}} d\mu_{12}\, d\mu_{13}}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[1 + \frac{\left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)' \mathbf{LNL}' \left(\overleftrightarrow{\boldsymbol{\mu}} - \mathbf{L}\overline{\mathbf{y}}\right)}{\left(n-3+\nu\right)c^2}\right]^{-\frac{n-3+\nu+2}{2}} d\mu_{12}\, d\mu_{13}}.$$

Then the probability that at least one of the comparisons will exceed the threshold, posterior to the data, is

$$\Pr\left(\mu_{12} > t_{1-\alpha} \cup \mu_{13} > t_{1-\alpha}\right) = p_{12} + p_{13} - p_{12,13} \tag{5.130}$$

with p_{12} calculated as before, p_{13} computed with a similar expression, and $p_{12,13}$ from the expression preceding (5.130).

There is an important difference with the frequentist approach. In the Bayesian analysis, the calculations involve posterior probabilities of events. In the classical methodology of multiple comparisons, the probabilities involve sampling distributions of some estimates, with calculations conducted as if the null hypothesis were true. In the Bayesian approach, there is no such thing as a "null or alternative" hypothesis, except in some metaphoric sense. The calculations indicated above would be carried out for two models representing different states of nature. Subsequently, the posterior odds ratio, as in Example 5.2, would quantify the strength of the evidence in favor of one of the two models. It is seen that the Bayesian probability calculus enables posing and solving the problem in a conceptually straightforward manner. ■

Example 5.13 *Joint modes in a Gaussian linear model*
Let the linear model be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{5.131}$$

where $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$ follow the independent distributions:

$$\begin{aligned} \boldsymbol{\beta}|\beta_0, \sigma^2_\beta &\sim N\left(\mathbf{1}\beta_0, \mathbf{I}_b\sigma^2_\beta\right), \\ \mathbf{u}|u_0, \sigma^2_u &\sim N\left(\mathbf{1}u_0, \mathbf{I}_t\sigma^2_u\right), \\ \mathbf{e} &\sim N\left(\mathbf{0}, \mathbf{I}_n\sigma^2_e\right), \end{aligned}$$

with the identity matrices having orders b, t, and n as indicated. The prior distribution of $\boldsymbol{\beta}$ postulates that all elements of this vector are i.i.d., with a common mean and variance. A similar assumption is made about the elements of $\mathbf{u}$. Above, β_0 and u_0 are unknown scalar parameters and σ^2_β, σ^2_u, and σ^2_e are unknown variance components. We shall adopt the independent prior densities

$$\begin{aligned} p\left(\beta_0\right) &= \frac{1}{\beta_{0,\max} - \beta_{0,\min}}, \\ p\left(u_0\right) &= \frac{1}{u_{0,\max} - u_{0,\min}}, \end{aligned}$$

where $(\min, \max)$ refer to $(upper, lower)$ boundaries for the appropriate parameters. The variance components are assigned independent scaled inverted chi-square distributions, with densities

$$\begin{aligned} p\left(\sigma^2_\beta|\nu_\beta, s^2_\beta\right) &\propto \left(\sigma^2_\beta\right)^{-\left(\frac{\nu_\beta+2}{2}\right)} \exp\left[-\frac{\nu_\beta s^2_\beta}{2\sigma^2_\beta}\right], \\ p\left(\sigma^2_u|\nu_u, s^2_u\right) &\propto \left(\sigma^2_u\right)^{-\left(\frac{\nu_u+2}{2}\right)} \exp\left[-\frac{\nu_u s^2_u}{2\sigma^2_u}\right], \\ p\left(\sigma^2_e|\nu_e, s^2_e\right) &\propto \left(\sigma^2_e\right)^{-\left(\frac{\nu_e+2}{2}\right)} \exp\left[-\frac{\nu_e s^2_e}{2\sigma^2_e}\right], \end{aligned}$$

where ν_β, ν_u, and ν_e are known "degree of belief" parameters, and s_β^2, s_u^2, and s_e^2 are some prior values of the dispersion components. An alternative representation of model (5.131) is

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ &= (\mathbf{X}\mathbf{1})\,\beta_0 + \mathbf{X}\boldsymbol{\xi}_\beta + (\mathbf{Z}\mathbf{1})\,u_0 + \mathbf{Z}\boldsymbol{\xi}_u + \mathbf{e} \\ &= \mathbf{m}\beta_0 + \mathbf{n}u_0 + \mathbf{X}\boldsymbol{\xi}_\beta + \mathbf{Z}\boldsymbol{\xi}_u + \mathbf{e}, \end{aligned} \tag{5.132}$$

where $\mathbf{m} = \mathbf{X}\mathbf{1}$ and $\mathbf{n} = \mathbf{Z}\mathbf{1}$ are incidence vectors of appropriate order and the new variables have distributions

$$\begin{aligned}\boldsymbol{\xi}_\beta|\sigma_\beta^2 &\sim N\left(\mathbf{0}, \mathbf{I}_b\sigma_\beta^2\right), \\ \boldsymbol{\xi}_u|\sigma_u^2 &\sim N\left(\mathbf{0}, \mathbf{I}_t\sigma_u^2\right).\end{aligned}$$

Hence, the entire parameter vector is then

$$\boldsymbol{\theta} = \left[\beta_0, u_0, \boldsymbol{\xi}_\beta', \boldsymbol{\xi}_u', \sigma_\beta^2, \sigma_u^2, \sigma_e^2\right]'.$$

Suppose we wish to find the joint mode of the posterior distribution of $\boldsymbol{\theta}$, having density

$$\begin{aligned}&p\left(\boldsymbol{\theta}|\beta_{0,\max}, \beta_{0,\min}, u_{0,\max}, u_{0,\min}, \nu_\beta, \nu_u, \nu_e, s_\beta^2, s_u^2, s_e^2, \mathbf{y}\right) \\ &\propto p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right) p\left(\beta_0|\beta_{0,\max}, \beta_{0,\min}\right) p\left(u_0|u_{0,\max}, u_{0,\min}\right) \\ &\quad \times p\left(\boldsymbol{\xi}_\beta|\sigma_\beta^2\right) p\left(\boldsymbol{\xi}_u|\sigma_u^2\right) \\ &\quad \times p\left(\sigma_\beta^2|\nu_\beta, s_\beta^2\right) p\left(\sigma_u^2|\nu_u, s_u^2\right) p\left(\sigma_e^2|\nu_e, s_e^2\right). \end{aligned} \tag{5.133}$$

In order to find a stationary point, the following first derivatives are needed, letting $L\left(\boldsymbol{\theta}|\mathbf{y}\right)$ be the logarithm of the joint posterior density.

Gradient for β_0:

$$\begin{aligned}\frac{\partial L\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial\beta_0} &= \frac{\partial}{\partial\beta_0}\left\{\log\left[p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right)\right]\right\} \\ &\quad + \frac{\partial}{\partial\beta_0}\left\{\log\left[p\left(\beta_0|\beta_{0,\max}, \beta_{0,\min}\right)\right]\right\} \\ &= -\frac{1}{2\sigma_e^2}\frac{\partial}{\partial\beta_0}\left(\mathbf{e}'\mathbf{e}\right) \\ &= \frac{\mathbf{m}'\left(\mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u\right)}{\sigma_e^2}\end{aligned}$$

where

$$\mathbf{e} = \mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u.$$

Gradient for u_0:

$$\begin{aligned}
&\frac{\partial L\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial u_0} \\
&= \frac{\partial}{\partial u_0}\left\{\log\left[p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right)\right]\right\} \\
&\quad + \frac{\partial}{\partial u_0}\left\{\log\left[p\left(u_0|u_{0,\max}, u_{0,\min}\right)\right]\right\} \\
&= -\frac{1}{2\sigma_e^2}\frac{\partial}{\partial u_0}\left(\mathbf{e}'\mathbf{e}\right) \\
&= \frac{\mathbf{n}'\left(\mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u\right)}{\sigma_e^2}
\end{aligned}$$

Gradient for ξ_β:

$$\begin{aligned}
\frac{\partial L\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\xi}_\beta} &= \frac{\partial}{\partial \boldsymbol{\xi}_\beta}\left\{\log\left[p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right)\right]\right\} \\
&\quad + \frac{\partial}{\partial \boldsymbol{\xi}_\beta}\left\{\log\left[p\left(\boldsymbol{\xi}_\beta|\sigma_\beta^2\right)\right]\right\} \\
&= -\frac{1}{2\sigma_e^2}\frac{\partial}{\partial \boldsymbol{\xi}_\beta}\left(\mathbf{e}'\mathbf{e}\right) - \frac{\partial}{\partial \boldsymbol{\xi}_\beta}\left(\frac{\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta}{2\sigma_\beta^2}\right) \\
&= \frac{\mathbf{X}'\left(\mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u\right)}{\sigma_e^2} - \frac{\boldsymbol{\xi}_\beta}{\sigma_\beta^2}.
\end{aligned}$$

Gradient for ξ_u:

$$\begin{aligned}
&\frac{\partial L\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\xi}_u} = \frac{\partial}{\partial \boldsymbol{\xi}_u}\left\{\log\left[p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right)\right]\right\} \\
&+ \frac{\partial}{\partial \boldsymbol{\xi}_u}\left\{\left[\log p\left(\boldsymbol{\xi}_u|\sigma_u^2\right)\right]\right\} \\
= \quad & -\frac{1}{2\sigma_e^2}\frac{\partial}{\partial \boldsymbol{\xi}_u}\left(\mathbf{e}'\mathbf{e}\right) - \frac{\partial}{\partial \boldsymbol{\xi}_u}\left(\frac{\boldsymbol{\xi}_u'\boldsymbol{\xi}_u}{2\sigma_u^2}\right) \\
= \quad & \frac{\mathbf{Z}'\left(\mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u\right)}{\sigma_e^2} - \frac{\boldsymbol{\xi}_u}{\sigma_u^2}.
\end{aligned}$$

Gradient for σ_e^2:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta}|\mathbf{y})}{\partial \sigma_e^2} &= \frac{\partial}{\partial \sigma_e^2}\left\{\log\left[p\left(\mathbf{y}|\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right)\right]\right\} \\ &+ \frac{\partial}{\partial \sigma_e^2}\left\{\log\left[p\left(\sigma_e^2|\nu_e, s_e^2\right)\right]\right\} \\ &= -\frac{n+\nu_e+2}{2\sigma_e^2} + \frac{\mathbf{e}'\mathbf{e}+\nu_e s_e^2}{2\sigma_e^4}.\end{aligned}$$

Gradient for σ_β^2:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta}|\mathbf{y})}{\partial \sigma_\beta^2} &= \frac{\partial}{\partial \sigma_\beta^2}\left\{\log\left[p\left(\boldsymbol{\xi}_\beta|\sigma_\beta^2\right)\right]\right. \\ &\left. + \log\left[p\left(\sigma_\beta^2|\nu_\beta, s_\beta^2\right)\right]\right\} \\ &= -\frac{b+\nu_\beta+2}{2\sigma_\beta^2} + \frac{\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta+\nu_\beta s_\beta^2}{2\sigma_\beta^4}.\end{aligned}$$

Gradient for σ_u^2:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta}|\mathbf{y})}{\partial \sigma_u^2} &= \frac{\partial}{\partial \sigma_u^2}\left\{\log\left[p\left(\boldsymbol{\xi}_u|\sigma_u^2\right)\right]\right. \\ &\left. + \log\left[p\left(\sigma_u^2|\nu_u, s_u^2\right)\right]\right\} \\ &= -\frac{t+\nu_u+2}{2\sigma_u^2} + \frac{\boldsymbol{\xi}_u'\boldsymbol{\xi}_u+\nu_u s_u^2}{2\sigma_u^4}.\end{aligned}$$

Setting all first derivatives simultaneously to 0 gives a system of equations that is not explicit in the solutions. A rearrangement of the system gives, after algebra, the following functional iteration (the superscript i in parentheses indicates round number):

$$\begin{bmatrix} \mathbf{m}'\mathbf{m} & \mathbf{m}'\mathbf{n} & \mathbf{m}'\mathbf{X} & \mathbf{m}'\mathbf{Z} \\ \mathbf{n}'\mathbf{m} & \mathbf{n}'\mathbf{n} & \mathbf{n}'\mathbf{X} & \mathbf{n}'\mathbf{Z} \\ \mathbf{X}'\mathbf{m} & \mathbf{X}'\mathbf{n} & \mathbf{X}'\mathbf{X}+\mathbf{I}\left(\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{(i)} & \mathbf{X}'\mathbf{X} \\ \mathbf{Z}'\mathbf{m} & \mathbf{Z}'\mathbf{n} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\left(\frac{\sigma_e^2}{\sigma_u^2}\right)^{(i)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ u_0 \\ \boldsymbol{\xi}_\beta \\ \boldsymbol{\xi}_u \end{bmatrix}^{(i)} = \begin{bmatrix} \mathbf{m}'\mathbf{y} \\ \mathbf{n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \tag{5.134}$$

$$\sigma_e^{2(i+1)} = \frac{(\mathbf{e}'\mathbf{e})^{(i)} + \nu_e s_e^2}{n+\nu_e+2},$$

$$\sigma_\beta^{2(i+1)} = \frac{\left(\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta\right)^{(i)} + \nu_\beta s_\beta^2}{b+\nu_\beta+2},$$

and:

$$\sigma_u^{2(i+1)} = \frac{\left(\boldsymbol{\xi}_u'\boldsymbol{\xi}_u\right)^{(i)} + \nu_u s_u^2}{t+\nu_u+2}.$$

The functional iteration starts by specifying starting values for σ_e^2, σ_β^2, and σ_u^2, and then solving (5.134), to obtain values for $\beta_0, u_0, \boldsymbol{\xi}_\beta$, and $\boldsymbol{\xi}_u$. The variance components are then updated with the three expressions below (5.134). If the iteration converges, it will produce a mode of the joint posterior distribution of all parameters. The iteration must be tailored such that values of β_0, u_0 stay within the boundaries stated in the probability model. ■

Example 5.14 *Marginal modes in a Gaussian linear model*
The setting is as in Example 5.13, but consider now finding the modal vector of the lower-dimensional distribution

$$\left[\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u | \beta_{0,\max}, \beta_{0,\min}, u_{0,\max}, u_{0,\min}, \nu_\beta, \nu_u, \nu_e, s_\beta^2, s_u^2, s_e^2, \mathbf{y}\right]. \tag{5.135}$$

The integral of the joint posterior density (5.133) with respect to all three variance components is needed to obtain the density of the desired lower-dimensional distribution. Omitting the dependency on hyperparameters in the notation, this integral is

$$\begin{aligned} p\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u | \mathbf{y}\right) \propto p\left(\beta_0\right) p\left(u_0\right) \int_0^\infty p\left(\mathbf{y} | \beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right) p\left(\sigma_e^2\right) d\sigma_e^2 \\ \times \int_0^\infty p\left(\boldsymbol{\xi}_\beta | \sigma_\beta^2\right) p\left(\sigma_\beta^2\right) d\sigma_\beta^2 \int_0^\infty p\left(\boldsymbol{\xi}_u | \sigma_u^2\right) p\left(\sigma_u^2\right) d\sigma_u^2. \end{aligned} \tag{5.136}$$

Each of the integrals is in a scaled inverted chi-squared form, and can be evaluated explicitly. We now evaluate each of the integrals in (5.136). Retaining only terms that vary with β_0, u_0, $\boldsymbol{\xi}_\beta$, and $\boldsymbol{\xi}_u$, we first get

$$\begin{aligned} &\int_0^\infty p\left(\mathbf{y} | \beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u, \sigma_e^2\right) p\left(\sigma_e^2\right) d\sigma_e^2 \\ &\propto \int_0^\infty \left(\sigma_e^2\right)^{-\left(\frac{n+\nu_e+2}{2}\right)} \exp\left[-\frac{\mathbf{e}'\mathbf{e} + \nu_e s_e^2}{2\sigma_e^2}\right] d\sigma_e^2 \\ &\propto \left[\mathbf{e}'\mathbf{e} + \nu_e s_e^2\right]^{-\frac{n+\nu_e}{2}} \\ &\propto \left[1 + \frac{\mathbf{e}'\mathbf{e}}{\nu_e s_e^2}\right]^{-\frac{n+\nu_e}{2}}, \end{aligned} \tag{5.137}$$

recalling that

$$\mathbf{e} = \left(\mathbf{y} - \mathbf{m}\beta_0 - \mathbf{n}u_0 - \mathbf{X}\boldsymbol{\xi}_\beta - \mathbf{Z}\boldsymbol{\xi}_u\right).$$

Similarly,

$$\int_0^\infty p\left(\boldsymbol{\xi}_\beta|\sigma_\beta^2\right) p\left(\sigma_\beta^2\right) d\sigma_\beta^2$$
$$\propto \int_0^\infty \left(\sigma_\beta^2\right)^{-\left(\frac{b+\nu_\beta+2}{2}\right)} \exp\left[-\frac{\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta+\nu_\beta s_\beta^2}{2\sigma_\beta^2}\right] d\sigma_\beta^2$$
$$\propto \left[1+\frac{\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta}{\nu_\beta s_\beta^2}\right]^{-\frac{b+\nu_\beta}{2}}. \tag{5.138}$$

Further, using a similar algebra,

$$\int_0^\infty p\left(\boldsymbol{\xi}_u|\sigma_u^2\right) p\left(\sigma_u^2\right) d\sigma_u^2 \propto \left[1+\frac{\boldsymbol{\xi}_u'\boldsymbol{\xi}_u}{\nu_u s_u^2}\right]^{-\frac{t+\nu_u}{2}}. \tag{5.139}$$

Each of the expressions in (5.137)-(5.139) is the kernel of a multivariate-t distribution. Now, collecting the integrals and using these in (5.136) we obtain, as posterior density, after suitable normalization,

$$p\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right) \propto \frac{\prod_{i=1}^{3}\left[1+\frac{\boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i}{c_i}\right]^{-\frac{d_i}{2}}}{\prod_{i=1}^{3}\int_{-\infty}^{\infty}\left[1+\frac{\boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i}{c_i}\right]^{-\frac{d_i}{2}} d\boldsymbol{\lambda}_i}, \tag{5.140}$$

where

$$\begin{aligned}
\boldsymbol{\lambda}_1 &= \mathbf{y}-\mathbf{m}\beta_0-\mathbf{n}u_0-\mathbf{X}\boldsymbol{\xi}_\beta-\mathbf{Z}\boldsymbol{\xi}_u, \quad c_1=\nu_e s_e^2, \quad d_1=n+\nu_e,\\
\boldsymbol{\lambda}_2 &= \boldsymbol{\xi}_\beta, \quad c_2=\nu_\beta s_\beta^2, \quad d_2=b+\nu_\beta,\\
\boldsymbol{\lambda}_3 &= \boldsymbol{\xi}_u, \quad c_3=\nu_u s_u^2, \quad d_3=t+\nu_u.
\end{aligned}$$

If the $\boldsymbol{\lambda}_i'$s were the random variables of interest, the distribution with density (5.140) would be a truncated (in the intervals $\beta_{0,\max}-\beta_{0,\min}$ and $u_{0,\max}-u_{0,\min}$) poly-t or product multivariate-t distribution (Box and Tiao, 1973); its properties are presented in Dickey (1968). For example, this distribution is known to be asymmetric and multimodal. However, note that $\boldsymbol{\lambda}_1$ involves all four random terms of interest, so the process of interest is not poly-t. Consider now finding the mode of the marginal distribution of concern. Differentiation of the logarithm of (5.140), $L\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right)$, with respect to each of the four terms, yields,

$$\frac{\partial L\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right)}{\partial \beta_0} = \frac{d_1\mathbf{m}'\mathbf{e}}{\left[c_1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1\right]},$$

$$\frac{\partial L\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right)}{\partial u_0} = \frac{d_1\mathbf{n}'\mathbf{e}}{\left[c_1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1\right]},$$

$$\frac{\partial L\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right)}{\partial \boldsymbol{\xi}_\beta} = \frac{d_1\mathbf{X}'\mathbf{e}}{\left[c_1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1\right]} - \frac{d_2\boldsymbol{\xi}_\beta}{\left[c_2 + \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2\right]},$$

and

$$\frac{\partial L\left(\beta_0, u_0, \boldsymbol{\xi}_\beta, \boldsymbol{\xi}_u|\mathbf{y}\right)}{\partial \boldsymbol{\xi}_u} = \frac{d_1\mathbf{Z}'\mathbf{e}}{\left[c_1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1\right]} - \frac{d_3\boldsymbol{\xi}_u}{\left[c_3 + \boldsymbol{\lambda}_3'\boldsymbol{\lambda}_3\right]}.$$

Setting all differentials simultaneously to zero gives

$$\mathbf{m}'\mathbf{m}\beta_0 + \mathbf{m}'\mathbf{n}u_0 + \mathbf{m}'\mathbf{X}\boldsymbol{\xi}_\beta + \mathbf{m}'\mathbf{Z}\boldsymbol{\xi}_u = \mathbf{m}'\mathbf{y},$$

$$\mathbf{n}'\mathbf{m}\beta_0 + \mathbf{n}'\mathbf{n}u_0 + \mathbf{n}'\mathbf{X}\boldsymbol{\xi}_\beta + \mathbf{n}'\mathbf{Z}\boldsymbol{\xi}_u = \mathbf{n}'\mathbf{y},$$

$$\mathbf{X}'\mathbf{m}\beta_0 + \mathbf{X}'\mathbf{n}u_0 + \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{w_1}{w_2}\right)\boldsymbol{\xi}_\beta + \mathbf{X}'\mathbf{Z}\boldsymbol{\xi}_u = \mathbf{X}'\mathbf{y},$$

$$\mathbf{Z}'\mathbf{m}\beta_0 + \mathbf{Z}'\mathbf{n}u_0 + \mathbf{Z}'\mathbf{X}\boldsymbol{\xi}_\beta + \left(\mathbf{Z}'\mathbf{Z} + \mathbf{I}\frac{w_1}{w_3}\right)\boldsymbol{\xi}_u = \mathbf{Z}'\mathbf{y},$$

where

$$w_1 = \frac{c_1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}{d_1} = \frac{\mathbf{e}'\mathbf{e} + \nu_e s_e^2}{n + \nu_e},$$

$$w_2 = \frac{c_2 + \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2}{d_2} = \frac{\boldsymbol{\xi}_\beta'\boldsymbol{\xi}_\beta + \nu_\beta s_\beta^2}{b + \nu_\beta},$$

and

$$w_3 = \frac{c_3 + \boldsymbol{\lambda}_3'\boldsymbol{\lambda}_3}{d_3} = \frac{\boldsymbol{\xi}_u'\boldsymbol{\xi}_u + \nu_u s_u^2}{t + \nu_u},$$

can be construed as "estimates" of variance components in a Gaussian linear mixed effects model (Henderson, 1973; Searle et al., 1992). Clearly, the equations that need to be solved simultaneously are not explicit in the solutions. As in Example 5.13, a functional iteration can be constructed, which can be expressed in matrix form as

$$\begin{bmatrix} \mathbf{m}'\mathbf{m} & \mathbf{m}'\mathbf{n} & \mathbf{m}'\mathbf{X} & \mathbf{m}'\mathbf{Z} \\ \mathbf{n}'\mathbf{m} & \mathbf{n}'\mathbf{n} & \mathbf{n}'\mathbf{X} & \mathbf{n}'\mathbf{Z} \\ \mathbf{X}'\mathbf{m} & \mathbf{X}'\mathbf{n} & \mathbf{X}'\mathbf{X} + \mathbf{I}\frac{w_1^{(i)}}{w_2^{(i)}} & \mathbf{X}'\mathbf{X} \\ \mathbf{Z}'\mathbf{m} & \mathbf{Z}'\mathbf{n} & \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}\frac{w_1^{(i)}}{w_3^{(i)}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ u_0 \\ \boldsymbol{\xi}_\beta \\ \boldsymbol{\xi}_u \end{bmatrix}^{(i+1)} = \begin{bmatrix} \mathbf{m}'\mathbf{y} \\ \mathbf{n}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{5.141}$$

where the w's depend on the unknowns, and change values from iterate to iterate. The algorithm starts by assigning starting values to the pseudo-variances (perhaps the means of the prior distributions of the variance components), and then calculating the first set of w's. These are used to obtain revised values for the location effects, and so on. The properties

of the algorithm are unknown. If it converges, it will locate one of the possibly many stationary points of the poly$-t$ distributions. Note that (5.141) are structurally similar to (5.134). However, the algorithm does not involve estimating equations for the variance components, as these have been integrated out of the joint posterior distribution. While it is extremely difficult to show that the modal vector of the joint distribution of all parameters differs from the mode of the lower-dimensional distribution with density (5.140), this example illustrates at least that the modes have different forms in the two cases. ■

5.6.4 Posterior Mean Vector and Covariance Matrix

Mean Vector

The mean (mean vector) and variance (covariance matrix) of posterior distributions have been identified in several of the highly stylized examples discussed before. It is convenient, however, to recall the pervasive presence of unknown nuisance parameters, so the following notation is helpful in this respect. The mean of the posterior distribution of a vector $\boldsymbol{\theta}_1$, when the statistical model posits the presence of a nuisance parameter $\boldsymbol{\theta}_2$, can be expressed as

$$E\left(\boldsymbol{\theta}_1|\mathbf{y}\right) = E_{\boldsymbol{\theta}_2|y}\left[E\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\mathbf{y}\right)\right]. \tag{5.142}$$

The inner expectation gives the posterior mean value of the parameter of interest at specified values of the nuisance parameters, as if these were known. The outer expectation averages over the marginal posterior distribution of the nuisances, thus incorporating whatever uncertainty exists about their values. Representation (5.142) is also useful in connection with sampling methods. As seen in Chapter 1, the process of Rao–Blackwellization enables one to obtain a more precise Monte Carlo estimator of $E\left(\boldsymbol{\theta}_1|\mathbf{y}\right)$ by drawing m samples from the posterior distribution of the nuisance parameter (whenever this is feasible), and then estimating the desired posterior mean as

$$\widehat{E}\left(\boldsymbol{\theta}_1|\mathbf{y}\right) = \frac{1}{m}\sum_{i=1}^{m} E\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(i)},\mathbf{y}\right),$$

where $\boldsymbol{\theta}_2^{(1)},\boldsymbol{\theta}_2^{(2)},\ldots,\boldsymbol{\theta}_2^{(m)}$ are draws from $[\boldsymbol{\theta}_2|\mathbf{y}]$. The preceding estimator is at least as precise as

$$\widehat{\widehat{E}}\left(\boldsymbol{\theta}_1|\mathbf{y}\right) = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}_1^{(i)},$$

where $\boldsymbol{\theta}_1$ is a direct draw from the posterior distribution $[\boldsymbol{\theta}_1|\mathbf{y}]$.

Optimality of the Mean Vector

Use of the mean as an "estimator" of the parameter has a decision-theoretic justification. Suppose the loss function is quadratic, with the form

$$L\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1\right) = \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)' \mathbf{Q} \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right),$$

where $\mathbf{Q}$ is a known symmetric, positive-definite matrix. The "error" of estimation is $\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1$, and the form adopted above for $L\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1\right)$ indicates that the penalty accruing from estimating the parameter with error is proportional to the square of such error. Using well-known formulas for expected values of quadratic forms, the expected posterior loss is

$$\begin{aligned} E_{\boldsymbol{\theta}_1|\mathbf{y}}\left[L\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1\right)\right] = &\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right]' \mathbf{Q} \left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right] \\ &+ tr\left[\mathbf{Q}\, Var\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right]. \end{aligned} \tag{5.143}$$

The second term does not involve $\widehat{\boldsymbol{\theta}}_1$, and the first term is nonnegative. Hence, the expected loss is minimized by taking as "optimal estimator":

$$\widehat{\boldsymbol{\theta}}_1 = E\left(\boldsymbol{\theta}_1|\mathbf{y}\right) \tag{5.144}$$

which is the posterior mean.

Relationship Between the Posterior Mean and the "Best" Predictor

Consider adopting a frequentist point of view, that is, let $[\boldsymbol{\theta}_1, \mathbf{y}|\boldsymbol{\theta}_2]$ be a joint distribution having a long-run frequency interpretation, where $\boldsymbol{\theta}_2$ is a known parameter (this is a very strong assumption; in practice, such a parameter is unknown, most often). In this setting, $\boldsymbol{\theta}_1$ is an unobservable random variable, or "random effect" in the usual frequentist sense. Then, using a logic similar to the preceding one, we will show that the frequentist expected loss

$$E_{\boldsymbol{\theta}_1, \mathbf{y}|\boldsymbol{\theta}_2}\left[L\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1\right)\right] = E_{\boldsymbol{\theta}_1, \mathbf{y}|\boldsymbol{\theta}_2}\left[\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)' \mathbf{Q} \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)\right] \tag{5.145}$$

is minimized by $\widehat{\boldsymbol{\theta}}_1 = E\left(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2\right)$, among all possible predictors. In (5.145), expectations are taken with respect to the joint distribution $[\boldsymbol{\theta}_1, \mathbf{y}|\boldsymbol{\theta}_2]$, with the nuisance parameter $\boldsymbol{\theta}_2$ treated as a known constant. Using the theorem

of double expectation

$$\begin{aligned} E_{\boldsymbol{\theta}_1,\mathbf{y}|\boldsymbol{\theta}_2}\left[L\left(\widehat{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_1\right)\right] &= E_{\mathbf{y}|\boldsymbol{\theta}_2}\left\{E_{\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2}\left[L\left(\widehat{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_1\right)\right]\right\} \\ &= E_{\mathbf{y}|\boldsymbol{\theta}_2}\left\{\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]' \mathbf{Q}\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]\right. \\ &\qquad \left. + \left[\mathbf{Q}\, Var\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]\right\} \\ &= E_{\mathbf{y}|\boldsymbol{\theta}_2}\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]' \mathbf{Q}\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right] \\ &\qquad + E_{\mathbf{y}|\boldsymbol{\theta}_2} tr\left[\mathbf{Q}\, Var\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]. \end{aligned} \tag{5.146}$$

The second term in (5.146) does not involve $\widehat{\boldsymbol{\theta}}_1$; thus, minimizing expected (frequentist) quadratic loss is achieved by minimizing the first term. Now, the latter is a weighted average, where the weight function is the density of the distribution $[\mathbf{y}|\boldsymbol{\theta}_2]$. It turns out that, if one minimizes,

$$\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]' \mathbf{Q}\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right] \tag{5.147}$$

for each realization of $\mathbf{y}$, this will also minimize

$$E_{\mathbf{y}|\boldsymbol{\theta}_2}\left\{\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]' \mathbf{Q}\left[\widehat{\boldsymbol{\theta}}_1 - E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right]\right\}$$

and, hence, (5.146). Using the same argument as in the preceding section, it follows that the minimizer for each $\mathbf{y}$ is the conditional expectation $E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)$ which, superficially, "looks similar" to our Bayesian $E\left(\boldsymbol{\theta}_1|\mathbf{y}\right)$. The similarity does not stand scrutiny, however, as $E\left(\boldsymbol{\theta}_1|\mathbf{y}\right)$ incorporates the uncertainty about the nuisance parameter $\boldsymbol{\theta}_2$, this is clearly not the case for $\widehat{\boldsymbol{\theta}}_1 = E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)$.

The "best predictor" is, thus, the conditional mean of the unobservable random effects given the observations, but assuming known parameters of the joint distribution of the data and of the random effects (Henderson, 1973). This statistic was introduced in Example 1.22 of Chapter 1. In addition to minimizing the expected squared error of prediction (5.145), the conditional mean has some frequentist properties of interest. For example, it is "unbiased", in some sense. This can be verified by taking the expectations of the conditional mean over the distribution of the observations

$$E_{\mathbf{y}|\boldsymbol{\theta}_2}\left(\widehat{\boldsymbol{\theta}}_1\right) = E_{\mathbf{y}|\boldsymbol{\theta}_2}\left[E\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2\right)\right] = E\left(\boldsymbol{\theta}_1\right), \tag{5.148}$$

recalling that the parameter $\boldsymbol{\theta}_2$ is not a random variable in the frequentist sense, so the dependence on it can be omitted in the notation. This is the definition of unbiasedness in the frequentist context of prediction of random effects. Now, since the conditional mean is an unbiased predictor, it follows that it minimizes prediction error variance as well. This can be seen simply

by putting $\mathbf{Q} = \mathbf{I}$ in (5.145), and by considering a scalar element of $\boldsymbol{\theta}_1$; then, the expected loss is the prediction error variance. A third property is

$$Cov\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1'\right) = Var\left(\widehat{\boldsymbol{\theta}}_1\right). \tag{5.149}$$

This results from the fact that

$$\begin{aligned} Cov\left(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1'\right) &= E_{\mathbf{y},\boldsymbol{\theta}_1|\boldsymbol{\theta}_2}\left(\widehat{\boldsymbol{\theta}}_1\boldsymbol{\theta}_1'\right) - E_{\mathbf{y},\boldsymbol{\theta}_1|\boldsymbol{\theta}_2}\left(\widehat{\boldsymbol{\theta}}_1\right) E\left(\boldsymbol{\theta}_1'\right) \\ &= E_{\mathbf{y}|\boldsymbol{\theta}_2}\left[E_{\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2}\left(\widehat{\boldsymbol{\theta}}_1\boldsymbol{\theta}_1'\right)\right] - E\left(\widehat{\boldsymbol{\theta}}_1\right) E\left(\widehat{\boldsymbol{\theta}}_1'\right) \\ &= E_{\mathbf{y}|\boldsymbol{\theta}_2}\left[\widehat{\boldsymbol{\theta}}_1 E_{\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2}\left(\boldsymbol{\theta}_1'\right)\right] - E\left(\widehat{\boldsymbol{\theta}}_1\right) E\left(\widehat{\boldsymbol{\theta}}_1'\right) \\ &= E\left[\widehat{\boldsymbol{\theta}}_1\widehat{\boldsymbol{\theta}}_1'\right] - E\left(\widehat{\boldsymbol{\theta}}_1\right) E\left(\widehat{\boldsymbol{\theta}}_1'\right) = Var\left(\widehat{\boldsymbol{\theta}}_1\right). \end{aligned}$$

An immediate consequence is that the prediction error $\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1$ and the predictor $\widehat{\boldsymbol{\theta}}_1$ have a null covariance. Additional consequences of (5.149), for any scalar element of $\boldsymbol{\theta}_1$, say θ_{1i}, are that

$$Corr\left(\widehat{\theta}_{1i}, \theta_{1i}\right) = \sqrt{\frac{Var\left(\widehat{\theta}_{1i}\right)}{Var\left(\theta_{1i}\right)}},$$

and

$$Var\left(\widehat{\theta}_{1i} - \theta_{1i}\right) = Var\left(\theta_{1i}\right)\left[1 - Corr^2\left(\widehat{\theta}_{1i}, \theta_{1i}\right)\right].$$

In general,

$$Var\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) = Var\left(\boldsymbol{\theta}_1\right) - Var\left(\widehat{\boldsymbol{\theta}}_1\right). \tag{5.150}$$

Example 5.15 *Best prediction of quadratic genetic merit*
Suppose we are in the setting of Example 5.3, and that the function

$$T = k_0 a + k_1 a^2$$

is to be predicted. As before, a is the additive genetic value of a certain farm animal for some trait, and k_0 and k_1 are known constants. This function is what is called "quadratic merit" in animal breeding (Wilton et al., 1968), and it postulates that the breeding worth of a potential parent is proportional to the square of its additive genetic value. The expectation of T is

$$\begin{aligned} E\left(T\right) &= k_0 E\left(a\right) + k_1 E^2\left(a\right) + k_1 v_a \\ &= k_1 v_a \end{aligned}$$

since it is assumed that $E\left(a\right) = 0$. As in Example 5.3 suppose that μ, v_a, and v_e are known. The frequentist conditional distribution of a, given the

phenotypic value y, coincides in this case with the Bayesian conditional posterior distribution of a. As shown in Example 5.3, this distribution is

$$a|y,\mu,v_a,v_e \sim N\left(h^2\left(y-\mu\right),v_a\left(1-h^2\right)\right),$$

where $h^2 = v_a/\left(v_a+v_e\right)$. From results in the preceding section, the best predictor of T is

$$\begin{aligned}\widehat{T} &= E\left(T|y,\mu,v_a,v_e\right) = k_0E\left(a|y,\mu,v_a,v_e\right)+k_1E\left(a^2|y,\mu,v_a,v_e\right)\\ &= k_0h^2\left(y-\mu\right)+k_1\left[h^2\left(y-\mu\right)\right]^2+k_1v_a\left(1-h^2\right),\end{aligned}$$

and this can be evaluated readily. It is easy to verify that the predictor is unbiased, as it has the same expectation as the predictand. Taking expectations over the distribution of y:

$$\begin{aligned}E_yE\left(T|y,\mu,v_a,v_e\right) &= k_0E\left(a\right)+k_1E\left[E\left(a^2|y,\mu,v_a,v_e\right)\right]\\ &= k_0E\left(a\right)+k_1E\left(a^2\right)=k_1v_a.\end{aligned}$$

Using (5.150), the prediction error variance is

$$Var\left(\widehat{T}-T\right)=Var\left(T\right)-Var\left(\widehat{T}\right).$$

In the preceding expression

$$\begin{aligned}Var\left(T\right) &= Var\left(k_0a+k_1a^2\right)\\ &= k_0^2v_a+k_1^2\,Var(a^2)+2k_0k_1\,Cov(a,a^2)\\ &= k_0^2v_a+2k_1^2v_a^2\end{aligned}$$

with this being so, because if $a\sim N\left(0,v_a\right)$, then $Var\left(a^2\right)=2v_a^2$ and $Cov(a,a^2)=0$ (Searle, 1971). Further, letting

$$\widehat{a}=E\left(a|y,\mu,v_a,v_e\right),$$

and using similar arguments, it follows that

$$\begin{aligned}Var\left(\widehat{T}\right) &= Var\left\{k_0h^2\left(y-\mu\right)+k_1\left[h^2\left(y-\mu\right)\right]^2+k_1v_a\left(1-h^2\right)\right\}\\ &= Var\left(k_0\widehat{a}+k_1\widehat{a}^2\right)\\ &= k_0^2\,Var\left(\widehat{a}\right)+2k_1^2\,Var^2\left(\widehat{a}\right)\end{aligned}$$

with

$$Var\left(\widehat{a}\right)=h^4\left(v_a+v_e\right).$$

Collecting terms,

$$\begin{aligned}Var\left(\widehat{T}-T\right) &= k_0^2\left[v_a-Var\left(\widehat{a}\right)\right]+2k_1^2\left[v_a^2-Var^2\left(\widehat{a}\right)\right]\\ &= \left[v_a-Var\left(\widehat{a}\right)\right]\left\{k_0^2+2k_1^2\left[v_a+Var\left(\widehat{a}\right)\right]\right\}.\end{aligned}$$

■

Dispersion Matrix of the Posterior Distribution

The posterior variance–covariance matrix between parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is

$$\begin{aligned} Cov\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right) &= \int\int \left(\boldsymbol{\theta}_1 - \overline{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta}_2 - \overline{\boldsymbol{\theta}}_2\right)' p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 \\ &= \int\int \boldsymbol{\theta}_1 \boldsymbol{\theta}_2' p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 - \overline{\boldsymbol{\theta}}_1 \overline{\boldsymbol{\theta}}_2, \end{aligned} \tag{5.151}$$

where

$$\overline{\boldsymbol{\theta}}_i = E\left(\boldsymbol{\theta}_i|\mathbf{y}\right), \quad i = 1, 2.$$

The posterior correlation between any pair of parameters can be deduced readily from (5.151). There are situations in which a high degree of posterior intercorrelation between parameters can be modified drastically by some suitable reparameterization. This can have an impact in the numerical behavior of Markov chain Monte Carlo (MCMC) algorithms for sampling from posterior distributions. We shall return to this in the MCMC part of this book.

If the probability model includes a third vector of nuisance parameters, say $\boldsymbol{\theta}_3$, a sometimes useful representation of the variance–covariance matrix, of the joint posterior distribution of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is

$$\begin{aligned} Cov\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}\right) &= E_{\boldsymbol{\theta}_3|\mathbf{y}}\left[Cov\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{\theta}_3, \mathbf{y}\right)\right] \\ &+ Cov_{\boldsymbol{\theta}_3|\mathbf{y}}\left[E\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_3, \mathbf{y}\right), E\left(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3, \mathbf{y}\right)\right]. \end{aligned} \tag{5.152}$$

The basic principles of Bayesian analysis have been covered at this point using relatively simple settings. In the following chapter, a detailed discussion of the linear regression model and of the mixed linear model will be presented from a Bayesian perspective.

This page intentionally left blank

6
Bayesian Analysis of Linear Models

6.1 Introduction

A review of the basic tenets of Bayesian inference was given in the preceding chapter. Here the treatment is extended by presenting the Bayesian analysis of some standard linear models used in quantitative genetics, and appropriate analytical solutions are derived whenever these exist. First, the standard linear regression model is discussed from a Bayesian perspective. Subsequently, the Bayesian mixed effects model under Gaussian assumptions is contrasted with its frequentist counterpart. The chapter finishes with a presentation of several marginal and conditional distributions of interest. The developments presented give a necessary background for a fully Bayesian analysis of linear models via Markov chain Monte Carlo methods, a topic to be discussed subsequently in this book.

6.2 The Linear Regression Model

Arguably, linear regression analysis is one of the most widely used statistical methods and its use in genetics probably dates back to Galton (1885). For example, one of the simplest methods for estimating heritability is based on regressing the mean value for a quantitative trait measured on several offspring from a pair of parents on the midparental mean (Falconer and Mackay, 1996). While regression models have been discussed earlier in this book, a more general tour of the Bayesian analysis of such models is

presented in this section. Most of the needed notation has been presented before, so we concentrate on essentials only.

A Gaussian linear regression model, making a distinction between two distinct sets of coefficients $\boldsymbol{\beta}_1$, $(p_1 \times 1)$ and $\boldsymbol{\beta}_2$, $(p_2 \times 1)$, is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} \\ &= \left[\begin{array}{cc} \mathbf{X}_1 & \mathbf{X}_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{array}\right] + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \end{aligned} \tag{6.1}$$

where the error term is $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2\right)$ and σ^2 is an unknown dispersion parameter. The likelihood function is then

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]. \tag{6.2}$$

It has been seen before that

$$(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2) = S_e + S_\beta,$$

where

$$\begin{aligned} S_e &= \left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right)' \left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\right), \\ S_\beta &= \left[\begin{array}{cc} \left(\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1\right)' & \left(\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_2\right)' \end{array}\right] \mathbf{C} \left[\begin{array}{c} \boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_2 \end{array}\right], \end{aligned}$$

$$\left[\begin{array}{c} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{array}\right] = \mathbf{C}^{-1} \left[\begin{array}{c} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{array}\right],$$

and

$$\mathbf{C} = \left[\begin{array}{cc} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{array}\right].$$

This decomposition leads rather directly to the joint, marginal and conditional posterior distributions of interest. Results from Bayesian analyses under two different prior specifications will be presented separately.

6.2.1 Inference under Uniform Improper Priors

Joint Posterior Density

Suppose a uniform distribution is adopted as joint prior for all elements of the parameter vector $\boldsymbol{\theta} = \left[\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma^2\right]'$. Unless such a uniform prior does not have finite boundaries, it is improper. At first sight improper priors do not seem to make sense in the light of probability theory, at least as presented so far. However, improper priors have played a role in an area usually referred to as "objective Bayesian analysis", an approach founded

by Jeffreys (1961). For some parameters and models (Box and Tiao, 1973; Bernardo and Smith, 1994), an improper uniform distribution can be shown to introduce as little information as possible, in some sense, beyond that contained in the data.

A specification based on improper uniform priors will be developed in this section. Here, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are allowed to take any values in the p_1- and p_2-dimensional spaces $\Re^{p_1}$ and $\Re^{p_2}$ respectively, while σ^2 falls between 0 and ∞. Hence, the joint posterior density is strictly proportional to the likelihood function. After normalization (assuming the integrals exist), the joint density can be written as

$$p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) = \frac{\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right]}{\int\limits_{\Re^{p_1}} \int\limits_{\Re^{p_2}} \int\limits_{0}^{\infty} \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right] d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2 \, d\sigma^2}. \tag{6.3}$$

The limits of integration are omitted in the notation hereinafter.

Conditional Posterior Distributions of the Regression Coefficients

These can be found directly from the joint posterior density by retaining the part that varies with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. One gets

$$p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \sigma^2, \mathbf{y}\right) \propto \exp\left[-\frac{S_\beta}{2\sigma^2}\right].$$

It follows that the joint posterior distribution of the regression coefficients, given σ^2, is

$$\left[\begin{array}{c} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{array}\right] \Bigg| \sigma^2, \mathbf{y} \propto N\left(\left[\begin{array}{c} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{array}\right], \left[\begin{array}{cc} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{array}\right]^{-1} \sigma^2\right). \tag{6.4}$$

Distributions of individual elements of $\boldsymbol{\beta}$ are, thus, normal, with mean given directly by the corresponding component of the mean vector, and with variance equal to the appropriate element of the inverse matrix above, times σ^2. Posterior distributions of linear combinations of the regression coefficients are normal as well, given σ^2.

If, in addition to σ^2, one treats either $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$ as known, the resulting conditional posterior distributions are

$$\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2, \sigma^2, \mathbf{y} \propto N\left(\widetilde{\boldsymbol{\beta}}_1, \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} \sigma^2\right) \tag{6.5}$$

and

$$\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1, \sigma^2, \mathbf{y} \propto N\left(\widetilde{\boldsymbol{\beta}}_2, \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1} \sigma^2\right), \tag{6.6}$$

where

$$\widetilde{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i' \left(\mathbf{y} - \mathbf{X}_j \boldsymbol{\beta}_j\right), \quad i = 1, 2, \; i \neq j.$$

The conditional posterior distribution of an individual regression coefficient, β_k, given σ^2 and all other regression coefficients $\left(\boldsymbol{\beta}_{-k}\right)$ is also normal, with mean

$$\widetilde{\beta}_k = \frac{\mathbf{x}_k' \left(\mathbf{y} - \mathbf{X}_{-k}\boldsymbol{\beta}_{-k}\right)}{\mathbf{x}_k'\mathbf{x}_k}, \tag{6.7}$$

where $\mathbf{x}_k$ is the kth column of $\mathbf{X}$, and $\mathbf{X}_{-k}$ is $\mathbf{X}$ without column $\mathbf{x}_k$. The variance is

$$Var\left(\beta_k|\boldsymbol{\beta}_{-k}, \sigma^2, \mathbf{y}\right) = \frac{\sigma^2}{\mathbf{x}_k'\mathbf{x}_k}. \tag{6.8}$$

Conditional Posterior Distribution of σ^2

If the joint density is viewed now only as a function of σ^2, with the regression parameters treated as constants, one gets

$$p\left(\sigma^2|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right]. \tag{6.9}$$

This is the kernel of a scaled inverse chi-square process with degree of belief parameter equal to $n-2$ (this "loss" of two degrees of freedom is a curious consequence of the uniform prior adopted), and scale parameter $\left(S_e + S_\beta\right)/\left(n-2\right)$. It has mean and variance equal to

$$E\left(\sigma^2|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}\right) = \frac{\left(S_e + S_\beta\right)}{n-4}$$

and to

$$Var\left(\sigma^2|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}\right) = \frac{2\left(S_e + S_\beta\right)^2}{\left(n-4\right)^2\left(n-6\right)}.$$

For the conditional posterior distribution of σ^2 to exist, it is necessary that $n > 2$. The mean of the distribution exists if $n > 4$, and the variance of the process is defined only if $n > 6$. Thus, given certain constraints on sample size, the conditional posterior distribution is proper even if the prior is not so.

Marginal Distributions of the Regression Coefficients

Integration of the joint density over σ^2 gives

$$p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{y}\right) \propto \int \left(\sigma^2\right)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right] d\sigma^2$$

$$\propto \left(S_e + S_\beta\right)^{-\left(\frac{n-2}{2}\right)} \propto \left[1 + \frac{S_\beta}{\left(n-2-p_1-p_2\right)\frac{S_e}{\left(n-2-p_1-p_2\right)}}\right]^{-k}, \tag{6.10}$$

where $k = (n - 2 - p_1 - p_2 + p_1 + p_2)/2$. This is the kernel of a $p_1 + p_2$-dimensional multivariate-t distribution, with mean $\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2']'$, degrees of freedom $n - 2 - p_1 - p_2$, and variance–covariance matrix

$$Var\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{y}\right) = \frac{S_e}{(n - p_1 - p_2 - 4)} \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1}.$$

The distribution is proper only if $n > p_1 + p_2 + 2$.

Since the joint distribution of the regression coefficients is multivariate t, it follows that any marginal and conditional distributions of the regression coefficients (after integrating σ^2 out) are either univariate or multivariate-t (see Chapter 1). The same is true for the posterior distribution of any linear combination of the coefficients.

Marginal Distribution of σ^2

Integrating the joint density with respect to the regression coefficients gives

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{S_e}{2\sigma^2}\right] \int\int \exp\left[-\frac{S_\beta}{2\sigma^2}\right] d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2.$$

The integrand is the kernel of the density of the multivariate normal distribution (6.4), so the integral evaluates to

$$(2\pi)^{\frac{p_1+p_2}{2}} \left|\mathbf{C}^{-1}\sigma^2\right|^{\frac{1}{2}}.$$

Using this in the previous expression and retaining only those terms varying with σ^2, one gets

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\left(\frac{n-p_1-p_2-2}{2}+1\right)} \exp\left(-\frac{S_e}{2\sigma^2}\right). \tag{6.11}$$

This indicates that the marginal posterior distribution of the variance is a scaled inverted chi-square process with degree of belief parameter $n - p_1 - p_2 - 2$, mean equal to

$$E\left(\sigma^2|\mathbf{y}\right) = \frac{S_e}{n - p_1 - p_2 - 4}, \tag{6.12}$$

and variance

$$Var\left(\sigma^2|\mathbf{y}\right) = \frac{2S_e^2}{\left(n - p_1 - p_2 - 4\right)^2 \left(n - p_1 - p_2 - 6\right)}. \tag{6.13}$$

The distribution is proper provided that $n > p_1 + p_2 + 2 = rank\left(\mathbf{X}\right) + 2$.

Posterior Distribution of the Residuals

The residuals contain information about the quality of fit of the model (Draper and Smith, 1981; Bates and Watts, 1988). Hence, an analysis of the residuals is often indicated as a starting point for exploring the adequacy of the assumptions, for example, of the functional form adopted. From a Bayesian perspective, this is equivalent to evaluating the posterior distribution of the residuals. If the residual for an observation is not centered near zero, this can be construed as evidence that the model is probably inconsistent with such observation (Albert and Chib, 1995). Let the residual for datum i be

$$e_i = y_i - \mathbf{x}_i'\boldsymbol{\beta},$$

where $\mathbf{x}_i'$ is now the i^{th} row of $\mathbf{X}$. Since y_i is fixed in Bayesian analysis, then this is a linear combination of the random variable $\mathbf{x}_i'\boldsymbol{\beta}$. The latter, a posteriori, is distributed as univariate-t on $n - 2 - p_1 - p_2$ degrees of freedom, with mean $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, and variance

$$Var\left(\mathbf{x}_i'\boldsymbol{\beta}|\mathbf{y}\right) = \frac{S_e\mathbf{x}_i'\mathbf{C}^{-1}\mathbf{x}_i}{(n - p_1 - p_2 - 4)}. \tag{6.14}$$

It follows that the posterior distribution of e_i is univariate-t, also on $n - 2 - p_1 - p_2$ degrees of freedom, with mean:

$$E\left(e_i|\mathbf{y}\right) = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \tag{6.15}$$

and variance equal to (6.14). The topic of model fit based on analysis of residuals is taken up again in Chapter 8.

Predictive Distributions

In Bayesian analysis, a distinction needs to be made between two predictive distributions that play different roles. The first one is the prior predictive distribution, which assigns densities or probabilities to the data points before these are observed. If a Bayesian model postulates that the data are generated according to the process $[\mathbf{y}|\boldsymbol{\theta}]$, and if $\boldsymbol{\theta}$ has a prior density indexed by hyperparameter H, then the prior predictive distribution is given by the mixture

$$\begin{aligned} p\left(\mathbf{y}|H\right) &= \int p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|H\right) d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}}\left[p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right], \end{aligned} \tag{6.16}$$

with the prior acting as mixing distribution. Hence, the prior predictive distribution results from averaging out the sampling model $[\mathbf{y}|\boldsymbol{\theta}]$ over all possible values that the parameter vector can take, with the relative weights

being the prior density at each value of $\boldsymbol{\theta}$. In other words, the prior predictive distribution gives the total probability of observing the actual data, unconditionally on parameter values, but given H.

The prior predictive distribution does not exist unless the prior is proper; otherwise, the integral in (6.16) would not be defined. Hence, there is no prior predictive distribution for the Bayesian regression model with an unbounded flat prior. When it exists, as will be the case with the second set of priors discussed later on, the predictive distribution provides a basis for model comparison. We shall return to this later but, for the moment, consider the following statement:

> "One ought to be inclined towards choosing a model over a competing one if the former predicts (before data collection) the observations that will occur with a higher probability than the latter."

On the other hand, the posterior predictive process is the distribution or density of future observations, given past data, and unconditionally with respect to parameter values. Thus, a natural application of this distribution is in forecasting problems. The distribution will exist whenever the posterior distribution of the parameters is proper. In the context of the linear regression model, suppose that one wishes to forecast a vector of future observations, $\mathbf{y}_f$, of order $n_f \times 1$, under the assumption that these will be generated according to model (6.1). Since $\mathbf{y}_f$ is unobservable, it can be included as an unknown in Bayes theorem, with this being a particular case of the technique called "data augmentation" (Tanner and Wong, 1987), discussed in Chapter 11. The joint posterior density of all unknowns is:

$$\begin{aligned} p\left(\mathbf{y}_f, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) &= p\left(\mathbf{y}_f | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2, \mathbf{y}\right) p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) \\ &= p\left(\mathbf{y}_f | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2\right) p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right), \end{aligned}$$

because, given the parameters, the future and current observations are mutually independent. The distribution $\left[\mathbf{y}_f | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2\right]$ is as postulated by the sampling model, that is,

$$\mathbf{y}_f | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 \sim N\left(\mathbf{X}_f \boldsymbol{\beta}, \mathbf{I}_f \sigma^2\right),$$

where $\mathbf{X}_f$ and $\mathbf{I}_f$ have suitable dimensions. The posterior predictive density is then

$$p\left(\mathbf{y}_f | \mathbf{y}\right) = \int \int \int p\left(\mathbf{y}_f | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2\right) p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}\right) d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2 \, d\sigma^2. \quad (6.17)$$

Thus, this distribution is a mixture of the sampling model for the future observations using the joint posterior distribution of the parameters (based

on past observations) as a mixing process. Now note that the sampling model for future observations implies that

$$\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \mathbf{e}_f,$$

where the future errors have the distribution $\mathbf{e}_f \sim N\left(\mathbf{0}, \mathbf{I}_f\sigma^2\right)$. The mean of the posterior predictive distribution is then

$$E\left(\mathbf{y}_f|\mathbf{y}\right) = E\left[E\left(\mathbf{y}_f|\boldsymbol{\beta}\right)\right] = E\left(\mathbf{X}_f\boldsymbol{\beta}\right) = \mathbf{X}_f\widehat{\boldsymbol{\beta}}, \tag{6.18}$$

where the outer expectation is taken with respect to the posterior distribution. The variance–covariance matrix of the predictive distribution, using a similar conditioning and deconditioning (outer expectations and variances taken over the posterior distribution of the parameters), is:

$$\begin{aligned} Var\left(\mathbf{y}_f|\mathbf{y}\right) &= Var\left[E\left(\mathbf{y}_f|\boldsymbol{\beta},\sigma^2\right)\right] + E\left[Var\left(\mathbf{y}_f|\boldsymbol{\beta},\sigma^2\right)\right] \\ &= \mathbf{X}_f\, Var\left(\boldsymbol{\beta}|\mathbf{y}\right)\mathbf{X}_f' + \mathbf{I}_f E\left(\sigma^2|\mathbf{y}\right) \\ &= \left(\mathbf{X}_f\mathbf{C}^{-1}\mathbf{X}_f' + \mathbf{I}_f\right)\frac{S_e}{(n-p_1-p_2-4)}. \end{aligned} \tag{6.19}$$

In order to specify completely the predictive distribution, return to (6.17) and rewrite it as

$$\begin{aligned} p\left(\mathbf{y}_f|\mathbf{y}\right) = \int\int\int p\left(\mathbf{y}_f|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma^2\right) \\ \times p\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2|\sigma^2,\mathbf{y}\right)p\left(\sigma^2|\mathbf{y}\right)d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2\, d\sigma^2. \end{aligned} \tag{6.20}$$

The first two densities under the integral sign are in normal forms. Now the quadratics in the exponents involve the regression coefficients and can be put as follows (we shall not distinguish between the two sets of regressions here)

$$\begin{aligned} &\left(\mathbf{y}_f - \mathbf{X}_f\boldsymbol{\beta}\right)'\left(\mathbf{y}_f - \mathbf{X}_f\boldsymbol{\beta}\right) + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)'\mathbf{C}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) \\ &= \left(\mathbf{y}_f - \mathbf{X}_f\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}_f - \mathbf{X}_f\widehat{\boldsymbol{\beta}}_f\right) + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right)'\mathbf{X}_f'\mathbf{X}_f\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right) \\ &\quad + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)'\mathbf{C}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) \\ &= S_{e_f} + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right)'\mathbf{X}_f'\mathbf{X}_f\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right) \\ &\quad + \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)'\mathbf{C}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right), \end{aligned} \tag{6.21}$$

where

$$\widehat{\boldsymbol{\beta}}_f = \left(\mathbf{X}_f'\mathbf{X}_f\right)^{-1}\mathbf{X}_f'\mathbf{y}_f$$

and

$$S_{e_f} = \left(\mathbf{y}_f - \mathbf{X}_f\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}_f - \mathbf{X}_f\widehat{\boldsymbol{\beta}}\right).$$

The first term in (6.21) does not involve the regression coefficients. The second and third terms can be combined, using (5.56) and (5.57), as

$$\begin{aligned}&\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_f\right)' \mathbf{X}_f'\mathbf{X}_f \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_f\right) + \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)' \mathbf{C} \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\\ &= \left(\boldsymbol{\beta}-\widehat{\widehat{\boldsymbol{\beta}}}_f\right)' \left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right) \left(\boldsymbol{\beta}-\widehat{\widehat{\boldsymbol{\beta}}}_f\right)\\ &+ \left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)' \mathbf{X}_f'\mathbf{X}_f \left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{X} \left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right). \qquad (6.22)\end{aligned}$$

Here,

$$\begin{aligned}\widehat{\widehat{\boldsymbol{\beta}}}_f &= \left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}_f'\mathbf{X}_f\widehat{\boldsymbol{\beta}}_f+\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}\right)\\ &= \left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}_f'\mathbf{y}_f+\mathbf{X}'\mathbf{y}\right).\end{aligned}$$

Let

$$\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}=\mathbf{C}_+.$$

Using (6.22) in (6.21) and, further, employing the ensuing result in density (6.20), one gets

$$\begin{aligned}p\left(\mathbf{y}_f|\mathbf{y}\right) \propto \int \Bigg\{ &\left(\sigma^2\right)^{-\left(\frac{n_f+p_1+p_2}{2}\right)} \exp\left[-\frac{S_{e_f}+(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}})'\mathbf{X}_f'\mathbf{X}_f\mathbf{C}_+^{-1}\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}})}{2\sigma^2}\right]\\ &\times \int \exp\left[-\frac{\left(\boldsymbol{\beta}-\widehat{\widehat{\boldsymbol{\beta}}}_f\right)'\mathbf{C}_+\left(\boldsymbol{\beta}-\widehat{\widehat{\boldsymbol{\beta}}}_f\right)}{2\sigma^2}\right] d\boldsymbol{\beta}\Bigg\} p\left(\sigma^2|\mathbf{y}\right) d\sigma^2.\end{aligned}$$

The last integral involves a Gaussian kernel and is equal to

$$(2\pi)^{\frac{p_1+p_2}{2}} \left|\mathbf{C}_+^{-1}\sigma^2\right|^{\frac{1}{2}} = \left(2\pi\sigma^2\right)^{\frac{p_1+p_2}{2}} \left|\mathbf{C}_+^{-1}\right|^{\frac{1}{2}}.$$

Making use of this, the predictive density can be written as

$$\begin{aligned}&p\left(\mathbf{y}_f|\mathbf{y}\right)\\ &\propto \int \left(\sigma^2\right)^{-\frac{n_f}{2}} \exp\left[-\frac{S_{e_f}+\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)'\mathbf{X}_f'\mathbf{X}_f\mathbf{C}_+^{-1}\mathbf{X}'\mathbf{X}\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)}{2\sigma^2}\right]\\ &\times p\left(\sigma^2|\mathbf{y}\right) d\sigma^2. \qquad (6.23)\end{aligned}$$

Writing the marginal density of σ^2 explicitly, as shown in (6.11), and letting

$$\mathbf{Q}_+=\mathbf{X}_f'\mathbf{X}_f\mathbf{C}_+^{-1}\mathbf{X}'\mathbf{X},$$

the predictive density is expressible as

$$p\left(\mathbf{y}_f|\mathbf{y}\right) \propto \int \left(\sigma^2\right)^{-\left(\frac{n_f+n-p_1-p_2-2}{2}+1\right)} \times \exp\left[-\frac{S_{e_f}+\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Q}_+\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)+S_e}{2\sigma^2}\right] d\sigma^2.$$

The integrand is the kernel of a scaled inverse chi-square (or inverted gamma) density, and this type of integral has been encountered a number of times before. Upon integrating over σ^2, one obtains

$$p\left(\mathbf{y}_f|\mathbf{y}\right) \propto \left[S_{e_f}+\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Q}_+\left(\widehat{\boldsymbol{\beta}}_f-\widehat{\boldsymbol{\beta}}\right)+S_e\right]^{-\left(\frac{n_f+n-p_1-p_2-2}{2}\right)}. \tag{6.24}$$

After lengthy matrix manipulations (see, e.g., Zellner, 1971), the posterior predictive density can be put in the form

$$p\left(\mathbf{y}_f|\mathbf{y}\right) \propto \left[1+\frac{\left(\mathbf{y}_f-\mathbf{X}_f\widehat{\boldsymbol{\beta}}\right)'\mathbf{P}_f\left(\mathbf{y}_f-\mathbf{X}_f\widehat{\boldsymbol{\beta}}\right)}{S_e}\right]^{-\left(\frac{n_f+\eta}{2}\right)}, \tag{6.25}$$

where $\eta = n-p_1-p_2-2$, and

$$\mathbf{P}_f = \mathbf{I}-\mathbf{X}_f\left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_f.$$

This indicates that the distribution is a multivariate-t process of order n_f, having mean vector $\mathbf{X}_f\widehat{\boldsymbol{\beta}}$, η degrees of freedom, and variance–covariance matrix

$$\frac{S_e}{n-p_1-p_2-4}\left[\mathbf{I}-\mathbf{X}_f\left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_f\right]^{-1}. \tag{6.26}$$

In order to show that (6.19) and (6.26) are equal, use can be made of the matrix identity

$$\left(\mathbf{A}+\mathbf{BGB}'\right)^{-1} = \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}+\mathbf{G}^{-1}\right)^{-1}\mathbf{B}'\mathbf{A}^{-1} \tag{6.27}$$

provided that the inverses involved exist. Then

$$\left(\mathbf{X}_f\mathbf{C}^{-1}\mathbf{X}_f'+\mathbf{I}_f\right)^{-1} = \mathbf{I}_f-\mathbf{X}_f\left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_f',$$

recalling that $\mathbf{C}=\mathbf{X}'\mathbf{X}$. This implies that, in (6.26),

$$\left[\mathbf{I}-\mathbf{X}_f\left(\mathbf{X}_f'\mathbf{X}_f+\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_f\right]^{-1} = \mathbf{X}_f\mathbf{C}^{-1}\mathbf{X}_f'+\mathbf{I}_f,$$

which is the matrix expression entering into (6.19). There is a much simpler way of arriving at (6.25), which will be presented in the following section on inference under conjugate priors.

6.2.2 *Inference under Conjugate Priors*

Suppose now that a scaled inverse chi-square prior is adopted for σ^2 and that a multivariate normal prior is assigned to $\boldsymbol{\beta}$, as follows (again, let H be a set of hyperparameters):

$$p\left(\boldsymbol{\beta}, \sigma^2 | H\right) = p\left(\boldsymbol{\beta} | \sigma^2, H_{\boldsymbol{\beta}}\right) p\left(\sigma^2 | H_{\sigma^2}\right).$$

Here H_{σ^2} indexes the scaled inverse chi-square process, and $H_{\boldsymbol{\beta}}$ is the set of hyperparameters of the multivariate normal distribution for $\boldsymbol{\beta}$, given σ^2. Write

$$p\left(\sigma^2 | H_{\sigma^2} = \nu^*, s^{*2}\right) \propto \left(\sigma^2\right)^{-\left(\frac{\nu^*}{2}+1\right)} \exp\left(-\frac{\nu^* s^{*2}}{2\sigma^2}\right),$$

and

$$p\left(\boldsymbol{\beta} | \sigma^2, H_{\boldsymbol{\beta}} = \mathbf{m}_\beta, \mathbf{V}_\beta\right) \propto \left|\mathbf{V}_\beta \sigma^2\right|^{-\frac{1}{2}} \exp\left[-\frac{\left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)' \mathbf{V}_\beta^{-1} \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)}{2\sigma^2}\right],$$

where $\mathbf{m}_\beta$ is the prior mean and $\mathbf{V}_\beta \sigma^2$ is the covariance matrix of the conditional (given σ^2) prior distribution. The joint prior density is then

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 | H\right) &\propto \left(\sigma^2\right)^{-\left(\frac{\nu^* + p_1 + p_2}{2}+1\right)} \\ \times \exp\left[-\frac{\left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)' \mathbf{V}_\beta^{-1} \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right) + \nu^* s^{*2}}{2\sigma^2}\right]. \end{aligned} \tag{6.28}$$

Using the likelihood function in (6.2), in conjunction with the prior given above, yields as joint posterior density of all unknown parameters

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, H\right) &\propto \left(\sigma^2\right)^{-\left(\frac{n + \nu^* + p_1 + p_2}{2}+1\right)} \\ \times \exp\left[-\frac{S_\beta + \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)' \mathbf{V}_\beta^{-1} \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right) + S_e + \nu^* s^{*2}}{2\sigma^2}\right]. \end{aligned} \tag{6.29}$$

Now the quadratic forms in $\boldsymbol{\beta}$ can be combined, using (5.56) and (5.57), as

$$\begin{aligned} & S_\beta + \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)' \mathbf{V}_\beta^{-1} \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right) \\ &= \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)' \mathbf{C} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right)' \mathbf{V}_\beta^{-1} \left(\boldsymbol{\beta} - \mathbf{m}_\beta\right) \\ &= \left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\right)' \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\right) + S_{\widehat{\boldsymbol{\beta}}}, \end{aligned}$$

where

$$\begin{aligned} \overline{\boldsymbol{\beta}} &= \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}} + \mathbf{V}_\beta^{-1}\mathbf{m}_\beta\right) \\ &= \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \left(\mathbf{X}'\mathbf{y} + \mathbf{V}_\beta^{-1}\mathbf{m}_\beta\right), \end{aligned} \tag{6.30}$$

and

$$S_{\widehat{\boldsymbol{\beta}}} = \left(\widehat{\boldsymbol{\beta}} - \mathbf{m}_\beta\right)' \mathbf{C} \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right) \mathbf{V}_\beta^{-1} \left(\widehat{\boldsymbol{\beta}} - \mathbf{m}_\beta\right).$$

Employing this in (6.29), the joint posterior density is expressible as

$$p\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, H\right) \propto \left(\sigma^2\right)^{-\left(\frac{n+\nu^*+p_1+p_2}{2}+1\right)}$$
$$\times \exp\left[-\frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)' \left(\mathbf{C}+\mathbf{V}_\beta^{-1}\right)\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right) + S_{\widehat{\boldsymbol{\beta}}} + S_e + \nu^* s^{*2}}{2\sigma^2}\right]. \quad (6.31)$$

Note that this is in the same mathematical form as the joint prior (6.28). This property is called conjugacy, meaning that the process remains in the same family of distributions, a posteriori, so only the parameters need to be updated, as the posterior has the same form as the prior. This implies that the marginal posterior distribution of σ^2 must be scaled inverted chi-square, and that the conditional posterior distribution of $\boldsymbol{\beta}$, given σ^2, must be multivariate normal. We now proceed to verify that this is so.

Conditional Posterior Distribution of the Regression Coefficients

This can be found via the usual procedure of examining the joint posterior density, fixing some parameters, and then letting those whose distribution is sought, vary. If (6.31) is viewed as a function of $\boldsymbol{\beta}$, it is clear that the conditional posterior distribution of the regression coefficients, given σ^2, is the multivariate normal process

$$\boldsymbol{\beta} | \sigma^2, \mathbf{y}, H \sim N\left(\overline{\boldsymbol{\beta}}, \left(\mathbf{X}'\mathbf{X} + \mathbf{V}_\beta^{-1}\right)^{-1} \sigma^2\right). \quad (6.32)$$

Hence, it is verified that conditional posterior has the same form as the conditional prior distribution, except that the mean vector and covariance matrix have been modified as a consequence of learning from the data. Since the stochastic process is multivariate normal, all posterior distributions of either individual or sets of regression coefficients are normal as well, given σ^2. The means and (co)variances of conditional posterior distributions of regression coefficients, given some other such coefficients, can be arrived at using the formulas given in Chapter 1.

It is instructive writing (6.30) as

$$\overline{\boldsymbol{\beta}} = \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \mathbf{V}_\beta^{-1} \mathbf{m}_\beta + \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \mathbf{C}\widehat{\boldsymbol{\beta}}.$$

Following O'Hagan (1994), now let

$$\begin{aligned} \mathbf{W} &= \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \mathbf{C}, \\ \mathbf{I} - \mathbf{W} &= \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1} \mathbf{V}_\beta^{-1}, \end{aligned}$$

so the posterior mean becomes

$$\begin{aligned}\overline{\boldsymbol{\beta}} &= (\mathbf{I}-\mathbf{W})\,\mathbf{m}_\beta + \mathbf{W}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{m}_\beta + \mathbf{W}\left(\widehat{\boldsymbol{\beta}} - \mathbf{m}_\beta\right).\end{aligned}$$

This representation indicates that the posterior mean is a matrix-weighted average of the prior mean $\mathbf{m}_\beta$ and of the ML estimator $\widehat{\boldsymbol{\beta}}$, with the latter receiving more weight as $\mathbf{W}$ increases, in some sense. When the information in the data is substantial relative to the prior information, so that $\mathbf{C}$ is much larger than $\mathbf{V}_\beta^{-1}$, then

$$\mathbf{W} = \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1}\mathbf{C} \to \mathbf{I}.$$

The implication is that the prior mean is overwhelmed by the ML estimator, as the former receives an effective weight of $\mathbf{0}$. On the other hand, when "prior precision" (the inverse of the covariance matrix) is large relative to the information contributed by the data, then $\mathbf{W} \to \mathbf{0}$, as $\mathbf{V}_\beta$ has "small" elements. In this case, the posterior mean would be expected to be very close to the prior mean, indicating a mild modification of prior opinions about the value of $\boldsymbol{\beta}$, after having observed the data.

Similarly, using (6.27), the posterior covariance matrix can be written as

$$\left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1}\sigma^2 = \left[\mathbf{V}_\beta - \mathbf{V}_\beta\left(\mathbf{V}_\beta + \mathbf{C}^{-1}\right)^{-1}\mathbf{V}_\beta\right]\sigma^2.$$

The representation on the right-hand side illustrates that the posterior variances are always smaller than or equal to the prior variances, at least when the prior and posterior distributions are normal. When $\mathbf{V}_\beta^{-1} \to \mathbf{0}$, so prior information becomes increasingly diffuse (large elements of $\mathbf{V}_\beta^{-1}$), the posterior covariance tends to $\mathbf{C}^{-1}\sigma^2$, which is in the same form as the variance covariance matrix of the ML estimator; the same occurs when the information in the data is very large relative to that contributed by the prior distribution. Hence, we see that either when prior information becomes relatively weaker and weaker, or when the relative contribution of the prior to knowledge about the regression coefficients is much smaller than that made by the data, then the conditional posterior distribution tends to

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim N\left(\widehat{\boldsymbol{\beta}}, \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2\right). \tag{6.33}$$

Hence, in the limit, our conditional posterior distribution is centered at the ML estimator, and the posterior covariance matrix is identical in form to the asymptotic covariance matrix of the ML estimator. This is a particular case of a more general result on asymptotic approximations to posterior distributions under regularity conditions and will be discussed in Chapter 7. A technical point must be highlighted here: recall that finding the asymptotic covariance matrix of the ML estimator requires taking expectations

over conceptual repeated sampling. However, Fisher's information matrix (see Chapter 3) does not play the same role in Bayesian asymptotics; this is because the paradigm does not involve repeated sampling over the joint distribution $[\boldsymbol{\theta}, \mathbf{y}]$. It will be seen that in Bayesian asymptotics, the observed information plays a role similar to that of expected information in ML estimation. In the specific case of the linear regression model discussed here, the matrix of negative second derivatives of the log-posterior with respect to $\boldsymbol{\beta}$ does not involve the observations. Thus, the observed information is a constant and, therefore, is equal to its expected value taken over the distribution of the data. However, this would retrieve the form of the asymptotic variance–covariance matrix of the ML estimator only when $\mathbf{C}$ overwhelms $\mathbf{V}_{\beta}^{-1}$.

Example 6.1 *Ridge regression from Bayesian and frequentist points of view*

Suppose the conditional prior of the regressions has the form

$$\left[\begin{array}{c} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{array} \right] \Bigg| \sigma^2, H_{\boldsymbol{\beta}} \sim N \left(\left[\begin{array}{c} \mathbf{m}_1 \\ \mathbf{m}_2 \end{array} \right], \left[\begin{array}{cc} \mathbf{I}\frac{\sigma^2_{\beta_1}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\frac{\sigma^2_{\beta_2}}{\sigma^2} \end{array} \right] \sigma^2 \right),$$

so the two sets of coefficients are independent, a priori. Then the mean of the conditional posterior distribution of the regression coefficients, using (6.30) and (6.32), is

$$\left[\begin{array}{c} \overline{\boldsymbol{\beta}}_1 \\ \overline{\boldsymbol{\beta}}_2 \end{array} \right] = \left[\begin{array}{cc} \mathbf{X}_1'\mathbf{X}_1 + \mathbf{I}\frac{\sigma^2}{\sigma^2_{\beta_1}} & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 + \mathbf{I}\frac{\sigma^2}{\sigma^2_{\beta_2}} \end{array} \right]^{-1} \left[\begin{array}{c} \mathbf{X}_1'\mathbf{y} + \mathbf{m}_1\frac{\sigma^2}{\sigma^2_{\beta_1}} \\ \mathbf{X}_1'\mathbf{y} + \mathbf{m}_2\frac{\sigma^2}{\sigma^2_{\beta_2}} \end{array} \right].$$

When there is a single set of regression coefficients and when the prior mean is a null vector, this reduces to

$$\overline{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}k\right)^{-1}\mathbf{X}'\mathbf{y},$$

where

$$k = \frac{\sigma^2}{\sigma^2_{\beta}}$$

is a positive scalar. In the regression literature, the linear function of the data $\overline{\boldsymbol{\beta}}$ is known as the "ridge regression estimator", after Hoerl and Kennard (1970). Seemingly, it did not evolve from a Bayesian argument. Now, using the notation where the posterior mean is expressed as a weighted average of the ML estimator and of the prior mean (null in this case), one can write symbolically

$$\begin{aligned} \overline{\boldsymbol{\beta}} &= \mathbf{0} + \mathbf{W}\left(\widehat{\boldsymbol{\beta}} - \mathbf{0}\right) \\ &= \left(\mathbf{X}'\mathbf{X} + \mathbf{I}k\right)^{-1}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}. \end{aligned}$$

Hence, the ridge regression "shrinks" the ML estimator toward zero, with a strength that depends on the value of k, the ratio between the two "variance components". From a Bayesian perspective, if $\sigma^2_\beta \to \infty$, indicating a vague prior opinion about the value of the regression coefficients, then there is little shrinkage, as k is near 0. Here the ridge estimator approaches the ML estimator. Thus, if $\mathbf{X}'\mathbf{X}$ is nearly singular, as in models with extreme collinearity, the posterior distribution is also nearly singular when $k = 0$. In this case, there would be extremely large posterior variances, as the diagonal elements of

$$(\mathbf{X}'\mathbf{X} + \mathbf{I}k)^{-1} \to \infty.$$

On the other hand, when $\sigma^2_\beta \to 0$, that is, in a situation where prior information is very precise, then the ridge regression estimator tends toward the prior mean, which is null. Hence, by incorporating prior information, i.e., by increasing the value of k, a nearly singular distribution becomes better conditioned. However, the "improvement" in condition does not depend on the data; rather, it is a consequence of bringing external information into the picture. Although this has a natural interpretation in a Bayesian context, the arguments are less transparent from a frequentist perspective.

For non-Bayesian descriptions of the ridge estimator, see Bunke (1975), Bibby and Toutenburg (1977), and Toutenburg (1982). In particular, and from a frequentist point of view, the ridge regression estimator has a Gaussian distribution, since it is a linear combination of Gaussian observations. Its mean value, taking expectations over the sampling model, is

$$\begin{aligned} E_{y|\beta,\sigma^2}\left[\left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}\right] &= \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta} - \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\frac{\sigma^2}{\sigma^2_\beta}\boldsymbol{\beta}. \end{aligned}$$

Thus, the ridge regression estimator is biased for $\boldsymbol{\beta}$, with the bias vector being

$$-\left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\frac{\sigma^2}{\sigma^2_\beta}\boldsymbol{\beta},$$

which goes to $\mathbf{0}$ as σ^2_β goes to infinity. The variance–covariance matrix of its sampling distribution is

$$\begin{aligned} & Var_{y|\beta,\sigma^2}\left[\left(\mathbf{X}'\mathbf{X}+\mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}\right] \\ = & \left(\mathbf{X}'\mathbf{X}+\mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\mathbf{I}\frac{\sigma^2}{\sigma^2_\beta}\right)^{-1}\sigma^2. \end{aligned}$$

What can be gained by using the biased estimator, instead of the ML statistic? The answer resides in the possibility of attaining a smaller mean squared error of estimation. The mean squared error matrix, that is, the sum of the covariance matrix plus the product of the bias vector times is transpose, is

$$\mathbf{M}(\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}\left[\boldsymbol{\beta}\boldsymbol{\beta}'k^2+\mathbf{X}'\mathbf{X}\sigma^2\right](\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}.$$

A "global" measure of the squared error of estimation is given by the sum of the diagonal elements of the mean squared error matrix

$$\begin{aligned} tr\left[\mathbf{M}(\boldsymbol{\beta})\right] &= tr\left[(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'k^2(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}\right] \\ &\quad + tr\left[(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}\mathbf{X}'\mathbf{X}\sigma^2(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-1}\right] \\ &= k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-2}\boldsymbol{\beta}+\sigma^2\, tr\left[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+\mathbf{I}k)^{-2}\right], \end{aligned}$$

after appropriate cyclical commutation of matrices under the trace operator. When $k = 0$ (no shrinkage at all), then

$$tr\left[\mathbf{M}(\boldsymbol{\beta})\right] = \sigma^2\, tr\left[(\mathbf{X}'\mathbf{X})^{-1}\right],$$

as one would expect, as then the global mean squared error is the sum of the sampling variances of the ML estimates of individual regression coefficients. On the other hand, when $k \to \infty$, that is, when there is strong shrinkage toward 0, then $tr\left[\mathbf{M}(\boldsymbol{\beta})\right]$ goes to $\boldsymbol{\beta}'\boldsymbol{\beta}$.

In order to illustrate, consider the case of a single parameter specification, for example, a model with a regression line going through the origin. Here the model is

$$y_i = \beta x_i + e_i.$$

The ridge regression estimator takes the simple form

$$\overline{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + k}.$$

Its expectation and sampling variance are

$$E_{y|\beta,\sigma^2}\left(\overline{\beta}\right) = \frac{\beta}{1+\frac{k}{\sum_{i=1}^{n} x_i^2}},$$

and

$$Var_{y|\beta,\sigma^2}\left(\overline{\beta}\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 \left(1+\frac{k}{\sum_{i=1}^{n} x_i^2}\right)^2},$$

respectively. Clearly, the ridge estimator (or posterior mean of β for the prior under discussion) is less variable than the ML estimator. The mean squared error, after rearrangement, is

$$M(\beta) = \frac{\beta^2 k^2 + \sigma^2 \left(\sum_{i=1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)^2 \left(1+\frac{k}{\sum_{i=1}^{n} x_i^2}\right)^2}.$$

Viewed as a function of k, at fixed β, the mean squared error approaches $\sigma^2/\left(\sum_{i=1}^{n} x_i^2\right)$ when k approaches 0, and β^2 when k approaches ∞. A question of interest is whether there is a value of k making $M(\beta)$ minimum. Taking derivatives of the logarithm of $M(\beta)$ with respect to k and then solving for the "optimum" value yields

$$k(\beta) = \frac{\sigma^2}{\beta^2}.$$

However, β is unknown, so the optimum k must be estimated. An intuitively appealing procedure is given by the following iterative algorithm. Set the functional iteration

$$\beta^{[i+1]} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \frac{\sigma^2}{\left(\beta^{[i]}\right)^2}}.$$

Start with $\beta^{[0]}$ equal to the ML estimator, and then iterate until values stabilize. In practice, σ^2 must be estimated as well, and natural candidates are the ML or the REML estimators of the variance for this regression model. The frequentist properties of this "empirical minimum mean squared error estimator" are difficult to evaluate analytically. ■

Conditional Posterior Distribution of σ^2

If one regards the joint density (6.29) as a function of σ^2, with the regression coefficients fixed, then it is clear that the conditional posterior distribution of the variance is a scaled inverse chi-square process with degree of belief parameter equal to $n + \nu^* + p_1 + p_2$, and with mean value and variance given by

$$E\left(\sigma^2|\boldsymbol{\beta},\mathbf{y},H\right) = \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+\left(\boldsymbol{\beta}-\mathbf{m}_\beta\right)'\mathbf{V}_\beta^{-1}\left(\boldsymbol{\beta}-\mathbf{m}_\beta\right)+\nu^* s^{*2}}{n+\nu^*+p_1+p_2-2},$$

and

$$Var\left(\sigma^2|\boldsymbol{\beta},\mathbf{y},H\right) = \frac{2\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+\left(\boldsymbol{\beta}-\mathbf{m}_\beta\right)'\mathbf{V}_\beta^{-1}\left(\boldsymbol{\beta}-\mathbf{m}_\beta\right)+\nu^* s^{*2}\right]^2}{\left(n+\nu^*+p_1+p_2-2\right)^2\left(n+\nu^*+p_1+p_2-4\right)},$$

respectively.

Marginal Distribution of the Regression Coefficients

Using properties of the inverse gamma distribution, the joint density in (6.31) can be integrated over σ^2 to obtain the following explicit form as marginal density of the regression coefficients

$$\begin{aligned} p\left(\boldsymbol{\beta}|\mathbf{y},H\right) &\propto \int \left(\sigma^2\right)^{-\left(\frac{n+\nu^*+p_1+p_2}{2}+1\right)} \\ &\times \exp\left[-\frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'\left(\mathbf{C}+\mathbf{V}_\beta^{-1}\right)\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)+S_{\widehat{\boldsymbol{\beta}}}+S_e+\nu^* s^{*2}}{2\sigma^2}\right] d\sigma^2 \\ &\propto \left[\frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'\left(\mathbf{C}+\mathbf{V}_\beta^{-1}\right)\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)+S_{\widehat{\boldsymbol{\beta}}}+S_e+\nu^* s^{*2}}{2}\right]^{-\frac{n+\nu^*+p_1+p_2}{2}} \\ &\propto \left[1+\frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'\left(\mathbf{C}+\mathbf{V}_\beta^{-1}\right)\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)}{\left(n+\nu^*\right)\left(\frac{S_{\widehat{\boldsymbol{\beta}}}+S_e+\nu^* s^{*2}}{n+\nu^*}\right)}\right]^{-\frac{p_1+p_2+n+\nu^*}{2}}. \end{aligned} \tag{6.34}$$

This is the kernel of a multivariate-t density function of order $p_1 + p_2$. The corresponding distribution has mean vector $\overline{\boldsymbol{\beta}}$, covariance matrix

$$Var\left(\boldsymbol{\beta}|\mathbf{y},H\right) = \frac{S_{\widehat{\boldsymbol{\beta}}}+S_e+\nu^* s^{*2}}{n+\nu^*-2}\left(\mathbf{C}+\mathbf{V}_\beta^{-1}\right)^{-1},$$

and $n+\nu^*$ degrees of freedom. Thus, all marginal distributions of individual or of subsets of regression coefficients are either univariate or multivariate-t.

The same holds for any linear combination or any conditional distribution of the regression coefficients, given some other coefficients.

As a side point, note that at first sight there does not seem to be a "loss of degrees of freedom", relative to n (sample size), in the process of accounting for uncertainty about the variance. In fact, however, there is a "hidden loss" of $p_1 + p_2$ degrees of freedom, which is canceled by the contribution made by the conditional prior of the regression coefficients, which involves σ^2; see (6.28). If the prior for $\boldsymbol{\beta}$ had not involved σ^2, with $\mathbf{V}_\beta$ then being the prior variance–covariance matrix (assumed known), then the degrees of freedom of the marginal posterior distribution of the regression coefficients would have been $n + \nu^* - p_1 - p_2$. If, in addition, the "degree of belief" parameter of the prior distribution for σ^2 had been taken to be $\nu^* = 0$, the degrees of freedom would have been $n - p_1 - p_2$, that is, the number of degrees of freedom arising in a standard classical analysis of linear regression. Note, however, that the Bayesian assignment $\nu^* = 0$ produces the improper prior distribution

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2},$$

which is some sort of "noninformative" prior. In this case, the posterior distribution of the regression coefficients would be proper if $n - p_1 - p_2 > 0$.

Marginal Distribution of σ^2

Integrating the joint density (6.31) with respect to the regressions, to obtain the marginal density of the variance, gives:

$$\begin{aligned} p\left(\sigma^2|\mathbf{y}, H\right) &\propto \left(\sigma^2\right)^{-\left(\frac{n+\nu^*+p_1+p_2}{2}+1\right)} \exp\left[-\frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{2\sigma^2}\right] \\ &\times \int\int \exp\left[-\frac{\left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\right)' \left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right) \left(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\right)}{2\sigma^2}\right] d\boldsymbol{\beta} \\ &\propto \left(\sigma^2\right)^{-\left(\frac{n+\nu^*}{2}+1\right)} \exp\left[-\frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{2\sigma^2}\right]. \end{aligned} \tag{6.35}$$

Hence, the posterior process is a scaled inverted chi-square distribution, with mean

$$E\left(\sigma^2|\mathbf{y}, H\right) = \frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{n + \nu^* - 2},$$

mode equal to

$$Mode\left(\sigma^2|\mathbf{y}, H\right) = \frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{n + \nu^* + 2},$$

and variance

$$Var\left(\sigma^2|\mathbf{y}, H\right) = \frac{2\left(S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}\right)^2}{\left(n + \nu^* - 2\right)^2 \left(n + \nu^* - 4\right)}.$$

Again, it is verified that the posterior distribution of σ^2 has the same form as the prior process, because of the conjugacy property mentioned earlier. Bayesian learning updates the parameter ν^* in the prior distribution to $n + \nu^*$, posterior to the data. Similarly, the parameter s^{*2} becomes

$$\frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{n + \nu^*}$$

in the posterior distribution.

Posterior Distribution of the Residuals

Using a similar argument to that employed when inference under improper uniform priors was discussed, the posterior distribution of a residual is also univariate-t, with degrees of freedom $n + \nu^*$, mean value equal to

$$E\left(e_i|\mathbf{y}, \mathbf{H}\right) = y_i - \mathbf{x}_i'\overline{\boldsymbol{\beta}},$$

and variance

$$Var\left(e_i|\mathbf{y}, \mathbf{H}\right) = \frac{S\widehat{\boldsymbol{\beta}} + S_e + \nu^* s^{*2}}{n + \nu^* - 2}\mathbf{x}_i'\left(\mathbf{C} + \mathbf{V}_\beta^{-1}\right)^{-1}\mathbf{x}_i.$$

Predictive Distributions

In the regression model with proper priors, the two predictive distributions exist. The prior predictive distribution can be arrived at directly by taking expectations of the model over the joint prior distribution of the parameters. First, we take expectations, given σ^2, and then decondition with respect to the variance. In order to do this, note that the observations are a linear combination of the regression coefficients (which, given σ^2, have a multivariate normal prior distribution) and of the errors, with the latter being normal as well. Hence, it follows that, conditionally on σ^2, the prior predictive distribution is the Gaussian process

$$\mathbf{y}|\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{V}_\beta, \sigma^2 \sim N\left(\mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}, \left(\mathbf{X}\mathbf{V}_\beta\mathbf{X}' + \mathbf{I}\right)\sigma^2\right).$$

Since the prior distribution of σ^2 is scaled inverted chi-square, deconditioning the normal distribution above requires evaluating the integral

$$p\left(\mathbf{y}|\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{V}_\beta, s^{2*}, \nu^*\right) \propto \int \left(\sigma^2\right)^{-\frac{n+\nu^*}{2}+1}$$

$$\times \exp\left[-\frac{\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right)\left(\mathbf{X}\mathbf{V}_\beta\mathbf{X}' + \mathbf{I}\right)^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right) + \nu^* s^{*2}}{2\sigma^2}\right] d\sigma^2$$

$$\propto \left[\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right)\left(\mathbf{X}\mathbf{V}_\beta\mathbf{X}' + \mathbf{I}\right)^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right) + \nu^* s^{*2}\right]^{-\frac{n+\nu^*}{2}}$$

$$\propto \left[1 + \frac{\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right)\left(\mathbf{X}\mathbf{V}_\beta\mathbf{X}' + \mathbf{I}\right)^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}\right)}{\nu^* s^{*2}}\right]^{-\frac{n+\nu^*}{2}},$$

after making use of the results given in Chapter 1, as seen several times before. Hence, the prior predictive distribution is the multivariate-t process on ν^* degrees of freedom

$$\mathbf{y}|\mathbf{H} \sim t_n \left(\mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}, \nu^*, \frac{(\mathbf{X}\mathbf{V}_\beta\mathbf{X}' + \mathbf{I})\, s^{*2}\nu^*}{\nu^* - 2} \right). \tag{6.36}$$

The third term in the argument of the distribution is the variance–covariance matrix. Thus, (6.36) gives the probability distribution of the data before observation takes place. If a model assigns low probability to the observations that actually occur, this can be taken as evidence against such a model, so some revision would be needed. This issue will be elaborated further in the discussion about model comparisons.

The posterior predictive distribution applies to future observations, as generated by the model

$$\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \mathbf{e}_f.$$

The argument employed for the prior predictive distribution can also be used here. Conditionally on σ^2, the model for the future observations is a linear combination of normals, with $\boldsymbol{\beta}$ following the normal conditional posterior distribution given in (6.32). The future errors have the usual normal distribution, which is independent of the posterior distribution of $\boldsymbol{\beta}$, since the latter depends on past errors only. Then, given σ^2, the predictive distribution of future observations is the normal process

$$\mathbf{y}_f|\mathbf{y}, \sigma,^2 H \sim N \left(\mathbf{X}_f\overline{\boldsymbol{\beta}}, \mathbf{X}_f \left(\mathbf{X}'\mathbf{X} + \mathbf{V}_\beta^{-1} \right)^{-1} \mathbf{X}_f'\sigma^2 + \mathbf{I}_f\sigma^2 \right).$$

Next, we must decondition, but this time integrating over the marginal posterior distribution of σ^2, with the corresponding density given in (6.35). Using similar algebra to that employed for the prior predictive distribution, one arrives at

$$\mathbf{y}_f|\mathbf{y}, H \sim t \left(\mathbf{X}_f\overline{\boldsymbol{\beta}}, n + \nu^*, \frac{\mathbf{X}_f\left(\mathbf{X}'\mathbf{X}+\mathbf{V}_\beta^{-1}\right)^{-1}\mathbf{X}_f'\left(S_{\widehat{\boldsymbol{\beta}}}+S_e+\nu^* s^{*2}\right)}{n+\nu^*-2} \right), \tag{6.37}$$

with this distribution being an n_f-dimensional one. The predictions made by the model can then be contrasted with future observations in some test of predictive ability.

6.2.3 Orthogonal Parameterization of the Model

Consider again the linear model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}, \end{aligned} \tag{6.38}$$

where the residuals follow the usual normal process, and suppose the $\boldsymbol{\beta}'$s have been assigned some prior distribution. It may be that there is a strong intercorrelation between the ML estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In Bayesian analysis, this would probably be reflected in a high degree of correlation between parameters in the posterior distribution, unless sharp independent priors are adopted for each of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. A more comprehensive discussion of the role of the prior on inferences will be presented in the next chapter. For the time being, it suffices to recall (5.8) where it was seen that as data contribute more and more information, the influence of the prior vanishes asymptotically. At any rate, a strong posterior inter-correlation tends to hamper interpretation and to impair numerical behavior, as well as the performance of MCMC algorithms (these techniques, however, would probably have doubtful value in this simple linear model, since an analytical solution exists).

An alternative is to entertain an alternative parameterization such that the new model reproduces the same location and dispersion structure as (6.38). Ideally, it is desirable to render the new parameters independent from each other, either in the ML analysis (in which case uncorrelated ML estimates are obtained), or a posteriori, in the Bayesian setting. Under the usual normality assumption, with i.i.d. residuals, the ML estimator of $\boldsymbol{\beta}$ is

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}. \tag{6.39}$$

After eliminating $\widehat{\boldsymbol{\beta}}_1$ from the estimating equations, the ML estimator of $\boldsymbol{\beta}_2$ can be written as

$$\widehat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

where

$$\mathbf{M}_1 = \left[\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right].$$

Recall that this matrix is idempotent. The fitted residuals are given by

$$\begin{aligned} \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y} \\ &= \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right](\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\ &= \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{e}. \end{aligned}$$

Hence, the distribution of the fitted residuals is independent of the $\boldsymbol{\beta}$ parameters. The independence of the distribution of the residuals with respect to $\boldsymbol{\beta}$ is a consequence of the fact that

$$\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{X} = \mathbf{0},$$

and it is said that $\left[\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]$ and $\mathbf{X}$ are orthogonal to each other. In the context of (6.38), it follows that

$$\mathbf{M}_1\mathbf{X}_1=\left[\mathbf{I}-\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}'\right]\mathbf{X}_1=\mathbf{0}$$

and, consequently, that

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_1=\mathbf{X}_2'\left[\mathbf{I}-\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{X}_1=\mathbf{0}.$$

Hence, $\mathbf{X}_2'\mathbf{M}_1$ is orthogonal to $\mathbf{X}_1$. Then, for

$$\begin{aligned}\mathbf{W} &= [\mathbf{W}_1,\mathbf{W}_2]\\ &= [\mathbf{X}_1,\mathbf{M}_1\mathbf{X}_2],\end{aligned}$$

one has that

$$\mathbf{W}'\mathbf{W}=\begin{bmatrix}\mathbf{W}_1'\mathbf{W}_1 & \mathbf{W}_1'\mathbf{W}_2\\ \mathbf{W}_2'\mathbf{W}_1 & \mathbf{W}_2'\mathbf{W}_2\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\end{bmatrix}.$$

Now, if instead of (6.38), one fits the model

$$\mathbf{y}=\mathbf{W}_1\boldsymbol{\alpha}_1+\mathbf{W}_2\boldsymbol{\alpha}_2+\mathbf{e}, \tag{6.40}$$

it turns out the ML estimator of the new regression coefficients is

$$\begin{aligned}\begin{bmatrix}\widehat{\boldsymbol{\alpha}}_1\\ \widehat{\boldsymbol{\alpha}}_2\end{bmatrix} &= \begin{bmatrix}\mathbf{W}_1'\mathbf{W}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_2'\mathbf{W}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{W}_1'\mathbf{y}\\ \mathbf{W}_2'\mathbf{y}\end{bmatrix}\\ &= \begin{bmatrix}\left(\mathbf{W}_1'\mathbf{W}_1\right)^{-1}\mathbf{W}_1'\mathbf{y}\\ \left(\mathbf{W}_2'\mathbf{W}_2\right)^{-1}\mathbf{W}_2'\mathbf{y}\end{bmatrix}\\ &= \begin{bmatrix}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y}\\ \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}\end{bmatrix}.\end{aligned} \tag{6.41}$$

The first term of the vector (6.41) gives the ML estimator of $\boldsymbol{\beta}_1$ in a model ignoring $\boldsymbol{\beta}_2$, whereas the second term gives the ML estimator of $\boldsymbol{\beta}_2$ in the full model (6.38). Taking expectations of the ML estimators of the "new" parameters, with respect to the original model (6.38),

$$\begin{aligned}E\left(\widehat{\boldsymbol{\alpha}}_1\right) &= E\left[\left(\mathbf{W}_1'\mathbf{W}_1\right)^{-1}\mathbf{W}_1'\mathbf{y}\right]=\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'E\left(\mathbf{y}\right)\\ &= \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\left(\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2\right)\\ &= \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1+\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2\\ &= \boldsymbol{\beta}_1+\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.\end{aligned}$$

This reveals the parametric relationship

$$\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2. \tag{6.42}$$

Further,

$$E\left(\widehat{\boldsymbol{\alpha}}_2\right) = E\left[\left(\mathbf{W}_2'\mathbf{W}_2\right)^{-1}\mathbf{W}_2'\mathbf{y}\right] = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1 E\left(\mathbf{y}\right)$$
$$= \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_1\boldsymbol{\beta}_1 + \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2.$$

so that

$$\boldsymbol{\alpha}_2 = \boldsymbol{\beta}_2. \tag{6.43}$$

Hence, models (6.38) and (6.40) have the same expectation, since

$$\begin{aligned}\mathbf{W}_1\boldsymbol{\alpha}_1 + \mathbf{W}_2\boldsymbol{\alpha}_2 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 \\ &\quad - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2. \end{aligned} \tag{6.44}$$

Thus, the two models generate the same probability distribution of the observations, and this can be represented by the statement

$$\begin{aligned}&\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_1, \sigma^2 \sim N\left(\mathbf{W}_1\boldsymbol{\alpha}_1 + \mathbf{W}_2\boldsymbol{\alpha}_2, \mathbf{I}\sigma^2\right) \\ &\equiv \mathbf{y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 \sim N\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \mathbf{I}\sigma^2\right).\end{aligned}$$

Note that for model (6.40):

$$Var\left(\left[\begin{array}{c}\widehat{\boldsymbol{\alpha}}_1 \\ \widehat{\boldsymbol{\alpha}}_2\end{array}\right]\right) = \left[\begin{array}{cc}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\sigma^2 & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\sigma^2\end{array}\right],$$

whereas in model (6.38) it is fairly direct to show that:

$$Var\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2\end{array}\right]\right) = \left[\begin{array}{cc}\left(\mathbf{X}_1'\mathbf{M}_1\mathbf{X}_1\right)^{-1} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}' & \left(\mathbf{X}_2'\mathbf{M}_2\mathbf{X}_2\right)^{-1}\end{array}\right]\sigma^2,$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'$ and

$$\mathbf{Q}_{12} = \left(\mathbf{X}_1'\mathbf{M}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_1\mathbf{M}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{M}_2\mathbf{X}_2\right)^{-1}.$$

Consider now a Bayesian implementation. Model (6.38) requires assigning a prior distribution to the β-coefficients, whereas a prior must be assigned to the $\alpha's$ in the alternative model (6.40). These two priors must be probabilistically consistent, that is, uncertainty statements about the α's must translate into equivalent statements on the β's, and vice-versa.

Suppose that one works with (6.40), and that the prior distribution of the α-coefficients is the normal process

$$\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \right) \sigma^2,$$

where the hyperparameters are known. Further, σ^2 is assumed to follow a scaled inverted chi-square distribution with parameters ν and s^2. Using (6.34), the density of the posterior distribution of the α-coefficients is

$$p\left(\boldsymbol{\alpha}|\mathbf{y},\mathbf{H}\right) \propto \left[1 + \frac{\left(\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}}\right)' \left(\mathbf{W}'\mathbf{W}+\mathbf{V}^{-1}\right)\left(\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}}\right)}{(n+\upsilon)\left(\frac{S\widehat{\boldsymbol{\alpha}}+S_e+\nu s^2}{n+\upsilon}\right)} \right]^{-\frac{p_1+p_2+n+\nu}{2}}, \quad (6.45)$$

where

$$\left(\mathbf{W}'\mathbf{W}+\mathbf{V}^{-1}\right) = \begin{bmatrix} \mathbf{W}_1'\mathbf{W}_1+\mathbf{V}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2'\mathbf{W}_2+\mathbf{V}_2^{-1} \end{bmatrix},$$

$$\overline{\boldsymbol{\alpha}} = \begin{bmatrix} \overline{\boldsymbol{\alpha}}_1 \\ \overline{\boldsymbol{\alpha}}_2 \end{bmatrix} = \begin{bmatrix} \left(\mathbf{W}_1'\mathbf{W}_1+\mathbf{V}_1^{-1}\right)^{-1}\left(\mathbf{W}_1'\mathbf{y}+\mathbf{V}_1^{-1}\mathbf{a}_1\right) \\ \left(\mathbf{W}_2'\mathbf{W}_2+\mathbf{V}_2^{-1}\right)^{-1}\left(\mathbf{W}_2'\mathbf{y}+\mathbf{V}_2^{-1}\mathbf{a}_2\right) \end{bmatrix},$$

and

$$S\widehat{\boldsymbol{\alpha}} = \sum_{i=1}^{2} \left(\widehat{\boldsymbol{\alpha}}_i-\mathbf{a}_i\right)' \mathbf{W}_i'\mathbf{W}_i \left(\mathbf{W}_i'\mathbf{W}_i+\mathbf{V}_i^{-1}\right)^{-1} \mathbf{V}_i^{-1}\left(\widehat{\boldsymbol{\alpha}}_i-\mathbf{a}_i\right),$$

with

$$S_e = \mathbf{y}'\mathbf{y} - \sum_{i=1}^{2} \widehat{\boldsymbol{\alpha}}_i'\mathbf{W}_i'\mathbf{y}.$$

The two sets of regression coefficients are also uncorrelated, a posteriori, because the variance–covariance matrix of the multivariate-t distribution with density (6.45) is diagonal. However, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are not stochastically independent, even with independent priors and with an orthogonal parameterization. This is because the process of deconditioning with respect to σ^2, renders the α's mutually dependent. Recall that the density of a multivariate-t distribution with a diagonal covariance matrix cannot be written as the product of the intervening marginal densities.

Since, presumably, inferential interest centers on the original regression coefficients, the joint posterior density of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be obtained by

effecting the change of variables implicit in (6.42) and (6.43). In matrix notation

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{bmatrix} = \mathbf{T}\boldsymbol{\alpha},$$

with inverse

$$\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \mathbf{T}^{-1}\boldsymbol{\beta}.$$

The determinant of the Jacobian matrix is

$$|\mathbf{J}| = \det\left(\mathbf{T}^{-1}\right) = 1$$

since the determinant of a triangular matrix is equal to the product of its diagonal elements, in this case all being equal to 1. Then the joint p.d.f. of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, using the inverse transformation in conjunction with (6.45), is

$$p\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{H}\right) \propto \left[1 + \frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'\mathbf{T}'^{-1}\left(\mathbf{W}'\mathbf{W}+\mathbf{V}^{-1}\right)\mathbf{T}^{-1}\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)}{(n+\upsilon)\left(\frac{S\widehat{\boldsymbol{\alpha}}+S_e+\nu s^2}{n+\upsilon}\right)}\right]^{-\frac{p+n+\nu}{2}}$$

$$\propto \left\{1 + \frac{\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)'\left[\mathbf{T}\left(\mathbf{W}'\mathbf{W}+\mathbf{V}^{-1}\right)^{-1}\mathbf{T}'\right]^{-1}\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)}{(n+\upsilon)\left(\frac{S\widehat{\boldsymbol{\alpha}}+S_e+\nu s^2}{n+\upsilon}\right)}\right\}^{-\frac{p+n+\nu}{2}}, \quad (6.46)$$

where $p = p_1 + p_2$, and

$$\overline{\boldsymbol{\beta}} = \mathbf{T}\overline{\boldsymbol{\alpha}}.$$

Hence, the posterior distribution of the original regression coefficients is also multivariate$-t$, with mean vector $\overline{\boldsymbol{\beta}}$, variance–covariance matrix:

$$Var\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{H}\right) = \frac{S\widehat{\boldsymbol{\alpha}}+S_e+\nu s^2}{n+\upsilon-2}\mathbf{T}\left(\mathbf{W}'\mathbf{W}+\mathbf{V}^{-1}\right)^{-1}\mathbf{T}'$$

and $n+\upsilon$ degrees of freedom.

It cannot be overemphasized that if one works with the parameterization on the β's, and that if independent priors are assigned to the two sets of regressions, the resulting multivariate-t distribution would be different from that with density (6.46). This is so because, then, the induced prior for the α's implies that the two sets of regressions are correlated, with prior covariance matrix

$$\begin{aligned} Cov\left(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2'\right) &= Cov\left[\boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2, \boldsymbol{\beta}_2'\right] \\ &= Cov\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2'\right) + \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\, Var\left(\boldsymbol{\beta}_2\right). \end{aligned}$$

Thus, if the β's are taken as independently distributed a priori, this would not be the case for the α's. Hence, one would not obtain the same probability statements as those resulting from a model where the α's are independent, a priori.

6.3 The Mixed Linear Model

The mixed linear model, or (co)variance components model (e.g., Henderson, 1953; Searle, 1971), is one that includes fixed and random effects entering linearly into the conditional (given the random effects) expectation of the observations. Typically, as seen several times before, the random effects and the model residuals are assigned Gaussian distributions which depend on components of variance or covariance. These dispersion parameters may be regarded as known (e.g., as in the case of best linear unbiased prediction) or unknown. The distinction between fixed and random effects does not arise naturally in the Bayesian framework. However, many frequentist and likelihood-based results can be obtained as special (and limiting) cases of the Bayesian linear model. In this section, a linear model with two variance components will be discussed in detail from a Bayesian perspective. Models with more than two components of variance, or multivariate linear models where several vectors of location parameters enter into the conditional expectation structure, require a somewhat more involved analytical treatment. However, Bayesian implementations are possible through Monte Carlo methods, as discussed in subsequent chapters. Here, use will be made of results presented in Lindley and Smith (1972), Gianola and Fernando (1986), and Gianola et al. (1990), as well as developments given in the present and in the preceding chapters.

6.3.1 Bayesian View of the Mixed Effects Model

Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\boldsymbol{\beta}$, $(p \times 1)$, and $\mathbf{u}$, $(q \times 1)$, are location vectors related to observations $\mathbf{y}$, $(n \times 1)$, through the nonstochastic matrices $\mathbf{X}$ and $\mathbf{Z}$, respectively. Further, let

$$\mathbf{e}|\sigma_e^2 \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right),$$

be a vector of independently distributed residuals. Hence, given the residual variance, the sampling model is

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma_e^2\right). \tag{6.47}$$

Suppose that elicitation yields the prior distribution

$$\boldsymbol{\beta}|\sigma_\beta^2 \sim N\left(\mathbf{0}, \mathbf{B}\sigma_\beta^2\right),$$

where $\mathbf{B}$ is a known, nonsingular matrix and σ_β^2 is a hyperparameter. In the Bayesian setting, thus, the "fixed" (in the frequentist view) vector $\boldsymbol{\beta}$ is random, since it is assigned a distribution, this being a normal one in this case. Further, take

$$\mathbf{u}|\mathbf{A}\sigma_u^2 \sim N\left(\mathbf{0}, \mathbf{A}\sigma_u^2\right),$$

where, in a quantitative genetics context, $\mathbf{A}$ would be a matrix of additive relationships (also assumed to be nonsingular) whenever $\mathbf{u}$ is a vector of additive genetic effects and σ_u^2 is the additive genetic variance. In the classical mixed model, the preceding distribution, rather than viewed as a Bayesian prior, is interpreted as one resulting from a long-run process of sampling vectors from some conceptual population, with fixed $\mathbf{A}$ and σ_u^2. Such a sampling process generates a distribution with null mean and covariance matrix $\mathbf{A}\sigma_u^2$. It follows that the location structure, $\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}$, has only one random component ($\mathbf{u}$) in a frequentist setting, whereas it has two ($\boldsymbol{\beta}$ and $\mathbf{u}$) in the Bayesian probability model.

Assume that the prior distributions elicited for the variance components are independent scaled inverted chi-square processes with parameters $\left(s_e^2,\nu_e\right)$ and $\left(s_u^2,\nu_u\right)$ for σ_e^2 and σ_u^2, respectively. The consequences of these Bayesian assumptions follow:

- Unconditionally to the residual variance, the sampling model, given $\boldsymbol{\beta}$ and $\mathbf{u}$, is multivariate-t, with density

$$p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},s_e^2,\nu_e\right) \propto \left[1+\frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)}{\nu_e s_e^2}\right]^{-\frac{n+\nu_e}{2}}. \tag{6.48}$$

- The marginal distribution of the additive effects, unconditionally to the additive genetic variance, is a multivariate-t process having density

$$p\left(\mathbf{u}|s_u^2,\nu_u\right) \propto \left[1+\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{\nu_u s_u^2}\right]^{-\frac{q+\nu_u}{2}}. \tag{6.49}$$

- In the frequentist mixed effects linear model, the marginal distribution of the observations is indexed by the "fixed" parameter $\boldsymbol{\beta}$ and by the variance components σ_u^2 and σ_e^2, and this is

$$\mathbf{y}|\boldsymbol{\beta},\sigma_u^2,\sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta},\mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_u^2+\mathbf{I}\sigma_e^2\right). \tag{6.50}$$

  In the Bayesian model, on the other hand, the marginal distribution of the observations is obtained by integrating over all unknown parameters $\left(\boldsymbol{\theta}=\boldsymbol{\beta},\mathbf{u},\sigma_u^2,\sigma_e^2\right)$ entering in the model. If the joint prior of all parameters has the form

$$\begin{aligned} &p\left(\boldsymbol{\beta},\mathbf{u},\sigma_u^2,\sigma_e^2|\sigma_\beta^2,s_u^2,\nu_u,s_e^2,\nu_e\right)\\ &=p\left(\boldsymbol{\beta}|\sigma_\beta^2\right)p\left(\mathbf{u}|\sigma_u^2\right)p\left(\sigma_u^2|s_u^2,\nu_u\right)p\left(\sigma_e^2|s_e^2,\nu_e\right), \end{aligned}$$

then the marginal density of the observations can be written as

$$\begin{aligned} & p\left(\mathbf{y}|\sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) \\ &= \int\int\int\int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2\right) p\left(\boldsymbol{\theta}|\sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) d\boldsymbol{\theta} \\ &= \int\int\left[\int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2\right) p\left(\sigma_e^2|s_e^2, \nu_e\right) d\sigma_e^2\right] \\ &\times\left[\int p\left(\mathbf{u}|\sigma_u^2\right) p\left(\sigma_u^2|s_u^2, \nu_u\right) d\sigma_u^2\right] p\left(\boldsymbol{\beta}|\sigma_\beta^2\right) d\boldsymbol{\beta}\, d\mathbf{u} \end{aligned} \tag{6.51}$$

with the integrals in brackets evaluating to (6.48) and (6.49), respectively. Hence, the marginal density of the observations, or prior predictive distribution, cannot be written in closed form when the variance components are unknown.

- It is possible to go further than in (6.51) when the dispersion parameters are known. In the Bayesian model, one has to contemplate the variability (uncertainty) of the $\boldsymbol{\beta}$ values introduced by the prior distribution. In this case, the marginal density of the observations would be

$$p\left(\mathbf{y}|\sigma_\beta^2, \sigma_u^2, \sigma_e^2\right) = \int\int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2\right) p\left(\boldsymbol{\beta}|\sigma_\beta^2\right) p\left(\mathbf{u}|\sigma_u^2\right) d\boldsymbol{\beta}\, d\mathbf{u}. \tag{6.52}$$

Since the three densities in the integrand are normal, the exponents can be combined as

$$\begin{aligned} & \frac{1}{\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\right. \\ & \qquad \left.+\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\frac{\sigma_e^2}{\sigma_u^2}\right] \\ & = \frac{1}{\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}\right)'\left(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}\right)+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right], \end{aligned} \tag{6.53}$$

where $\boldsymbol{\alpha} = \left[\boldsymbol{\beta}', \mathbf{u}'\right]'$, $\mathbf{W} = \left[\begin{array}{cc} \mathbf{X} & \mathbf{Z} \end{array}\right]$, and

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{cc} \mathbf{B}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{array}\right].$$

Now let

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\mathbf{W}'\mathbf{y}. \tag{6.54}$$

Hence,

$$\begin{aligned} & (\mathbf{y}-\mathbf{W}\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha})+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \\ = \; & \mathbf{y}'\mathbf{y}-2\boldsymbol{\alpha}'\mathbf{W}'\mathbf{y}+\boldsymbol{\alpha}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\alpha} \\ = \; & \mathbf{y}'\mathbf{y}-2\boldsymbol{\alpha}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\widehat{\boldsymbol{\alpha}}+\boldsymbol{\alpha}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\alpha} \\ & +\widehat{\boldsymbol{\alpha}}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\widehat{\boldsymbol{\alpha}} \end{aligned}$$

where the quadratic in $\widehat{\boldsymbol{\alpha}}$ is added and subtracted in order to complete a "square". One can then write

$$\begin{aligned} & \mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)\widehat{\boldsymbol{\alpha}}+ \\ & (\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}) \\ = \; & \mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y}+(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}}). \end{aligned} \tag{6.55}$$

Using this in (6.53), it turns out that the integrand in (6.52) is expressible as

$$\begin{aligned} p\left(\mathbf{y}|\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right) & \propto \int \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y}\right.\right. \\ & \left.\left.+(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})\right]\right\}d\boldsymbol{\alpha} \\ \propto \exp\left(-\frac{\mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y}}{2\sigma_e^2}\right) & \int \exp\left[-\frac{(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})'\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha}-\widehat{\boldsymbol{\alpha}})}{2\sigma_e^2}\right]d\boldsymbol{\alpha}. \end{aligned} \tag{6.56}$$

The integral in (6.56) involves the kernel of a $p+q$-dimensional Gaussian distribution, and it evaluates to

$$(2\pi)^{\frac{p+q}{2}}\left|\left(\mathbf{W}'\mathbf{W}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\sigma_e^2\right|^{\frac{1}{2}}.$$

Since this does not involve the data, it gets absorbed in the integration constant, yielding, as marginal density of the data,

$$p\left(\mathbf{y}|\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right)=\frac{\exp\left(-\dfrac{\mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y}}{2\sigma_e^2}\right)}{\displaystyle\int\exp\left(-\frac{\mathbf{y}'\mathbf{y}-\widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y}}{2\sigma_e^2}\right)d\mathbf{y}}. \tag{6.57}$$

This can be shown to be the density of the normal process

$$\mathbf{y}|\sigma_\beta^2,\sigma_u^2,\sigma_e^2\sim N\left(\mathbf{0},\mathbf{XBX}'\sigma_\beta^2+\mathbf{ZAZ}'\sigma_u^2+\mathbf{I}\sigma_e^2\right).$$

A comparison of this with (6.50) reveals that the Bayesian linear model cannot be construed as its frequentist counterpart. In the latter, the marginal distribution of the observations requires knowledge

of the fixed effects and of the variances in order to compute probabilities. In the Bayesian model, the "fixed" effects have been integrated out (with respect to the prior), and the marginal process of the observations also depends on the known hyperparameter σ^2_β. The Bayesian model differs in other important respects, as illustrated in the following sections. For example, finding the marginal distribution of the observations (unconditionally to all parameters), or prior predictive distribution requires additional integration with respect to σ^2_u and σ^2_e.

6.3.2 Joint and Conditional Posterior Distributions

Under the assumptions made, the joint posterior density of all parameters is

$$p\left(\boldsymbol{\beta}, \mathbf{u}, \sigma^2_u, \sigma^2_e | \mathbf{y}, \sigma^2_\beta, s^2_u, \nu_u, s^2_e, \nu_e\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma^2_e\right) p\left(\boldsymbol{\beta}|\sigma^2_\beta\right) p\left(\mathbf{u}|\sigma^2_u\right) \times p\left(\sigma^2_u|s^2_u, \nu_u\right) p\left(\sigma^2_e|s^2_e, \nu_e\right). \tag{6.58}$$

All fully conditional posterior distributions can be identified directly from the joint posterior density, by viewing the latter as a function of the parameter(s) of interest, while keeping all other parameters and data $\mathbf{y}$ ("*ELSE*") fixed. All such densities can be written in closed form, as shown below.

- The conditional posterior density of the additive genetic variance, σ^2_u, given all other parameters, is

$$p\left(\sigma^2_u|\sigma^2_\beta, s^2_u, \nu_u, s^2_e, \nu_e, ELSE\right) \propto p\left(\mathbf{u}|\sigma^2_u\right) p\left(\sigma^2_u|s^2_u, \nu_u\right) \propto \left(\sigma^2_u\right)^{-\frac{q+\nu_u+2}{2}} \exp\left(-\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s^2_u}{2\sigma^2_u}\right). \tag{6.59}$$

  This is the kernel of a scaled inverted chi-square distribution with degree of belief parameter $\nu'_u = q + \nu_u$, and scale parameter

$$s'^2_u = \frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s^2_u}{q + \nu_u}.$$

  Similarly, the fully conditional posterior density of σ^2_e is

$$p\left(\sigma^2_e|\sigma^2_\beta, s^2_u, \nu_u, s^2_e, \nu_e, ELSE\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma^2_e\right) p\left(\sigma^2_e|s^2_e, \nu_e\right) \propto \left(\sigma^2_e\right)^{-\frac{n+\nu_e+2}{2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \nu_e s^2_e}{2\sigma^2_e}\right). \tag{6.60}$$

  Hence, the conditional posterior distribution of the residual variance is scaled inverted chi-square, with parameters $\nu'_e = n + \nu_e$, and

$$s'^2_e = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \nu_e s^2_e}{n + \nu_e}.$$

It is easy to verify that, given all other parameters, the two variance components are conditionally independent.

- The density of the conditional distribution of the "fixed" and "random" effects, given the variance components, is

$$p\left(\boldsymbol{\beta}, \mathbf{u}|\sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e, ELSE\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2\right) p\left(\boldsymbol{\beta}|\sigma_\beta^2\right) p\left(\mathbf{u}|\sigma_u^2\right).$$

Use can be made here of (6.53) and (6.55), to arrive at

$$\begin{aligned} & p\left(\boldsymbol{\beta}, \mathbf{u}|\sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e, ELSE\right) \\ \propto \exp & \left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\alpha}}'\mathbf{W}'\mathbf{y} + \left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\right)'\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\right)\right]\right\} \\ \propto \exp & \left[-\frac{\left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\right)'\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\right)}{2\sigma_e^2}\right], \end{aligned} \tag{6.61}$$

since the term that does not depend on $\boldsymbol{\alpha}$ is absorbed in the integration constant. It follows that the corresponding conditional posterior distribution is the multivariate normal process, as discussed in Chapter 1, Example 1.18,

$$\boldsymbol{\beta}, \mathbf{u}|\mathbf{y}, \sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e, \sigma_u^2, \sigma_e^2 \sim N\left(\widehat{\boldsymbol{\alpha}}, \left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\sigma_e^2\right). \tag{6.62}$$

Note that since σ_u^2 and σ_e^2 are assumed to be known, the distribution does not depend on parameters s_u^2, ν_u, s_e^2 and ν_e. However, these are left in the notation for the sake of symmetry in the presentation. We now write the mean of this distribution explicitly as

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} + \mathbf{B}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}.$$

If $\sigma_\beta^2 \to \infty$, which represents in some sense vague prior knowledge about $\boldsymbol{\beta}$, the above system of equations becomes what is known as Henderson's mixed model equations (Henderson et al., 1959; Henderson, 1973; Searle, 1971). In this situation, it can be shown that $\widehat{\boldsymbol{\beta}}$ above is the ML estimator of the fixed vector in a Gaussian mixed linear model with known variance components. Further, $\widehat{\mathbf{u}}$ is the best linear unbiased predictor of $\mathbf{u}$, or $BLUP$ (this holds under a mixed linear model even if the joint distribution of the random effects and of the observations is not Gaussian, but the dispersion parameters must be known). Hence, both the ML estimator and the $BLUP$ arise in the context of special cases of a more general Bayesian setting. If there is prior information about the "fixed" effects, σ_β^2 would be finite and, hence, the Bayesian model would effect some shrinkage toward

the prior mean (in this case assumed to be null) of the β vector. In fact, let $\widetilde{\boldsymbol{\alpha}}$ be the solution to the system

$$\mathbf{W}'\mathbf{W}\widetilde{\boldsymbol{\alpha}} = \mathbf{W}'\mathbf{y}.$$

The vector $\widetilde{\boldsymbol{\alpha}}$ can be interpreted as the ML estimator of $\boldsymbol{\alpha}$ in a model in which both $\boldsymbol{\beta}$ and $\mathbf{u}$ are treated as fixed (or as posterior mean in a model where both σ^2_β and σ^2_u go to infinity). Often, $\widetilde{\boldsymbol{\alpha}}$ is not defined uniquely, as $\mathbf{W}'\mathbf{W}$ may have deficient rank. On the other hand, $\widehat{\boldsymbol{\alpha}}$ is unique, provided that either $\mathbf{X}$ has full-column rank, or that σ^2_β is finite. Then, the posterior mean vector

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\mathbf{W}'\mathbf{y} \\ &= \left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{W}'\mathbf{W}\widetilde{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}^{-1}\mathbf{0}\right)\end{aligned}$$

can be viewed as a matrix weighted average of $\widetilde{\boldsymbol{\alpha}}$ (with this vector being a function of the data only) and of the mean of the prior distribution of $\boldsymbol{\alpha}$, which is $\mathbf{0}$. The matrix weights are $\mathbf{W}'\mathbf{W}$ (a measure of the precision of inferences contributed by the data) and $\boldsymbol{\Sigma}^{-1}$(a measure of prior precision), respectively.

- Result (6.62) implies that the marginal distributions of $\boldsymbol{\beta}$ and $\mathbf{u}$, given the variance components, are also multivariate normal, having mean vectors $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$, respectively. We will derive the form of these two densities systematically. First, note that the joint posterior density of $\boldsymbol{\beta}$ and $\mathbf{u}$, given the dispersion parameters, can be written as

$$\begin{aligned}p\left(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\sigma^2_\beta,\sigma^2_u,\sigma^2_e\right) \propto \exp\Bigg\{-\frac{1}{2\sigma^2_e}\Big[&(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \\ &+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma^2_e}{\sigma^2_\beta}+\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\frac{\sigma^2_e}{\sigma^2_u}\Big]\Bigg\}. \qquad (6.63)\end{aligned}$$

Now put $\mathbf{w}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and write, employing a decomposition similar to that in (6.55),

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\frac{\sigma^2_e}{\sigma^2_u} \\ &=[\mathbf{w}(\boldsymbol{\beta})-\mathbf{Z}\mathbf{u}]'[\mathbf{w}(\boldsymbol{\beta})-\mathbf{Z}\mathbf{u}]+\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\frac{\sigma^2_e}{\sigma^2_u} \\ &=\mathbf{w}'(\boldsymbol{\beta})\mathbf{w}(\boldsymbol{\beta})-\widehat{\widehat{\mathbf{u}}}(\boldsymbol{\beta})'\mathbf{Z}'\mathbf{w}(\boldsymbol{\beta}) \\ &+\left[\widehat{\widehat{\mathbf{u}}}(\boldsymbol{\beta})-\mathbf{u}\right]'\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right)\left[\widehat{\widehat{\mathbf{u}}}(\boldsymbol{\beta})-\mathbf{u}\right],\end{aligned}$$

where

$$\widehat{\widehat{\mathbf{u}}}(\boldsymbol{\beta}) = \left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right)^{-1}\mathbf{Z}'\mathbf{w}(\boldsymbol{\beta}).$$

Making use of this in the joint density (6.63) and integrating with respect to $\mathbf{u}$, to obtain the marginal posterior density of $\boldsymbol{\beta}$, given the variance components, yields

$$\begin{aligned}
&p\left(\boldsymbol{\beta}|\mathbf{y},\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right)\\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{w}'\left(\boldsymbol{\beta}\right)\mathbf{w}\left(\boldsymbol{\beta}\right)-\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)'\mathbf{Z}'\mathbf{w}\left(\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\right]\right\}\\
&\int \exp\left\{-\frac{1}{2\sigma_e^2}\left[\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)-\mathbf{u}\right]'\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2}\right)\left[\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)-\mathbf{u}\right]\right\}d\mathbf{u}\\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{w}'\left(\boldsymbol{\beta}\right)\mathbf{w}\left(\boldsymbol{\beta}\right)-\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)'\mathbf{Z}'\mathbf{w}\left(\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\right]\right\}\\
&\qquad\times(2\pi)^{\frac{q}{2}}\left|\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\tfrac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\sigma_e^2\right|^{\frac{1}{2}}\\
&\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{w}'\left(\boldsymbol{\beta}\right)\mathbf{w}\left(\boldsymbol{\beta}\right)-\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)'\mathbf{Z}'\mathbf{w}\left(\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\right]\right\}.
\end{aligned}\tag{6.64}$$

Now

$$\begin{aligned}
&\mathbf{w}'\left(\boldsymbol{\beta}\right)\mathbf{w}\left(\boldsymbol{\beta}\right)-\widehat{\widehat{\mathbf{u}}}\left(\boldsymbol{\beta}\right)'\mathbf{Z}'\mathbf{w}\left(\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\\
&=\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\\
&-\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\\
&=\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left[\mathbf{I}-\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\tfrac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\mathbf{Z}'\right]\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\\
&\qquad+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}.
\end{aligned}\tag{6.65}$$

Using matrix identity (6.27), it can be shown that

$$\mathbf{I}-\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\tfrac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\mathbf{Z}'=\left(\mathbf{Z}\mathbf{A}\mathbf{Z}'\tfrac{\sigma_u^2}{\sigma_e^2}+\mathbf{I}\right)^{-1}=\mathbf{V}^{-1}.$$

Making use of this in (6.65), the marginal density (6.64) can be written as

$$\begin{aligned}
&p\left(\boldsymbol{\beta}|\mathbf{y},\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right)\\
&\propto\left\{-\tfrac{1}{2\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\frac{\sigma_e^2}{\sigma_\beta^2}\right]\right\}.
\end{aligned}$$

Note that, as $\sigma_\beta^2\to\infty$, the kernel of the posterior density tends toward the likelihood of $\boldsymbol{\beta}$ in a Gaussian linear mixed model with

known variance components and covariance matrix as given in (6.50). The ML estimator of $\boldsymbol{\beta}$ is

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

and using the standard decomposition, one obtains

$$\begin{aligned} p\left(\boldsymbol{\beta}|\mathbf{y},\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right) &\propto \exp\left\{-\tfrac{1}{2\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}\right)\right.\right. \\ &\left.\left.+\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right]\right\} \\ &\propto \exp\left\{-\tfrac{1}{2\sigma_e^2}\left[\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right]\right\}. \quad (6.66) \end{aligned}$$

The quadratic forms in $\boldsymbol{\beta}$ can now be combined as

$$\begin{aligned} &\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)+\boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta}\tfrac{\sigma_e^2}{\sigma_\beta^2} \\ &=\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\mathbf{B}^{-1}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right) \\ &+\widetilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\mathbf{B}^{-1}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1}\mathbf{B}^{-1}\tfrac{\sigma_e^2}{\sigma_\beta^2}\widetilde{\boldsymbol{\beta}}, \qquad (6.67) \end{aligned}$$

where

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\mathbf{B}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

After some matrix manipulations, it is possible to show that this is precisely the $\boldsymbol{\beta}$-component of $\widehat{\boldsymbol{\alpha}}$ in (6.62). Using (6.67) in (6.66) and retaining the part that varies with $\boldsymbol{\beta}$ yields, as density of the marginal distribution of $\boldsymbol{\beta}$ (conditionally on the variance components),

$$\begin{aligned} &p\left(\boldsymbol{\beta}|\mathbf{y},\sigma_\beta^2,\sigma_u^2,\sigma_e^2\right) \\ &\propto \exp\left\{-\tfrac{1}{2\sigma_e^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\mathbf{B}^{-1}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\}. \end{aligned}$$

Thus, the marginal posterior density of $\boldsymbol{\beta}$ when the dispersion parameters are known, is the normal process

$$\boldsymbol{\beta}|\mathbf{y},\sigma_\beta^2,\sigma_u^2,\sigma_e^2 \sim N\left(\widehat{\boldsymbol{\beta}},\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\mathbf{B}^{-1}\tfrac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1}\sigma_e^2\right). \qquad (6.68)$$

Again, note that the mean of the distribution is a matrix weighted average of the ML estimator of $\boldsymbol{\beta}$ and of the mean of the prior distribution of this vector, assumed to be null in this case.

- Using similar algebra, it can be found that the marginal posterior density of $\mathbf{u}$, given the variance components, is

$$\mathbf{u}|\mathbf{y},\sigma^2_\beta,\sigma^2_u,\sigma^2_e \sim N\left(\widehat{\mathbf{u}},\left(\mathbf{Z}'\mathbf{T}^{-1}\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right)^{-1}\sigma^2_e\right), \tag{6.69}$$

with

$$\widehat{\mathbf{u}}=\left(\mathbf{Z}'\mathbf{T}^{-1}\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right)^{-1}\mathbf{Z}'\mathbf{T}^{-1}\mathbf{y},$$

and

$$\mathbf{T}=\mathbf{X}\mathbf{B}\mathbf{X}'\frac{\sigma^2_\beta}{\sigma^2_e}+\mathbf{I}.$$

The mean vector is equal to the $\mathbf{u}$-component of $\widehat{\boldsymbol{\alpha}}$ in (6.62).

- Since the joint distribution of $\boldsymbol{\beta}$ and $\mathbf{u}$ is jointly Gaussian, with parameters as in (6.62), it follows that the processes

$$\left[\boldsymbol{\beta}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e,\sigma^2_u,\sigma^2_e,\mathbf{u}\right],$$
$$\left[\mathbf{u}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e,\sigma^2_u,\sigma^2_e,\boldsymbol{\beta}\right],$$

are normal as well. Making use of results in Chapter 1, one can arrive at

$$\begin{aligned}&\boldsymbol{\beta}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e,\sigma^2_u,\sigma^2_e,\mathbf{u}\\&\sim N\left(\widehat{\boldsymbol{\beta}}(\mathbf{u}),\left[\mathbf{X}'\mathbf{X}+\mathbf{B}^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\sigma^2_e\right),\end{aligned} \tag{6.70}$$

where

$$\widehat{\boldsymbol{\beta}}(\mathbf{u})=\left[\mathbf{X}'\mathbf{X}+\mathbf{B}^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Z}\mathbf{u}).$$

Also

$$\begin{aligned}&\mathbf{u}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e,\sigma^2_u,\sigma^2_e,\boldsymbol{\beta}\\&\sim N\left(\widehat{\mathbf{u}}(\boldsymbol{\beta}),\left[\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right]^{-1}\sigma^2_e\right),\end{aligned} \tag{6.71}$$

where

$$\widehat{\mathbf{u}}(\boldsymbol{\beta})=\left[\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\frac{\sigma^2_e}{\sigma^2_u}\right]^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

6.3.3 Marginal Distribution of Variance Components

The location parameters can be integrated out analytically of the joint posterior density. First, write the density of the joint posterior distribution

of all parameters as

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y}, \sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) \\
&\propto p\left(\sigma_u^2 | s_u^2, \nu_u\right) p\left(\sigma_e^2 | s_e^2, \nu_e\right) \left(\sigma_e^2\right)^{-\frac{n}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \\
&\times \exp\left\{-\tfrac{1}{2\sigma_e^2}\left[\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\alpha}}_c'\mathbf{W}'\mathbf{y} + (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_c)'\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_c)\right]\right\} \\
&\propto p\left(\sigma_u^2 | s_u^2, \nu_u\right) p\left(\sigma_e^2 | s_e^2, \nu_e\right) \left(\sigma_e^2\right)^{-\frac{n}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \exp\left(-\frac{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\alpha}}_c'\mathbf{W}'\mathbf{y}}{2\sigma_e^2}\right) \\
&\times \exp\left[-\frac{(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_c)'\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}_c)}{2\sigma_e^2}\right],
\end{aligned}
$$

where $\widehat{\boldsymbol{\alpha}}_c$ is $\widehat{\boldsymbol{\alpha}}$ of (6.62), with the subscript placed to emphasize the dependence of this vector on the unknown variance components σ_u^2 and σ_e^2. The location parameters now appear in the kernel of a multivariate normal density and can be integrated out by analytical means. After integration, one gets

$$
\begin{aligned}
&p\left(\sigma_u^2, \sigma_e^2 | \mathbf{y}, \sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) \\
&\propto p\left(\sigma_u^2 | s_u^2, \nu_u\right) p\left(\sigma_e^2 | s_e^2, \nu_e\right) \left(\sigma_e^2\right)^{-\frac{n}{2}} \left(\sigma_u^2\right)^{-\frac{q}{2}} \exp\left(-\frac{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\alpha}}_c'\mathbf{W}'\mathbf{y}}{2\sigma_e^2}\right) \\
&\times \left|\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \sigma_e^2\right|^{\frac{1}{2}}.
\end{aligned}
$$

The joint density of the two variance components can be written explicitly as

$$
\begin{aligned}
&p\left(\sigma_u^2, \sigma_e^2 | \mathbf{y}, \sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) \propto \left(\sigma_u^2\right)^{-\frac{q+\nu_u+2}{2}} \left(\sigma_e^2\right)^{-\frac{n-p-q+\nu_e+2}{2}} \\
&\times \exp\left(-\frac{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\alpha}}_c'\mathbf{W}'\mathbf{y} + \nu_e s_e^2 + \nu_u s_u^2 \frac{\sigma_e^2}{\sigma_u^2}}{2\sigma_e^2}\right) \left|\left(\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}^{-1}\right)\right|^{-\frac{1}{2}}. \quad (6.72)
\end{aligned}
$$

It is not possible to go further in the level of marginalization. This is because the joint distribution involves ratios between variance components, both in the exponential expression and in the matrix $\boldsymbol{\Sigma}$, as part of the determinant. Note that the two variance components are not independent a posteriori (even if they are so, a priori).

6.3.4 Marginal Distribution of Location Parameters

The integral of the joint posterior density (6.58), with respect to the unknown variance components, so as to obtain the unconditional posterior

density of the location parameters, can be represented as

$$p\left(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e\right)\propto\left[\int p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\sigma^2_e\right)p\left(\sigma^2_e|s^2_e,\nu_e\right)d\sigma^2_e\right]$$
$$\times\left[\int p\left(\mathbf{u}|\sigma^2_u\right)p\left(\sigma^2_u|s^2_u,\nu_u\right)d\sigma^2_u\right]p\left(\boldsymbol{\beta}|\sigma^2_\beta\right).$$

Each of the two integrals is proportional to the kernel of a multivariate-t density of appropriate order, so one can write:

$$p\left(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\sigma^2_\beta,s^2_u,\nu_u,s^2_e,\nu_e\right)\propto\left[1+\frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)}{\nu_e s^2_e}\right]^{-\frac{n+\nu_e}{2}}$$
$$\times\left(1+\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{\nu_u s^2_u}\right)^{-\frac{q+\nu_u}{2}}p\left(\boldsymbol{\beta}|\sigma^2_\beta\right). \tag{6.73}$$

This density does not have a recognizable form. The first two expressions are multivariate-t, and their product would define a poly-t density, if it were not for the presence of $p\left(\boldsymbol{\beta}|\sigma^2_\beta\right)$.

Suppose now that σ^2_β goes to infinity, so that, in the limit, one is in a situation of vague prior knowledge about this location parameter. The prior density of $\boldsymbol{\beta}$ becomes flatter and flatter and, in the limit, it is proportional to a constant. However, this uniform process is not proper, as the integral over $\boldsymbol{\beta}$ is not finite. In this case, the joint distribution of $\boldsymbol{\beta}$ and $\mathbf{u}$ is in a poly-t form, and marginalization can be carried out one step further. Before integration of the variance components, the joint posterior density takes the form

$$p\left(\boldsymbol{\beta},\mathbf{u},\sigma^2_e,\sigma^2_u|\mathbf{y},s^2_u,\nu_u,s^2_e,\nu_e\right)$$
$$\propto p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\sigma^2_e\right)p\left(\sigma^2_e|s^2_e,\nu_e\right)p\left(\mathbf{u}|\sigma^2_u\right)p\left(\sigma^2_u|s^2_u,\nu_u\right),$$

since now the prior density of $\boldsymbol{\beta}$ is flat. One can write

$$\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)=\left[\mathbf{w}\left(\mathbf{u}\right)-\mathbf{X}\boldsymbol{\beta}\right]'\left[\mathbf{w}\left(\mathbf{u}\right)-\mathbf{X}\boldsymbol{\beta}\right]$$
$$=\left[\mathbf{w}\left(\mathbf{u}\right)-\mathbf{X}\widetilde{\boldsymbol{\beta}}\right]'\left[\mathbf{w}\left(\mathbf{u}\right)-\mathbf{X}\widetilde{\boldsymbol{\beta}}\right]+\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right),$$

where

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{w}\left(\mathbf{u}\right)\\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{u}\right)\end{aligned}$$

is the "regression" of $\mathbf{w}(\mathbf{u}) = \mathbf{y} - \mathbf{Z}\mathbf{u}$ on $\mathbf{X}$. Using this in the joint posterior density yields

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2, \sigma_u^2 | \mathbf{y}, s_u^2, \nu_u, s_e^2, \nu_e\right) \\
&\propto \left(\sigma_e^2\right)^{-\frac{n+\nu_e+2}{2}} \exp\left\{-\frac{\left[\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)\right] + \nu_e s_e^2}{2\sigma_e^2}\right\} \\
&\times \exp\left[-\frac{\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)}{2\sigma_e^2}\right]\left(\sigma_u^2\right)^{-\frac{q+\nu_u+2}{2}} \exp\left(-\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s_u^2}{2\sigma_u^2}\right).
\end{aligned}
$$

This expression can be integrated analytically with respect to $\boldsymbol{\beta}$, to obtain

$$
\begin{aligned}
&p\left(\mathbf{u}, \sigma_e^2, \sigma_u^2 | \mathbf{y}, s_u^2, \nu_u, s_e^2, \nu_e\right) \\
&\propto \left(\sigma_e^2\right)^{-\frac{n+\nu_e+2}{2}} \exp\left[-\frac{\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right) + \nu_e s_e^2}{2\sigma_e^2}\right] \\
&\times \left|\left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma_e^2\right|^{\frac{1}{2}}\left(\sigma_u^2\right)^{-\frac{q+\nu_u+2}{2}} \exp\left(-\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s_u^2}{2\sigma_u^2}\right) \\
&\propto \left(\sigma_e^2\right)^{-\frac{n-p+\nu_e+2}{2}} \exp\left[-\frac{\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right) + \nu_e s_e^2}{2\sigma_e^2}\right] \\
&\times \left(\sigma_u^2\right)^{-\frac{q+\nu_u+2}{2}} \exp\left(-\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s_u^2}{2\sigma_u^2}\right). \qquad (6.74)
\end{aligned}
$$

Note now that the two variance components appear in kernels of scaled inverted chi-square densities and that, given $\mathbf{u}$, their distributions are independent. Further, the dispersion parameters can be integrated out analytically, obtaining

$$
\begin{aligned}
&p\left(\mathbf{u} | \mathbf{y}, \sigma_\beta^2, s_u^2, \nu_u, s_e^2, \nu_e\right) \\
&\propto \left[\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right) + \nu_e s_e^2\right]^{-\frac{n-p+\nu_e}{2}} \\
&\times \left[\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u s_u^2\right]^{-\frac{q+\nu_u}{2}} \\
&\propto \left[1 + \frac{\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)}{\nu_e s_e^2}\right]^{-\frac{n-p+\nu_e}{2}} \\
&\times \left(1 + \frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{\nu_u s_u^2}\right)^{-\frac{q+\nu_u}{2}},
\end{aligned}
$$

as the hyperparameters are constant, and can be factored out of the expression. Finally, note that

$$\begin{aligned}\mathbf{w}(\mathbf{u}) - \mathbf{X}\widetilde{\boldsymbol{\beta}} &= \mathbf{y} - \mathbf{Z}\mathbf{u} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{Z}\mathbf{u}\right) \\ &= \mathbf{M}\left(\mathbf{y} - \mathbf{Z}\mathbf{u}\right),\end{aligned}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Using this in the marginal posterior density of $\mathbf{u}$ above gives

$$\begin{aligned}p\left(\mathbf{u}|\mathbf{y}, \sigma^2_\beta, s^2_u, \nu_u, s^2_e, \nu_e\right) \propto &\left[1 + \frac{(\mathbf{y} - \mathbf{Z}\mathbf{u})'\,\mathbf{M}\,(\mathbf{y} - \mathbf{Z}\mathbf{u})}{\nu_e s^2_e}\right]^{-\frac{n-p+\nu_e}{2}} \\ &\times \left(1 + \frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{\nu_u s^2_u}\right)^{-\frac{q+\nu_u}{2}}. \qquad (6.75)\end{aligned}$$

Hence, the marginal distribution of $\mathbf{u}$ is not in any easily recognizable form and further marginalization, e.g., with respect to a subvector of this location parameter, is not feasible by analytical means. However, it is possible to draw samples from the marginal distribution of each element of $\mathbf{u}$ by means of MCMC methods to be discussed in later chapters.

In summary, most of the analytical results available for a Bayesian analysis of linear models under Gaussian assumptions have been presented here. When the model contains more than one unknown variance component, it is not possible to arrive at the fully marginal distributions of individual parameters. The analytical treatment is even more involved when the model is nonlinear in the location parameters, or when the response variables are not Gaussian. In the next two chapters, additional topics in Bayesian analysis will be presented, such as the role of the prior distribution and Bayesian tools for model comparison.

7

The Prior Distribution and Bayesian Analysis

7.1 Introduction

In the preceding two chapters, some of the basic machinery for the development of Bayesian probability models, with emphasis on linear specifications, was presented. It was seen that the inferences drawn depend on the forms of the likelihood function and of the prior distribution. A natural question is: What is the impact of the prior on inferences? Clearly, a similar query could be raised about the effect of the likelihood function. Alternatively, one can pose the question: How much information is contributed by the data (or by the prior) about the quantities of interest? In this chapter, we begin with an example that illustrates the effect of the prior distribution on inferences. Subsequently, some measures of statistical information are presented and used to quantify what is encoded in the likelihood function, and in the prior and posterior processes. Another section describes how prior distributions, contributing "little" information relative to that contributed by data, can be constructed. This section includes a discussion of Jeffreys' prior, of the maximum entropy principle, and of what is called reference analysis. A step-by-step derivation of the associated reference prior, including the situation in which nuisance parameters are present in the model, is given at the end.

n	ML estimate	Posterior mode	Posterior mean
5	0.2	0.2	0.286
10	0.2	0.2	0.250
20	0.2	0.2	0.227
40	0.2	0.2	0.214

TABLE 7.1. Effect of sample size (n) on the mean and mode of the posterior distribution of the gene frequency: uniform prior.

7.2 An Illustration of the Effect of Priors on Inferences

Suppose that one wishes to infer the frequency of a certain allele (θ) in some homogeneous population. Further, assume that all that can be stated a priori is that the frequency is contained in the interval $[0,1]$. A random sample of n genes is drawn from the population and x copies of the allele of interest are observed in the sample. The ML estimator of θ is x/n, and its sample variance is $\theta(1-\theta)/n$ (this can be estimated empirically by replacing the unknown parameter by its ML estimate).

A reasonable Bayesian probability model consists of a uniform prior distribution in the said interval, plus a binomial sampling model, with x successes out of n independent trials. The prior distribution is centered at $1/2$, and the posterior density of the gene frequency is

$$p(\theta|x,n) \propto \theta^x (1-\theta)^{n-x}.$$

Hence, the posterior process is the $Be(x+1, n-x+1)$ distribution, with its mode being equal to the ML estimator, and the posterior mean being $(x+1)/(n+2)$. Table 7.1 gives a sequence of posterior distributions at increasing sample sizes; in each of these distributions the ratio of the number of successes to sample size is kept constant at 0.2, which is the ML estimate of the gene frequency. Note that the posterior mean gets closer to the ML estimate as n increases; in the limit, the posterior mean tends toward x/n, suggesting that the information from the sample overwhelms the prior, asymptotically.

Assume now that the prior distribution is the $Be(11,11)$ process. The posterior density is now

$$p(\theta|x,n) \propto \theta^{x+11-1} (1-\theta)^{n+11-x-1},$$

so the corresponding distribution is $Be(x+11, n+11-x)$, having mean $(x+11)/(n+22)$, and mode $(x+10)/(n+20)$. Note that as $n \to \infty$, both the mean and mode go towards x/n. Table 7.2 gives a sequence of posterior distributions similar to that displayed in Table 7.1. Again, both the posterior mean and mode move toward the ML estimator as sample size increases, but the influence of the beta prior assigned here is more marked

n	ML estimate	Posterior mode	Posterior mean
5	0.2	0.440	0.444
10	0.2	0.400	0.406
20	0.2	0.350	0.357
50	0.2	0.286	0.292

TABLE 7.2. Effect of sample size (n) on the mean and mode of the posterior distribution of the gene frequency: beta prior.

than that of the uniform distribution in the preceding case. The reason for this is that the $Be(11, 11)$ distribution is fairly sharp and assigns small prior probability to values of θ smaller than $\frac{1}{4}$. If $n = 1000$ and $x = 200$, thus keeping the ML estimator at $\frac{1}{5}$, the posterior mean and mode would be both approximately equal to 0.207, verifying the "asymptotic domination" of the prior by the likelihood function.

The dissipation of the influence of the prior as sample size increases was already seen in a discrete setting, when Bayes theorem was introduced. The result can be coined in a more general form as follows. Suppose that n independent draws are made from the same distribution $[y_i|\theta]$, and let the prior density be $p(\theta|H)$, where H could be a set of hyperparameters. The posterior density of θ is then

$$p(\theta|y_1, y_2, \ldots, y_n, H) \propto p(\theta|H) \prod_{i=1}^{n} p(y_i|\theta)$$
$$\propto \exp\left\{\sum_{i=1}^{n}\left(\log\left[p(y_i|\theta)\right] + \frac{\log\left[p(\theta|H)\right]}{n}\right)\right\}.$$

Then, as n increases,

$$\log\left[p(y_i|\theta)\right] + \frac{\log\left[p(\theta|H)\right]}{n} \to \log\left[p(y_i|\theta)\right],$$

and

$$p(\theta|y_1, y_2, ..., y_n, H) \to \frac{\exp\left\{\sum_{i=1}^{n} \log\left[p(y_i|\theta)\right]\right\}}{\int \exp\left\{\sum_{i=1}^{n} \log\left[p(y_i|\theta)\right]\right\} d\theta},$$

which is the normalized likelihood, assuming the integral exists. Hence, the contribution of the prior to the posterior becomes less and less important as the sample size grows. This can be expressed by saying that, given enough data, the prior is expected to have a small influence on inferences about θ. This is examined in more detail in the following section.

7.3 A Rapid Tour of Bayesian Asymptotics

Intuitively, the beliefs about a parameter θ, reflected in its posterior distribution, should become more concentrated about the true parameter value θ_0 as the amount of information increases (loosely speaking, as the number of observations $n \to \infty$). In this section, a summary of some relevant asymptotic results is presented, following Bernardo and Smith (1994) closely. The arguments and results parallel those for asymptotic theory in ML estimation, as given in Chapters 3 and 4. Hence, only the essentials are presented.

7.3.1 Discrete Parameter

Suppose θ takes one of several mutually exclusive and exhaustive states, so its prior distribution is discrete. The posterior distribution is

$$p(\theta_i|\mathbf{y}) = \frac{p(\mathbf{y}|\theta_i)\, p(\theta_i)}{\sum_i p(\mathbf{y}|\theta_i)\, p(\theta_i)},$$

where the sum is taken over all possible states of the parameter. Dividing both numerator and denominator by the likelihood conferred by the data to the true parameter value θ_0, one gets

$$p(\theta_i|\mathbf{y}) = \frac{\frac{p(\mathbf{y}|\theta_i)}{p(\mathbf{y}|\theta_0)} p(\theta_i)}{\sum_i \frac{p(\mathbf{y}|\theta_i)}{p(\mathbf{y}|\theta_0)} p(\theta_i)}.$$

Assuming the observations are independent, given the parameter, the posterior is expressible as

$$\begin{aligned} p(\theta_i|\mathbf{y}) &= \frac{\exp\left[\log \frac{p(\mathbf{y}|\theta_i)}{p(\mathbf{y}|\theta_0)} + \log p(\theta_i)\right]}{\sum_i \exp\left[\log \frac{p(\mathbf{y}|\theta_i)}{p(\mathbf{y}|\theta_0)} + \log p(\theta_i)\right]} \\ &= \frac{\exp\left[\sum_{j=1}^{n} \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} + \log p(\theta_i)\right]}{\sum_i \exp\left[\sum_{j=1}^{n} \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} + \log p(\theta_i)\right]}. \end{aligned}$$

Now, as $n \to \infty$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} &= \int \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} p(y_j|\theta_0)\, dy \\ &= -\int \log \frac{p(y_j|\theta_0)}{p(y_j|\theta_i)} p(y_j|\theta_0)\, dy. \end{aligned}$$

The integral immediately above is called the Kullback–Leibler distance or the discrepancy between two distributions, which is shown in Section 7.4.7 to be 0 when $\theta_i = \theta_0$ and positive otherwise. Hence, as $n \to \infty$,

$$\frac{1}{n}\sum_{j=1}^{n} \log \frac{p\left(y_j|\theta_i\right)}{p\left(y_j|\theta_0\right)} \to -\int \log \frac{p\left(y_j|\theta_0\right)}{p\left(y_j|\theta_i\right)} p\left(x_j|\theta_0\right) dy$$
$$= \begin{cases} 0, & \textit{for } \theta_i = \theta_0, \\ -\infty, & \textit{otherwise.} \end{cases}$$

Hence

$$\lim_{n\to\infty} p\left(\theta_i|\mathbf{y}\right) = \lim_{n\to\infty} \left\{ \frac{\exp\left[\sum_{j=1}^{n} \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} + \log p\left(\theta_i\right)\right]}{\sum_i \exp\left[\sum_{j=1}^{n} \log \frac{p(y_j|\theta_i)}{p(y_j|\theta_0)} + \log p\left(\theta_i\right)\right]} \right\}$$
$$= 0, \quad \textit{for all } \theta_i \neq \theta_0,$$

and is equal to 1 for $\theta_i = \theta_0$. This indicates that, as sample size grows, the posterior distribution becomes more and more concentrated around θ_0 and that, in the limit, all probability mass is placed on the true value. It can be shown that if one fails to assign positive prior probability to the true state of the parameter, the limiting posterior concentrates around the parameter value producing the smallest Kullback–Leibler discrepancy with the true model (Bernardo and Smith, 1994).

7.3.2 Continuous Parameter

The posterior density of a parameter vector can be represented as

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \propto \exp\left[\log p\left(\mathbf{y}|\boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right)\right].$$

Let $\widetilde{\boldsymbol{\theta}}$ be the prior mode and let $\widehat{\boldsymbol{\theta}}_n$ be the ML estimator based on a sample of size n. Taylor series expansions of the log-prior density and of the likelihood function yield (recall that the gradient is null when evaluated at the corresponding modal value):

$$\log p\left(\boldsymbol{\theta}\right) \approx \log p\left(\widetilde{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right)' \widetilde{\mathbf{H}} \left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right)$$

and

$$\log p\left(\mathbf{y}|\boldsymbol{\theta}\right) \approx \log p\left(\widehat{\boldsymbol{\theta}}_n\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)' \widehat{\mathbf{H}} \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right),$$

where

$$\widetilde{\mathbf{H}} = -\left.\frac{\partial^2 \log p\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}$$

and

$$\widehat{\mathbf{H}} = -\frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

is the observed information matrix. These two matrices of second derivatives, called Hessians, can be interpreted as the "precision" matrices of the prior distribution and of the sampling model, respectively. Under the usual regularity conditions, the remainder of the approximation to the log-likelihood is small in large samples, since the likelihood is expected to be sharp and concentrated near the true value of the parameter vector. Further, recall that in large samples the ML estimate $\widehat{\boldsymbol{\theta}}_n$ is expected to be close to its true value $\boldsymbol{\theta}_0$. Using the Taylor series approximations to the prior and the likelihood, the posterior density is, roughly,

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \exp\left\{\log\left[p\left(\widehat{\boldsymbol{\theta}}_n\right) p\left(\widetilde{\boldsymbol{\theta}}\right)\right]\right.$$
$$\left.-\frac{1}{2}\left[\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)' \widehat{\mathbf{H}}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right) + \left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)' \widetilde{\mathbf{H}}\left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)\right]\right\}.$$

After retaining only the terms that involve $\boldsymbol{\theta}$ one gets

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)' \widehat{\mathbf{H}}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right) + \left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)' \widetilde{\mathbf{H}}\left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)\right]\right\}. \tag{7.1}$$

The two quadratic forms on the parameter vector can be combined, via the formulas used repeatedly in the preceding chapter, to obtain

$$\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)' \widehat{\mathbf{H}}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right) + \left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)' \widetilde{\mathbf{H}}\left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)$$
$$= \left(\boldsymbol{\theta}-\widehat{\widetilde{\boldsymbol{\theta}}}_n\right)' \left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)\left(\boldsymbol{\theta}-\widehat{\widetilde{\boldsymbol{\theta}}}_n\right)$$
$$+ \left(\widehat{\boldsymbol{\theta}}_n-\widetilde{\boldsymbol{\theta}}\right)' \widehat{\mathbf{H}}\left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)^{-1}\widetilde{\mathbf{H}}\left(\widehat{\boldsymbol{\theta}}_n-\widetilde{\boldsymbol{\theta}}\right), \tag{7.2}$$

with

$$\widehat{\widetilde{\boldsymbol{\theta}}}_n = \left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)^{-1}\left(\widehat{\mathbf{H}}\widehat{\boldsymbol{\theta}}_n + \widetilde{\mathbf{H}}\widetilde{\boldsymbol{\theta}}\right). \tag{7.3}$$

Employing (7.2) in (7.1) and keeping only the terms involving $\boldsymbol{\theta}$ gives

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\widehat{\widetilde{\boldsymbol{\theta}}}_n\right)' \left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)\left(\boldsymbol{\theta}-\widehat{\widetilde{\boldsymbol{\theta}}}_n\right)\right\}. \tag{7.4}$$

Hence, under regularity conditions, the posterior distribution is asymptotically normal with mean $\widehat{\widetilde{\boldsymbol{\theta}}}_n$ and covariance matrix $\left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)^{-1}$. We write

$$\boldsymbol{\theta}|\mathbf{y} \sim N\left[\widehat{\widetilde{\boldsymbol{\theta}}}_n, \left(\widehat{\mathbf{H}}+\widetilde{\mathbf{H}}\right)^{-1}\right]. \tag{7.5}$$

Note that the approximation to the posterior mean is a matrix weighted average of the prior mode and of the ML estimator, with the weights being the corresponding precision matrices. The matrix sum $\widehat{\mathbf{H}} + \widetilde{\mathbf{H}}$ is a measure of posterior precision and its inverse reflects the posterior variances and covariances among elements of $\boldsymbol{\theta}$.

Another approximation can be obtained as follows. For large n, the precision matrix $\widehat{\mathbf{H}}$ will tend to be much larger than the prior precision matrix $\widetilde{\mathbf{H}}$. Then, roughly, $\widehat{\mathbf{H}} + \widetilde{\mathbf{H}} \approx \widehat{\mathbf{H}}$ and $\widehat{\widetilde{\boldsymbol{\theta}}}_n \approx \widehat{\boldsymbol{\theta}}_n$. Since the prior precision matrix is "dominated" by the Hessian corresponding to the sampling model, the ML estimator receives a much larger weight, so the mean of the approximate posterior distribution will be very close to the ML estimator. Hence, for large n,

$$\boldsymbol{\theta}|\mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}_n, \widehat{\mathbf{H}}^{-1}\right). \tag{7.6}$$

For conditionally i.i.d. observations,

$$-\frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = -\sum_{i=1}^{n} \frac{\partial^2 \log p\left(y_i|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = n\left[\frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2 \log p\left(y_i|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right].$$

By the weak law of large numbers, the term in square brackets, which is equal to $\frac{1}{n}\widehat{\mathbf{H}}$, converges in probability (symbolized " $\xrightarrow{p}$ ") to its expectation:

$$\frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2 \log p\left(y_i|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \xrightarrow{p} E\left[-\frac{\partial^2 \log p\left(y|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = \mathbf{I}_1\left(\boldsymbol{\theta}\right),$$

which is Fisher's information measure for a sample of size 1. Hence,

$$\widehat{\mathbf{H}} \xrightarrow{p} n\mathbf{I}_1\left(\boldsymbol{\theta}\right) = \mathbf{I}\left(\boldsymbol{\theta}\right).$$

That is, the observed information matrix converges in probability to Fisher's information matrix. Then, an alternative approximation to the posterior distribution is

$$\boldsymbol{\theta}|\mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}_n, \left[n\mathbf{I}_1\left(\boldsymbol{\theta}\right)\right]^{-1}\right).$$

It is important to stress that the heuristics presented above build on the existence of regularity conditions. It is definitely not always the case that the posterior distribution approaches normality as sample size increases. We refer the reader to Bernardo and Smith (1994), and references therein, for a discussion of this delicate subject.

Example 7.1 *Asymptotic distribution of location parameters in a normal model*

Suppose that n observations are taken from some strain of fish and that the sampling model for some attribute is

$$x_i|\theta_1, \sigma^2 \sim N\left(\theta_1, \sigma^2\right).$$

The residual dispersion is known. The location parameter has an uncertainty distribution according to the process $\theta_1 \sim N(0, \lambda_1)$, where the variance is also known. Another random sample of n observations is drawn from some other strain according to the model

$$y_i|\theta_2, \tau^2 \sim N\left(\theta_2, \tau^2\right),$$

with $\theta_2 \sim N(0, \lambda_2)$; again, λ_2 is known. The joint posterior density of θ_1 and θ_2 is

$$\begin{aligned} &p\left(\theta_1, \theta_2|\sigma,^2 \lambda_1, \tau^2, \lambda_2, \mathbf{x}, \mathbf{y}\right) \\ \propto &\left[\prod_{i=1}^{n} p\left(x_i|\theta_1, \sigma^2\right) p\left(\theta_1|\lambda_1\right)\right]\left[\prod_{i=1}^{n} p\left(y_i|\theta_2, \tau^2\right) p\left(\theta_2|\lambda_2\right)\right]. \end{aligned}$$

It can be seen that θ_1 and θ_2 are independently distributed, a posteriori, for any sample size. Since the prior and the posterior are normal, it follows that the posterior distribution of each θ is normal as well. For example, the posterior distribution of θ_1 has mean

$$\begin{aligned} E\left(\theta_1|\lambda_1, \sigma^2, \mathbf{x}\right) &= \left(n + \frac{\sigma^2}{\lambda_1}\right)^{-1} n\overline{x} \\ &= \left(1 + \frac{\sigma^2}{n\lambda_1}\right)^{-1} \overline{x} \end{aligned}$$

and variance

$$Var\left(\theta_1|\lambda_1, \sigma^2, \mathbf{x}\right) = \left(n + \frac{\sigma^2}{\lambda_1}\right)^{-1} \sigma^2.$$

As $n \to \infty$, the mean and variance tend to $\overline{x}$ and σ^2/n, respectively. The asymptotic joint posterior distribution of θ_1 and θ_2 can then be written as

$$\theta_1, \theta_2|\sigma^2, \tau^2, \mathbf{x}, \mathbf{y} \sim N\left(\left[\begin{array}{c} \overline{x} \\ \overline{y} \end{array}\right], \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & \tau^2 \end{array}\right] n^{-1}\right).$$

The same result is arrived at by employing (7.6) since $\overline{x}$ and $\overline{y}$ are the ML estimators of θ_1 and θ_2, respectively. In this example the observed and expected information matrices are equal because the variances are assumed to be known. ■

7.4 Statistical Information and Entropy

7.4.1 Information

A readable introduction to the concept of information theory in probability is in Applebaum (1996), and some of the ideas are adapted to our context

hereinafter. Consider the following two statements:
(1) In two randomly mated, randomly selected lines of finite size derived from the same population, mean values and gene frequencies differ after 50 generations.
(2) A new mutant is found at generation 2 with frequency equal to 15%.
Arguably, the second statement conveys more information than the first, because it involves an event having a very low prior probability. The first statement should not evoke surprise, as the corresponding event is expected in the light of well-established theory. Let E be an event and let $p(E)$ be its probability. Then using the above line of argument, the information contained in the observed event, $I(E)$, should be a decreasing function of its probability.

There are three conditions that an information measure must meet:
(1) It must be positive.
(2) The information from observing two events jointly must be at least as large as that from the observation of any of the single elementary events. For example, suppose there are two unlinked loci in a population in Hardy–Weinberg equilibrium and that the probability of observing an AA genotype is $p(AA)$, while that of observing Bb is $p(Bb)$. As the two events are independent, one has that

$$p(AA \cap Bb) = p(AA)\, p(Bb).$$

Since $p(AA \cap Bb) \leq p(AA)$ and $p(AA \cap Bb) \leq p(Bb)$, it follows that the information content $I(AA \cap Bb) \geq I(AA)$ and, similarly, $I(AA \cap Bb) \geq I(Bb)$.
(3) For independent events E_1 and E_2 (Applebaum, 1996):

$$I(E_1 \cap E_2) = I(E_1) + I(E_2).$$

To illustrate, suppose a machine reads: "genotype at first locus is AA", so the information is $I(AA)$, while another machine yields "genotype at second locus is Bb", with the information being $I(Bb)$. Hence, the information is $I(AA) + I(Bb)$. On the other hand, if the machine reads both genotypes simultaneously, we would not have information over and above $I(AA) + I(Bb)$.

From information theory, the function satisfying the three conditions given above must have the form

$$I(E) = -K \log_a [p(E)], \tag{7.7}$$

where K and a are positive constants. Since $0 \leq p(E) \leq 1$, it follows that this information measure is positive, as K is positive, thus meeting the first condition. If the event is certain, $I(E) = 0$, and no information is gained from knowing that the event took place (this would be known beforehand). On the other hand, if the event is impossible, $p(E) = 0$, the

information measure is not finite; this is viewed as reasonable, indicating the impossibility of obtaining information from events that do not occur. Second, for independent events

$$\begin{aligned} I(E_1 \cap E_2) &= -K \log_a [p(E_1 \cap E_2)] \\ &= -K \log_a [p(E_1)] - K \log_a [p(E_2)] \\ &= I(E_1) + I(E_2), \end{aligned}$$

satisfying the second and third conditions. Standard choices for the constants are $K = 1$ and $a = 2$, and the units in which information is measured are called "bits". For example, suppose one crosses genotypes Aa and aa; the offspring can be either Aa or aa, with equal probability. Hence, the information resulting from observing one of the two alternatives is:

$$I(Aa) = I(aa) = -\log_2\left(2^{-1}\right) = 1 \text{ bit.}$$

Example 7.2 *Cross between double heterozygotes*
Suppose that the cross $AaBb \times AaBb$ is made, with the two loci unlinked, as before. We calculate the information accruing from observation of each of the following events:

1. progeny is aa;
2. progeny is Bb;
3. an offspring $aabb$ is observed;
4. the offspring is $Aabb$; and
5. the outcome of the cross is $AaBb$.

Using Mendel's rules:

1. $p(aa) = p(bb) = \frac{1}{4}$, so $I(aa) = I(bb) = -\log_2\left(2^{-2}\right) = 2$ bits.
2. $p(Bb) = \frac{1}{2}$, so $I(Bb) = -\log_2\left(2^{-1}\right) = 1$ bit.
3. $p(aabb) = \frac{1}{16}$, so $I(aabb) = -\log_2\left(2^{-4}\right) = 4$ bits. Note that $I(aabb) = I(aa) + I(bb)$ since the two events are independent.
4. $p(Aabb) = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$ and $I(Aabb) = -\log_2\left(2^{-3}\right) = 3$ bits.
5. $p(AaBb) = \frac{1}{4}$ and $I(AaBb) = -\log_2\left(2^{-2}\right) = 2$ bits. ■

In the continuous case, heuristically, if one replaces "event" by "observed data y", the information measure becomes

$$\begin{aligned} I(y) &= -\log_2 [p(y|\theta)] \\ &= -\left[\frac{\log p(y|\theta)}{\log(2)}\right] \\ &= 0.69315 - \log p(y|\theta), \end{aligned}$$

where $p(y|\theta)$ is the density function indexed by parameter θ. Since 0.69315 is a constant, it can be dropped, as it is convenient to work with natural logarithms (that is, any calculation of information should be increased by 0.69315 to have the "correct" number of bits). Further, for n data points drawn independently from the same distribution, the information would be

$$I(\mathbf{y}) = 0.69315 - \sum_{i=1}^{n} \log p(y_i|\theta). \tag{7.8}$$

For example, if $y_i \sim NIID\left(\mu, \sigma^2\right)$, the information in a sample of size n would be

$$I(\mathbf{y}) = 0.69315 + \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2,$$

so information is related to the "deviance" $\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2$.

7.4.2 Entropy of a Discrete Distribution

Since the information content of an event depends on its probability (or density), it is technically more sensible to think in terms of information from the distribution of a random variable. It is important to keep this in mind, because in Bayesian analysis sometimes one wishes to find and use a prior distribution conveying "as little information as possible" (Box and Tiao, 1973; Zellner, 1971; Bernardo, 1979; Berger and Bernardo, 1992). The formal development requires finding a distribution that minimizes some information measure.

Consider one discrete random variable X having K mutually exclusive and exhaustive states and probability distribution p_i $(i = 1, 2, \ldots, K)$ (in this section, $\Pr(X = x_i) = p(x_i)$, the usual notation for a p.m.f. is replaced by p_i). Since X is random, one does not know beforehand how much information will be contained in a yet-to-occur observation. Hence, $I(X)$ is random as well, but a property of the distribution is the mean information

$$\begin{aligned} H(p_1, p_2, ..., p_K) &= E[I(X)] = E\{-\log[\Pr(X = x_i)]\} \\ &= -\sum_{i=1}^{K} p_i \log(p_i). \end{aligned} \tag{7.9}$$

This is called the entropy of a distribution, with its name due to the fact that this functional form appears in thermodynamics. In physics, entropy is used to refer to the degree of randomness or disorder in processes. Shannon (1948) coined the term information entropy to describe the tendency of communication to become more and more distorted by noise. For example, if some material is photocopied over again and in this process becomes illegible, the information is continually degraded. Thus, a distribution conveying

minimum information, given some constraints that we wish this distribution to reflect (e.g., the probabilities must add up to 1), can be viewed as a maximum entropy distribution. The role of entropy in Bayesian analysis in connection with the elicitation of priors, is discussed later in this chapter.

The entropy function is not defined when $p_i = 0$, and the term $p_i \log(p_i)$ is taken to be null in such a case (Applebaum, 1996). Since entropy involves an average of numbers that must be at least 0, it follows that $H(\cdot) \geq 0$, with the null value corresponding to the situation where there is no uncertainty whatsoever (so no information is gained). It follows, then, that a situation of "maximum entropy" is one where "a lot" of information is to be gained from observation. Entropy is a measure of the prior uncertainty of a random experiment associated with the probability distribution in question or, alternatively, of the information gained when the outcome is observed (Baldi and Brunak, 1998). If one is told about the outcome of an event, the entropy is reduced from $H(\cdot)$ to 0, so this measures the gain in information. In some sense the concept is counterintuitive (Durbin et al., 1998) because the more randomness, the higher the entropy and the information. The concept becomes clearer when one thinks in terms of a reduction of entropy after some information is received. Hence, what matters is the difference in entropy before and after observing data; this difference can be interpreted as the informational content provided by a set of observations. For example, one may wish to compute the difference in entropy between the prior and posterior distributions.

Example 7.3 *Entropy of a random DNA*

This is from Durbin et al. (1998): if each base (A, C, G, T) occurs equiprobably within a DNA sequence, the probability of a random base is $\frac{1}{4}$. The entropy per base is then

$$-\sum_{i=1}^{4} \frac{1}{4} \log_2 \left(\frac{1}{4}\right) = 2 \text{ bits.}$$

Durbin et al. (1998) state that this can be interpreted as the number of binary questions (with yes or no responses) needed to discover the outcome. For example, the first question would be: Is the basis a purine or a pyrimidine?. If the answer is "purine", the choice must be between A or G. The second question is: Which is the specific base? The more uncertainty there is, the more questions are needed to discover the outcome. ■

Example 7.4 *Entropy of a conserved position*

Suppose that a DNA sequence is expected to be random so that, before observation, the entropy is 2 bits, as in the preceding example of Durbin et al. (1998). It is observed that in a given position the frequency of A is 0.7, whereas that of G is 0.3. Using (7.9), the entropy after observation is

then

$$-\left[\frac{7}{10}\log_2\frac{7}{10}+\frac{3}{10}\log_2\frac{3}{10}\right]=0.88 \text{ bits.}$$

The information content of the position is given by the difference in entropy before and after observation: $2-0.88=1.12$ bits. The more conserved the position, the higher its information content is. ■

Example 7.5 *Sampling of genes*

Let allele A have frequency p in some population, so that all other possible alleles appear with probability $1-p$. Suppose that n alleles are drawn at random and that x are of the A form. The process is binomial and the entropy of the distribution of the random variable X is

$$H(p)=-\sum_{x=0}^{n}\left[\log\frac{n!\,p^x(1-p)^{n-x}}{x!\,(n-x)!}\right]\frac{n!\,p^x(1-p)^{n-x}}{x!\,(n-x)!}.$$

If $n=1$, this reduces to the entropy of a Bernoulli distribution

$$H(p)=-p\log(p)-(1-p)\log(1-p).$$

Taking derivatives with respect to p, to find the maximum value of the entropy for this distribution,

$$\frac{dH(p)}{dp}=-\log(p)+\log(1-p).$$

Setting to 0 and solving gives $p=\frac{1}{2}$ as the gene frequency giving maximum entropy, with the maximized entropy being equal to 1 bit when expressed in a $\log_2$ base. Hence the gene frequency distribution producing maximum entropy is that corresponding to the situation where allele A has the same frequency as that of all the other alleles combined (but without making a distinction between these).

Now consider the situation where there are three alleles with frequencies p_1, p_2, and $p_3=1-p_1-p_2$. The entropy of the gene frequency distribution is now

$$H(p_1,p_2)=-p_1\log(p_1)-p_2\log(p_2)-(1-p_1-p_2)\log(1-p_1-p_2).$$

To find the maximum entropy distribution we calculate the gradients

$$\frac{dH(p_1,p_2)}{dp_i}=-\log(p_i)+\log(1-p_1-p_2),\quad i=1,2.$$

Setting these derivatives to 0, one arrives at $p_1=p_2=\frac{1}{3}$. Again, the maximum entropy distribution is one where the three alleles are equally likely.

In general, for K states, the entropy in (7.9) has the following gradient, after introducing the constraint that $\sum_{i=1}^{K} p_i = 1$,

$$\frac{dH\left(p_1, p_2, ..., p_{K-1}\right)}{dp_i} = -\log\left(p_i\right) + \log\left(1 - p_1 - p_2 - \cdots - p_{K-1}\right).$$

After setting to 0, one obtains the solution $p_i = 1/K$ for all alleles. Hence, the maximum entropy distribution is uniform and the maximized entropy is equal to

$$H\left(\frac{1}{K}, \frac{1}{K}, ..., \frac{1}{K}\right) = -\sum_{i=1}^{K} \frac{1}{K} \log\left(\frac{1}{K}\right) = \log\left(K\right).$$

■

The preceding example illustrates that entropy is a measure of uncertainty. The entropy is null when there is complete certainty about the allele to be sampled, and it is maximum when one cannot make an informed choice about which of the allelic states is more likely. Further, for a number of states $K < M$, $\log\left(K\right) < \log\left(M\right)$, which implies that the entropy (uncertainty) grows monotonically with the number of choices that can be made.

7.4.3 Entropy of a Joint and Conditional Distribution

Suppose that the pair of random variables (X, Y) has a joint distribution with joint probabilities p_{xy}, $(x = 1, 2, \ldots, m,\ y = 1, 2, \ldots, n)$. The entropy of the joint distribution (Applebaum, 1996) is

$$\begin{aligned} H\left(p_{11}, p_{12}, \ldots, p_{mn}\right) &= -\sum_{x=1}^{m}\sum_{y=1}^{n} p_{xy} \log\left(p_{xy}\right) \\ &= -\sum_{x=1}^{m}\sum_{y=1}^{n} p_x p_{y|x} \log\left(p_x p_{y|x}\right). \end{aligned} \tag{7.10}$$

Further

$$\begin{aligned} H\left(p_{11}, p_{12}, \ldots, p_{mn}\right) &= -\sum_{x=1}^{m}\sum_{y=1}^{n} p_x p_{y|x} \left[\log\left(p_x\right) + \log\left(p_{y|x}\right)\right] \\ &= -\sum_{x=1}^{m} p_x \log\left(p_x\right) \sum_{y=1}^{n} p_{y|x} - \sum_{x=1}^{m} \left[\sum_{y=1}^{n} p_{y|x} \log\left(p_{y|x}\right)\right] p_x \\ &= H\left(\mathbf{p}_x\right) + \sum_{x=1}^{m} H\left(\mathbf{p}_{y|x}\right) p_x = H\left(\mathbf{p}_x\right) + \overline{H}\left(\mathbf{p}_{y|x}\right), \end{aligned} \tag{7.11}$$

where $\mathbf{p}_x$ is the vector of probabilities of the marginal distribution of X and $\mathbf{p}_{y|x}$ is the vector of probabilities of the conditional distribution of Y given X. This indicates that the entropy of the joint distribution is the sum of two components: the entropy of the distribution of X and the average $\overline{H}\left(\mathbf{p}_{y|x}\right)$ of the conditional entropies $H\left(\mathbf{p}_{y|x}\right)$, taken with respect to the distribution of X. The second term provides a measure of the uncertainty about Y knowing that X has been realized, but without being able to state what its value is; on the other hand, $H\left(\mathbf{p}_{y|x}\right)$ measures the uncertainty when one knows the value taken by X. Note that if X and Y are independent

$$\begin{aligned} H\left(p_{11}, p_{12}, \ldots, p_{mn}\right) &= -\sum_{x=1}^{m}\sum_{y=1}^{n} p_x p_y \left[\log\left(p_x\right) + \log\left(p_y\right)\right] \\ &= -\sum_{x=1}^{m} p_x \log\left(p_x\right) \sum_{y=1}^{n} p_y - \sum_{y=1}^{n} p_y \log\left(p_y\right) \sum_{x=1}^{m} p_x \\ &= H\left(\mathbf{p}_x\right) + H\left(\mathbf{p}_y\right), \end{aligned} \tag{7.12}$$

where $\mathbf{p}_y$ is the vector of probabilities of the distribution of Y.

7.4.4 Entropy of a Continuous Distribution

The entropy of a continuous distribution (Shannon, 1948; Jaynes, 1957) is defined to be

$$H\left[p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] = -\int \cdots \int \log\left[p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] p\left(\mathbf{y}|\boldsymbol{\theta}\right) d\mathbf{y}, \tag{7.13}$$

where $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$ is the density function of the random vector $\mathbf{y}$, indexed by a parameter $\boldsymbol{\theta}$. Note that entropy is not invariant under transformation. Suppose the random variable is scalar, and that one considers the one-to-one change of variables $z = f\left(y\right)$, where z increases monotonically with y. Then

$$p\left(z|\boldsymbol{\theta}\right) = p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right] \frac{df^{-1}\left(z\right)}{dz}.$$

The entropy of the distribution of z becomes

$$\begin{aligned} H\left[p\left(z|\boldsymbol{\theta}\right)\right] &= -\int \log\left\{p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right] \frac{df^{-1}\left(z\right)}{dz}\right\} p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right] \frac{df^{-1}\left(z\right)}{dz} dz \\ &= -\int \log\left\{p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right]\right\} p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right] \frac{df^{-1}\left(z\right)}{dz} dz \\ &\quad -\int \log\left[\frac{df^{-1}\left(z\right)}{dz}\right] p\left[f^{-1}\left(z\right)|\boldsymbol{\theta}\right] \frac{df^{-1}\left(z\right)}{dz} dz \\ &= -\int \left\{\log\left[p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right]\right\} p\left(\mathbf{y}|\boldsymbol{\theta}\right) dy - \int \log\left[\frac{df^{-1}\left(z\right)}{dz}\right] p\left(z|\boldsymbol{\theta}\right) dz \end{aligned}$$

$$= H\left[p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] - E_z\left\{\log\left[\frac{df^{-1}\left(z\right)}{dz}\right]\right\}.$$

This indicates that the information content is not invariant even under a one-to-one, monotonic, transformation, putting in question the usefulness of entropy as a measure of uncertainty in the continuous case (Bernardo and Smith, 1994). In Section 7.4.7 it is shown that a transformation invariant measure is provided by the relative entropy.

Example 7.6 *Entropy of a uniform distribution*
Suppose Y is a scalar variable distributed uniformly in the interval (a, b), so its density is $1/(b-a)$. Using (7.13) yields

$$\begin{aligned} H\left[p\left(y|a,b\right)\right] &= -\int_a^b \left[\log\left(\frac{1}{b-a}\right)\right]\frac{1}{b-a}dy \\ &= \log\left(b-a\right). \end{aligned}$$

When dealing with discrete random variables, it was pointed out earlier that entropy is at least null. This does not always carry to the continuous case (Applebaum, 1996). For example, note that if the difference between the bounds $b-a$ is < 1, then the entropy would be negative. ■

Example 7.7 *Entropy of a normal distribution*
Let now $Y \sim N\left(\mu, \sigma^2\right)$, so

$$\begin{aligned} H\left[p\left(y|\mu,\sigma^2\right)\right] &= -\int\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right]\right)\right\} \\ &\quad\times\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right]dy \\ &= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\int\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right]dy \\ &\quad+\frac{1}{2\sigma^2}\int\left(y-\mu\right)^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right]dy \\ &= \frac{1}{2}\left[1+\log\left(2\pi\sigma^2\right)\right]. \end{aligned}$$

Note that entropy increases with σ^2. Consider now the standardized multivariate normal distribution $\mathbf{y} \sim N_n\left(\mathbf{0}, \mathbf{R}\right)$, where $\mathbf{R}$ is a correlation matrix

having all off-diagonal elements equal to ρ. Then

$$\begin{aligned} H\left[p\left(\mathbf{y}|\mathbf{0},\mathbf{R}\right)\right] &= -\int\left\{\log\left[\frac{1}{(2\pi)^{\frac{n}{2}}\left|\mathbf{R}\right|^{\frac{1}{2}}}\exp-\left(\frac{\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}}{2}\right)\right]\right\} \\ &\quad\times\frac{1}{(2\pi)^{\frac{n}{2}}\left|\mathbf{R}\right|^{\frac{1}{2}}}\exp-\left(\frac{\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}}{2}\right)d\mathbf{y} \\ &= \frac{1}{2}\left[\log\left(2\pi\right)^{n}+\log\left|\mathbf{R}\right|\right]+E\left(\frac{\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}}{2}\right) \\ &= \frac{1}{2}\left[\log\left(2\pi\right)^{n}+\log\left|\mathbf{R}\right|+tr\mathbf{R}^{-1}Var\left(\mathbf{y}\right)\right] \\ &= \frac{1}{2}\left[\log\left(2\pi\right)^{n}+\log\left|\mathbf{R}\right|+n\right]. \end{aligned}$$

Now, using results in Searle et al. (1992),

$$\left|\mathbf{R}\right| = \left(1-\rho\right)\left[1+\left(n-1\right)\rho\right],$$

and employing this in the preceding yields

$$H\left[p\left(\mathbf{y}|\mathbf{0},\mathbf{R}\right)\right] = \frac{1}{2}\left\{\log\left(2\pi\right)^{n}+\log\left(1-\rho\right)+\log\left[1+\left(n-1\right)\rho\right]+n\right\}.$$

When $\rho = 0$, the entropy is equal to

$$H\left[p\left(\mathbf{y}|\mathbf{0},\mathbf{R}\right)\right] = \frac{n}{2}\left[1+\log\left(2\pi\right)\right],$$

which is n times larger than the entropy of a univariate standard normal distribution. ■

Example 7.8 *Entropy in a truncated normal model*

Consider calculating the entropies of the prior, likelihood, and posterior distributions in a model where the mean of a normal distribution, μ, is to be inferred; the variance is known. Suppose that n observations are collected and that these are i.i.d. as $N\left(\mu,\sigma^2\right)$. The p.d.f. of an observation is

$$p\left(y_j|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_j-\mu\right)^2\right],$$

where $j = 1, 2, \ldots, n$ identifies the data point. Assume that the prior distribution of μ is uniform between boundaries a and b. As in Example 7.6, the entropy of the prior distribution with density p_0 is

$$H\left(p_0\right) = -\int\log\left(p_o\right)p_o\,d\mu = \log\left(b-a\right).$$

Thus, the prior entropy increases with the distance between boundaries a and b. This means that the prior uncertainty about the values of μ increases as the boundaries are further apart.

The entropy of the likelihood for a sample of size n is

$$H\left[p\left(\mathbf{y}|\mu,\sigma^2\right)\right] = -\int\cdots\int \log\left[p\left(\mathbf{y}|\mu,\sigma^2\right)\right] p\left(\mathbf{y}|\mu,\sigma^2\right) d\mathbf{y}$$
$$= -\int\cdots\int \log\left\{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right]\right\} p\left(\mathbf{y}|\mu,\sigma^2\right) d\mathbf{y}$$
$$= \frac{n}{2}\left[1+\log\left(2\pi\sigma^2\right)\right].$$

Hence, the entropy of the likelihood increases as sample size increases. This is because the joint density becomes smaller and smaller, so one gets more "surprised", i.e., more information accrues, as more data are observed.

The density of the posterior distribution of μ is

$$p\left(\mu|\sigma^2,a,b,\mathbf{y}\right) \propto \exp\left[-\frac{n\left(\mu-\overline{y}\right)^2}{2\sigma^2}\right] I\left(a<\mu<b\right),$$

where $\overline{y}$ is the mean of all observations and $I\left(\cdot\right)$ is an indicator variable denoting the region where the parameter is allowed to take density. Hence, the posterior distribution is a truncated normal process between a and b. In the absence of truncation

$$\mu|a,b,\mathbf{y} \sim N\left(\overline{y},\sigma^2/n\right).$$

Now the normalized posterior density is

$$p\left(\mu|\sigma^2,a,b,\mathbf{y}\right) = \frac{\exp\left[-\frac{n(\mu-\overline{y})^2}{2\sigma^2}\right]}{\int_a^b \exp\left[-\frac{n(\mu-\overline{y})^2}{2\sigma^2}\right] d\mu}$$
$$= \frac{\exp\left[-\frac{n(\mu-\overline{y})^2}{2\sigma^2}\right]}{\sqrt{2\pi\sigma^2/n}\left[\Phi\left(\frac{b-\overline{y}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right]}, \tag{7.14}$$

with the integration constant being

$$c_n = \left\{\sqrt{2\pi\sigma^2/n}\left[\Phi\left(\frac{b-\overline{y}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right]\right\}^{-1}. \tag{7.15}$$

If $b=\infty$, so that the posterior takes nonnull density only between a and ∞, standard results from truncation selection in quantitative genetics (e.g., Falconer and Mackay, 1996) give

$$E\left(\mu|\sigma^2,a,b,\mathbf{y}\right) = \overline{y} + \frac{\phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}{1-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}\frac{\sigma}{\sqrt{n}} = \eta,$$

and

$$Var\left(\mu|\sigma^2,a,b,\mathbf{y}\right)$$
$$=\frac{\sigma^2}{n}\left\{1-\frac{\phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}{1-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}\left[\frac{\phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}{1-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)}-\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right]\right\}=\gamma,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively.

The entropy of the truncated normal posterior distribution is then

$$H\left[p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)\right]=-\int_a^b\log\left[p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)\right]p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)d\mu$$
$$=\frac{n}{2\sigma^2}\int_a^b\left(\mu-\overline{y}\right)^2p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)d\mu$$
$$+\log\left\{\sqrt{2\pi\sigma^2/n}\left[\Phi\left(\frac{b-\overline{y}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right]\right\}.$$

The expectation to be computed can be written as

$$E_{\mu|\sigma^2,a,b,\mathbf{y}}\left(\mu-\overline{y}\right)^2=\int_a^b\left(\mu-\overline{y}\right)^2p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)d\mu$$
$$=\int_a^b\left[(\mu-\eta)+(\eta-\overline{y})\right]^2p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)d\mu$$
$$=\int_a^b\left(\mu-\eta\right)^2p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)d\mu+\left(\eta-\overline{y}\right)^2=\gamma+\left(\eta-\overline{y}\right)^2,\qquad(7.16)$$

with the last term resulting because the expected value of $\mu-\eta$ is 0 under the posterior distribution. Recall that η and γ are the mean and variance of the posterior distribution, respectively. Now, using this in the last expression for entropy given above, it turns out that

$$H\left[p\left(\mu|\sigma^2,a,b,\mathbf{y}\right)\right]=\frac{n\left[\gamma+\left(\eta-\overline{y}\right)^2\right]}{2\sigma^2}$$
$$+\log\sqrt{2\pi\sigma^2/n}+\log\left[\Phi\left(\frac{b-\overline{y}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right].$$

The entropy of the posterior density goes to 0 as sample size goes to infinity. Algebraically, the first and second terms go to ∞ and $-\infty$ as $n\to\infty$, so

they cancel out. As $n \to \infty$, the third term goes to 0 because the posterior distribution becomes a point mass in the limit, with all density assigned to the "true" value of μ. In the limit, there is no longer "surprise", as the true value of μ is known with probability equal to 1. ■

7.4.5 Information about a Parameter

The concept of amount of information about a parameter provided by a sample of observations was discussed earlier in the book in connection with likelihood inference. However, the treatment presented was heuristic, without making use of information-theoretic arguments. Important contributions to the use of information theory in statistics are Fisher (1925), Shannon (1948), and Kullback (1968). Here an elementary introduction to the subject is provided, following closely some of the developments in Kullback (1968).

Kullback's Information Measure

Suppose that an observation y is made and that one wishes to evaluate two competing models or hypotheses H_i $(i = 1, 2)$, each having prior probability $p(H_i)$. For example, one of the hypotheses could be that the observation belongs to some distribution. The posterior probability of hypothesis 1 being true is

$$p(H_1|y) = \frac{p(y|H_1)\,p(H_1)}{p(y|H_1)\,p(H_1) + p(y|H_2)\,p(H_2)},$$

where $p(y|H_i)$ is the density of the observation under hypothesis H_i and $p(H_i)$ is the prior probability of H_i being true. The logarithm of the ratio of posterior probabilities, or posterior log-odds ratio, is then

$$\log \frac{p(H_1|y)}{p(H_2|y)} = \log \frac{p(H_1)}{p(H_2)} + \log \frac{p(y|H_1)}{p(y|H_2)},$$

where the ratio of densities is often known as the Bayes factor in favor of hypothesis 1 relative to hypothesis 2. Rearranging the preceding expression

$$\log \frac{p(y|H_1)}{p(y|H_2)} = \log \frac{p(H_1|y)}{p(H_2|y)} - \log \frac{p(H_1)}{p(H_2)}. \tag{7.17}$$

The difference between posterior and prior log-odds ratios was interpreted by Kullback (1968) as the information contained in y for discrimination in favor of H_1 against H_2. The difference can be either negative or positive.

The expected information for discrimination per observation from H_1 is

$$\begin{aligned} I(1:2) &= \int \log \frac{p(y|H_1)}{p(y|H_2)} p(y|H_1)\, dy \\ &= \int \left[\log \frac{p(H_1|y)}{p(H_2|y)} - \log \frac{p(H_1)}{p(H_2)}\right] p(y|H_1)\, dy \\ &= \int \left[\log \frac{p(H_1|y)}{p(H_2|y)}\right] p(y|H_1)\, dy - \log \frac{p(H_1)}{p(H_2)}. \end{aligned} \tag{7.18}$$

This is the difference between the mean value (taken with respect to the distribution $[y|H_1]$) of the logarithm of the posterior odds ratio and of the prior log-odds ratio. Similarly, the expected information per observation for discrimination in favor of H_2 is

$$I(2:1) = \int \log \frac{p(y|H_2)}{p(y|H_1)} p(y|H_2)\, dy. \tag{7.19}$$

Kullback (1968) defines the "divergence" between hypotheses (or distributions) as

$$\begin{aligned} J(1:2) &= I(1:2) + I(2:1) \\ &= \int \left[\log \frac{p(H_1|y)}{p(H_2|y)}\right] p(y|H_1)\, dy - \log \frac{p(H_1)}{p(H_2)} + \\ &\quad \int \left[\log \frac{p(H_2|y)}{p(H_1|y)}\right] p(y|H_2)\, dy - \log \frac{p(H_2)}{p(H_1)} \\ &= \int \left[\log \frac{p(H_1|y)}{p(H_2|y)}\right] p(y|H_1)\, dy + \int \left[\log \frac{p(H_2|y)}{p(H_1|y)}\right] p(y|H_2)\, dy. \end{aligned} \tag{7.20}$$

This measure has most of the properties of a distance (Kullback, 1968). All preceding definitions generalize naturally to the situation where the observation is a vector, instead of a scalar.

Example 7.9 *Correlation versus independence under normality assumptions*

Suppose there are two random variables having distributions $X \sim N\left(0, \sigma_X^2\right)$ and $Y \sim N\left(0, \sigma_Y^2\right)$. Let H_1 be the hypothesis that the variables have a joint normal distribution with correlation ρ, whereas H_2 will pose that their distributions are independent. Now let $p(x, y)$ be the bivariate normal density and let $g(x)$ and $h(y)$ be the corresponding marginal densities. Then, using the first representation leading to (7.18),

$$\begin{aligned} I(1:2) &= \int\int \log \left[\frac{p(x,y)}{g(x)\, h(y)}\right] p(x,y)\, dx\, dy \\ &= E_{H_1}\left\{\log\left[p(x,y)\right]\right\} - E_{H_1}\left\{\log\left[g(x)\right]\right\} - E_{H_1}\left\{\log\left[h(y)\right]\right\}. \end{aligned} \tag{7.21}$$

Now the form of the bivariate normal density gives

$$\log\left[p\left(x,y\right)\right] = -\log\left(2\pi\sigma_X\sigma_Y\right) - \frac{1}{2}\log\left(1-\rho^2\right) - \frac{1}{2\left(1-\rho^2\right)}\left(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y}\right).$$

Taking expectations under H_1:

$$\begin{aligned} & E_{H_1}\left\{\log\left[p\left(x,y\right)\right]\right\} \\ &= -\log\left(2\pi\sigma_X\sigma_Y\right) - \frac{1}{2}\log\left(1-\rho^2\right) - \frac{1}{2\left(1-\rho^2\right)}\left(2-2\rho^2\right) \\ &= -\log\left(2\pi\sigma_X\sigma_Y\right) - \frac{1}{2}\log\left(1-\rho^2\right) - 1. \end{aligned} \tag{7.22}$$

Likewise,

$$\begin{aligned} E_{H_1}\left\{\log\left[g\left(x\right)\right]\right\} &= E_{H_1}\left[-\log\left(\sigma_X\sqrt{2\pi}\right) - \frac{x^2}{2\sigma_X^2}\right] \\ &= -\log\left(\sigma_X\sqrt{2\pi}\right) - \frac{1}{2}, \end{aligned} \tag{7.23}$$

and

$$E_{H_1}\left\{\log\left[h\left(y\right)\right]\right\} = -\log\left(\sigma_Y\sqrt{2\pi}\right) - \frac{1}{2}. \tag{7.24}$$

Collecting (7.22) to (7.24) in (7.21)

$$I\left(1:2\right) = -\frac{1}{2}\log\left(1-\rho^2\right). \tag{7.25}$$

Hence, the expected information per observation for discrimination is a function of the correlation coefficient. Its value ranges from 0 (when the correlation is equal to 0) to ∞ when its absolute value is 1.

Similarly,

$$\begin{aligned} I\left(2:1\right) &= \int\int \log\left[\frac{g\left(x\right)h\left(y\right)}{p\left(x,y\right)}\right] g\left(x\right)h\left(y\right)dx\,dy \\ &= E_{H_2}\left\{\log\left[g\left(x\right)\right]\right\} + E_{H_2}\left\{\log\left[h\left(y\right)\right]\right\} - E_{H_2}\left\{\log\left[p\left(x,y\right)\right]\right\}. \end{aligned} \tag{7.26}$$

Now

$$E_{H_2}\left\{\log\left[g\left(x\right)\right]\right\} = -\log\left(\sigma_X\sqrt{2\pi}\right) - \frac{1}{2},$$

$$E_{H_2}\left\{\log\left[h\left(y\right)\right]\right\} = -\log\left(\sigma_Y\sqrt{2\pi}\right) - \frac{1}{2},$$

and

$$\begin{aligned} E_{H_2}\left\{\log\left[p\left(x,y\right)\right]\right\} &= -\log\left(2\pi\sigma_X\sigma_Y\right) - \frac{1}{2}\log\left(1-\rho^2\right) \\ &\quad -\frac{1}{2\left(1-\rho^2\right)}E_{H_2}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-2\rho\frac{xy}{\sigma_X\sigma_Y}\right) \\ &= -\log\left(2\pi\sigma_X\sigma_y\right) - \frac{1}{2}\log\left(1-\rho^2\right) - \frac{1}{\left(1-\rho^2\right)}. \end{aligned}$$

Hence,

$$I\left(2:1\right) = \frac{1}{2}\log\left(1-\rho^2\right) + \frac{1}{\left(1-\rho^2\right)} - 1. \tag{7.27}$$

The mean information per observation for discrimination in favor of the independence hypothesis ranges from 0, when the correlation is null, to ∞ when $\rho = 1$. The divergence measure in (7.20) is, thus,

$$\begin{aligned} J\left(1:2\right) &= I\left(1:2\right) + I\left(2:1\right) \\ &= \frac{\rho^2}{\left(1-\rho^2\right)} \end{aligned} \tag{7.28}$$

which ranges from 0 to ∞ as well. ■

Example 7.10 *Comparing hypotheses*

Following Kullback (1968), suppose that H_2 is a set of mutually exclusive and exhaustive hypotheses, one of which must be true, and that hypothesis H_1 is a member of such a set. We wish to calculate the information in y in favor of H_1. As stated, the problem implies necessarily that

$$p\left(H_2\right) = p\left(H_2|y\right) = 1,$$

this being so because the event "H_2 is true" is certain, a priori, so it must be certain a posteriori as well. Using this in (7.17) gives

$$\log\frac{p\left(y|H_1\right)}{p\left(y|H_2\right)} = \log p\left(H_1|y\right) - \log p\left(H_1\right).$$

If the observation y "proves" that H_1 is true, i.e., that $p\left(H_1|\ y\right) = 1$, then the information in y about H_1 is

$$\log\frac{p\left(y|H_1\right)}{p\left(y|H_2\right)} = -\log p\left(H_1\right).$$

This implies that if the prior probability of H_1 is small, the information resulting from its verification is large. On the other hand, if the prior probability is large, the information is small. This is reasonable on the intuitive grounds that much is learned if an implausible proposition is found to be true, as noted in Subsection 7.4.1.

If all possible hypotheses are $H_1, H_2, \ldots, H_n$, the information in y about H_i would be $-\log p(H_i)$. The mean value (taken over the prior distribution of the hypotheses) is

$$\mathcal{H} = -\sum_{i=1}^{n} [\log p(H_i)]\, p(H_i),$$

which is the entropy of the distribution of the hypotheses. ■

Example 7.11 *Information provided by an experiment*
Let $\boldsymbol{\theta}$ be a parameter vector having some prior distribution with density $h(\boldsymbol{\theta})$. Take $g(\mathbf{y}|\boldsymbol{\theta})$ to be the density of the data vector $\mathbf{y}$ under the sampling model posed, and take $g(\mathbf{y})$ as the marginal density of the observations, that is, the average of the density of the sampling model over the prior distribution of $\boldsymbol{\theta}$; let $h(\boldsymbol{\theta}|\mathbf{y})$ be the resulting posterior density. Put

$$H_1 = \boldsymbol{\theta} \text{ and } \mathbf{y} \text{ have a joint distribution with density } f(\boldsymbol{\theta}, \mathbf{y}),$$

in which case the data have something to say about the parameters, and

$$H_2 = \boldsymbol{\theta} \text{ and } \mathbf{y} \text{ are independent.}$$

Using (7.21), the expected information per observation for discrimination in favor of H_1 is

$$\begin{aligned} I(1:2) &= \int\int \log\left[\frac{f(\boldsymbol{\theta}, \mathbf{y})}{h(\boldsymbol{\theta})\, g(\mathbf{y})}\right] f(\boldsymbol{\theta}, \mathbf{y})\, d\boldsymbol{\theta}\, d\mathbf{y} \\ &= \int \left\{ \int \log\left[\frac{h(\boldsymbol{\theta}|\mathbf{y})}{h(\boldsymbol{\theta})}\right] h(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \right\} \mathbf{g}(\mathbf{y}) d\mathbf{y}. \end{aligned} \tag{7.29}$$

This measure was termed "information about a parameter provided by an experiment" by Lindley (1956). Further,

$$\begin{aligned} I(1:2) &= \int \left\{ \int [\log h(\boldsymbol{\theta}|\mathbf{y}) - \log h(\boldsymbol{\theta})]\, h(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \right\} \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &= \int \left\{ \int [\log h(\boldsymbol{\theta}|\mathbf{y})]\, h(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \right\} \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &\quad - \int\int [\log h(\boldsymbol{\theta})]\, h(\boldsymbol{\theta}|\mathbf{y})\, \mathbf{g}(\mathbf{y}) d\boldsymbol{\theta} d\mathbf{y} \\ &= \int \left\{ \int [\log h(\boldsymbol{\theta}|\mathbf{y})]\, h(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \right\} \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &\quad - \int [\log h(\boldsymbol{\theta})]\, h(\boldsymbol{\theta}) \int \mathbf{g}(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}. \end{aligned} \tag{7.30}$$

In the preceding expression, note that $\int \mathbf{g}(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = 1$. Further, recall that

$$-\int \log h(\boldsymbol{\theta})\, h(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathcal{H}(\boldsymbol{\theta})$$

is the prior entropy**,** and that

$$-\int \log h(\boldsymbol{\theta}|\mathbf{y})\, h(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} = \mathcal{H}(\boldsymbol{\theta}|\mathbf{y})$$

is the posterior entropy. Using these definitions in (7.30), the information about $\boldsymbol{\theta}$ in the experiment can be written as

$$I(1:2) = \mathcal{H}(\boldsymbol{\theta}) - E_y\left[\mathcal{H}(\boldsymbol{\theta}|\mathbf{y})\right], \tag{7.31}$$

where the second term is the average of the posterior entropy over all possible values that the data can take, should the experiment be repeated an infinite number of times. It follows that the information in an experiment can be interpreted as the decrease in entropy stemming from having observed data. ■

7.4.6 Fisher's Information Revisited

It will be shown here that Fisher's information measure (see Chapter 3) can be derived employing the concepts of mean information for discrimination and of divergence proposed by Kullback (1968). Although only the case of a single parameter will be discussed, the developments extend to the multiparameter situation in a straightforward manner.

Suppose that θ and $\theta+\Delta\theta$ are neighboring points in the parameter space. Now consider the mean information measure given in the first line of (7.18), and put

$$I(\theta : \theta+\Delta\theta) = \int \left[\log \frac{p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta+\Delta\theta)}\right] p(\mathbf{y}|\theta)\, d\mathbf{y} \tag{7.32}$$

and

$$I(\theta+\Delta\theta : \theta) = \int \left[\log \frac{p(\mathbf{y}|\theta+\Delta\theta)}{p(\mathbf{y}|\theta)}\right] p(\mathbf{y}|\theta+\Delta\theta)\, d\mathbf{y}. \tag{7.33}$$

The divergence, as defined in (7.20), is

$$\begin{aligned} J(\theta : \theta+\Delta\theta) &= I(\theta : \theta+\Delta\theta) + I(\theta+\Delta\theta : \theta) \\ &= \int \left[\log \frac{p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta+\Delta\theta)}\right] \left[p(\mathbf{y}|\theta) - p(\mathbf{y}|\theta+\Delta\theta)\right] d\mathbf{y}. \end{aligned} \tag{7.34}$$

Recall that Fisher's measure of information about θ (here, we shall use the notation $Inf(\theta)$, to avoid confusion with the notation for mean dis-

crimination) is

$$Inf(\theta) = E_{\mathbf{y}|\theta}\left[\frac{d\log p(\mathbf{y}|\theta)}{d\theta}\right]^2$$
$$= \int\left[\frac{1}{p(\mathbf{y}|\theta)}\frac{dp(\mathbf{y}|\theta)}{d\theta}\right]^2 p(\mathbf{y}|\theta)\,d\mathbf{y}. \tag{7.35}$$

Using a Taylor series, now expand the logarithm of $p(\mathbf{y}|\theta+\Delta\theta)$ about θ as follows:

$$\log p(\mathbf{y}|\theta+\Delta\theta) \approx \log p(\mathbf{y}|\theta) + \frac{d\log p(\mathbf{y}|\theta)}{d\theta}\Delta\theta$$
$$+\frac{1}{2}\frac{d^2\log p(\mathbf{y}|\theta)}{(d\theta)^2}(\Delta\theta)^2 + \frac{1}{6}\frac{d^3\log p(\mathbf{y}|\theta)}{(d\theta)^3}(\Delta\theta)^3.$$

Hence

$$\log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta+\Delta\theta)$$
$$\approx -\left[\frac{d\log p(\mathbf{y}|\theta)}{d\theta}\Delta\theta + \frac{d^2\log p(\mathbf{y}|\theta)}{2(d\theta)^2}(\Delta\theta)^2 + \frac{d^3\log p(\mathbf{y}|\theta)}{6(d\theta)^3}(\Delta\theta)^3\right]. \tag{7.36}$$

Using this, (7.32) can now be written as

$$I(\theta:\theta+\Delta\theta) = \int\left[\log\frac{p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta+\Delta\theta)}\right]p(\mathbf{y}|\theta)\,d\mathbf{y}$$
$$\approx -\Delta\theta\int\frac{d\log p(\mathbf{y}|\theta)}{d\theta}p(\mathbf{y}|\theta)\,d\mathbf{y} - \frac{(\Delta\theta)^2}{2}\int\frac{d^2\log p(\mathbf{y}|\theta)}{(d\theta)^2}p(\mathbf{y}|\theta)\,d\mathbf{y}$$
$$-\frac{(\Delta\theta)^3}{6}\int\frac{d^3\log p(\mathbf{y}|\theta)}{(d\theta)^3}p(\mathbf{y}|\theta)\,d\mathbf{y}.$$

Under regularity conditions, as seen in Chapter 3, the first term in the preceding expression vanishes because the expected value of the score of the log-likelihood is 0. The quadratic term involves the expected value of the second derivatives of the log-likelihood, or the negative of Fisher's information measure. Further, the cubic term can be neglected if the expansion is up to second-order (provided that the expected value of the third derivatives is bounded). Then, up to second-order,

$$I(\theta:\theta+\Delta\theta) \approx -\frac{(\Delta\theta)^2}{2}\int\frac{d^2\log p(\mathbf{y}|\theta)}{(d\theta)^2}p(\mathbf{y}|\theta)\,d\mathbf{y}$$
$$= \frac{(\Delta\theta)^2 Inf(\theta)}{2}. \tag{7.37}$$

Hence, the mean information for discrimination in favor of θ is proportional to Fisher's information measure and to the difference between the neighboring values of the parameter.

Now consider the divergence in (7.34)

$$J(\theta : \theta + \Delta\theta) = \int \left[\log \frac{p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta + \Delta\theta)}\right] [p(\mathbf{y}|\theta) - p(\mathbf{y}|\theta + \Delta\theta)]\, d\mathbf{y}$$
$$= \int \left[\log \frac{p(\mathbf{y}|\theta + \Delta\theta)}{p(\mathbf{y}|\theta)}\right] \left[\frac{p(\mathbf{y}|\theta + \Delta\theta) - p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta)}\right] p(\mathbf{y}|\theta)\, d\mathbf{y}. \tag{7.38}$$

Now

$$\log \frac{p(\mathbf{y}|\theta + \Delta\theta)}{p(\mathbf{y}|\theta)} = \log \left[1 + \frac{p(\mathbf{y}|\theta + \Delta\theta) - p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta)}\right]$$
$$\approx \frac{p(\mathbf{y}|\theta + \Delta\theta) - p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta)},$$

with this result following from an expansion of $\log(1 + x)$ about $x = 0$. Here the role of x is played by the relative difference between densities at the neighboring points, which is near zero, by assumption. Using the preceding result in (7.38)

$$J(\theta : \theta + \Delta\theta) \approx \int \left[\frac{p(\mathbf{y}|\theta + \Delta\theta) - p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta)}\right]^2 p(\mathbf{y}|\theta)\, d\mathbf{y}. \tag{7.39}$$

An additional approximation results from noting that, by definition,

$$\frac{dp(\mathbf{y}|\theta)}{d\theta} = \frac{p(\mathbf{y}|\theta + \Delta\theta) - p(\mathbf{y}|\theta)}{\Delta\theta},$$

as $\Delta\theta \to 0$. Making use of this in (7.39)

$$J(\theta : \theta + \Delta\theta) \approx \int \left[\frac{1}{p(\mathbf{y}|\theta)} \frac{dp(\mathbf{y}|\theta)}{d\theta} \Delta\theta\right]^2 p(\mathbf{y}|\theta)\, d\mathbf{y}$$
$$= (\Delta\theta)^2 \int \left[\frac{d \log p(\mathbf{y}|\theta)}{d\theta}\right]^2 p(\mathbf{y}|\theta)\, d\mathbf{y} = (\Delta\theta)^2 Inf(\theta). \tag{7.40}$$

Therefore the discrepancy is proportional to the square of the difference between the neighboring points and to Fisher's information measure.

7.4.7 Prior and Posterior Discrepancy

It is instructive to evaluate how much information is gained (equivalently, how much the posterior differs from the prior) as the process of Bayesian learning proceeds. A measure of "distance" between the prior and posterior distributions is given by the Kullback–Leibler discrepancy between the two

corresponding densities. Letting $p_0(\boldsymbol{\theta})$ and $p_1(\boldsymbol{\theta})$ be the prior and posterior densities, respectively, the discrepancy is (O'Hagan, 1994)

$$\begin{aligned} D(p_0, p_1) &= \int \cdots \int \left(\log \frac{p_1}{p_0}\right) p_1 \, d\boldsymbol{\theta} \\ &= E\left(\log \frac{p_1}{p_0}\right) \qquad (7.41) \\ &= -E\left(\log \frac{p_0}{p_1}\right), \qquad (7.42) \end{aligned}$$

with the expectation taken over the posterior distribution. Note that the Kullback–Leibler distance can be viewed as a relative entropy. This has two advantages over absolute entropy. First, as shown below, this relative entropy is always at least zero, contrary to the plain entropy of a continuous distribution; see Example 7.6 for an illustration of this problem. Second, relative entropy is invariant under transformation (Jaynes, 1994). Note that a change of variables from θ to λ (it suffices to consider the scalar situation to see that this holds true), with p_1^* and p_0^* representing the new densities, gives

$$\begin{aligned} D(p_0^*, p_1^*) &= -\int \log \left(\frac{p_1 \frac{d\theta}{d\lambda}}{p_0 \frac{d\theta}{d\lambda}}\right) p_1 \frac{d\theta}{d\lambda} d\lambda \\ &= D(p_0, p_1), \end{aligned}$$

since the differentials "cancel out".

As noted above, the discrepancy is greater than or equal to 0, being null only when the data do not contribute any information about the parameter, as $p_1 = p_0$ in this case. In order to show that $D(\cdot)$ is at least 0, recall Jensen's inequality (Subsubsection 3.7.1 in Chapter 3), stating that for a convex function g,

$$g[E(X)] \leq E[g(X)].$$

In our context, for g being the logarithmic function, this yields

$$\log E\left(\frac{p_1}{p_0}\right) \leq E\left(\log \frac{p_1}{p_0}\right).$$

Also,

$$\log E\left(\frac{p_0}{p_1}\right) \leq E\left(\log \frac{p_0}{p_1}\right).$$

Now

$$\begin{aligned} \log E\left(\frac{p_0}{p_1}\right) &= \log\left[\int \cdots \int \frac{p_0}{p_1} p_1 d\boldsymbol{\theta}\right] \\ &= \log\left[\int \cdots \int p_0 \, d\boldsymbol{\theta}\right] = \log(1) = 0, \end{aligned}$$

provided the prior is proper, so it integrates to 1. Hence, the Kullback–Leibler discrepancy is null or positive.

When the observations are conditionally independent and sample size is n, the Kullback–Leibler distance can be expressed as:

$$D(p_0, p_1) = \int \cdots \int \left(\log \frac{p_1}{p_0} \right) p_1 \, d\boldsymbol{\theta}$$

$$= \int \cdots \int \left[\log \frac{c_n g(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right] p_1 \, d\boldsymbol{\theta}$$

$$= \log(c_n) + \int \cdots \int \left[\log \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) \right] p_1 \, d\boldsymbol{\theta}$$

$$= \log(c_n) + \sum_{i=1}^{n} E\left[\log p(y_i|\boldsymbol{\theta}) \right], \tag{7.43}$$

where $p_0 = g(\boldsymbol{\theta})$ is the prior density and c_n is the integration constant of the posterior density based on n observations. Recall that the expectation in (7.43) is taken over the posterior distribution.

Example 7.12 *Kullback–Leibler distance in a truncated normal model* Consider again the truncated normal model in Example 7.8. Then, from (7.43),

$$D(p_0, p_1) = \log(c_n) - \frac{n \log\left(2\pi\sigma^2\right)}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} E(y_i - \mu)^2, \tag{7.44}$$

with the expectation taken over the posterior distribution of μ. From (7.16), in the truncated normal model, one has

$$E(y_i - \mu)^2 = \gamma + (\eta - y_i)^2.$$

Using this in (7.44), the Kullback–Leibler discrepancy is

$$D(p_0, p_1) = \log(c_n) - \frac{n \log\left(2\pi\sigma^2\right)}{2} - \frac{n\gamma + \sum_{i=1}^{n} (\eta - y_i)^2}{2\sigma^2}$$

$$= \log(c_n) - \frac{n \log\left(2\pi\sigma^2\right)}{2} - \frac{n\left[\gamma + (\eta - \overline{y})^2\right] + \sum_{i=1}^{n} (y_i - \overline{y})^2}{2\sigma^2}.$$

Making use of the integration constant given in (7.15), and with $b = \infty$, the distance between the prior and the posterior distributions can be expressed

as

$$D(p_0, p_1) = -\frac{1}{2}\log\left(2\pi\sigma^2/n\right) - \log\left[1 - \Phi\left(\frac{a-\overline{y}}{\sigma/\sqrt{n}}\right)\right]$$
$$-\frac{1}{2}n\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\{n\left[\gamma + (\eta - \overline{y})^2\right] + \sum_{i=1}^{n}(y_i - \overline{y})^2\right\}.$$

■

7.5 Priors Conveying Little Information

One of the criticisms often made of Bayesian inference is the potential effect that a possibly subjective, arbitrary or misguided prior can have on inferences. Hence, efforts have been directed at arriving at "objective priors", by this meaning prior distributions that say "little" relative to the contributions made by the data (Jeffreys, 1961; Box and Tiao, 1973; Zellner, 1971). It has been seen already that the effect of the prior dissipates as sample size increases, so the problem is essentially one affecting finite sample inferences. Further, when the observations in the sample are correlated or when the model involves many parameters, it is not always clear how large the sample should be for any influences of the prior to be overwhelmed by the data. The problem of finding objective or noninformative priors is an extremely difficult one, and consensus seems to be lacking between researchers that have worked in this area. Here we will present a short review of some of the approaches that have been suggested. For additional detail, see Bernardo (1979), Berger and Bernardo (1992), Bernardo and Smith (1994), O'Hagan (1994), and Leonard and Hsu (1999).

7.5.1 The Uniform Prior

The most widely used (and abused) "noninformative" prior is that based on the Bayes-Laplace "principle of insufficient reason". This states that, in the absence of evidence to the contrary, all possibilities should have the same prior probability (e.g., Bernardo and Smith, 1994). For example, if θ takes one of K possible values, the noninformative prior indicated by this principle is the uniform distribution

$$\left\{\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right\}.$$

As noted in Example 7.5 this is also a maximum entropy distribution when all that is known is that there are K mutually exclusive and exhaustive states.

In the continuous case the counterpart is the continuous uniform distribution, but this leads to inconsistencies. Suppose that θ is assigned a uniform prior distribution to convey lack of knowledge about the values of this parameter. Then, the density of the distribution of a parameter resulting from a monotone transformation $\lambda = f(\theta)$ is

$$\begin{aligned} p(\lambda) &= p\left[f^{-1}(\lambda)\right]\left|\frac{f^{-1}(\lambda)}{d\lambda}\right| \\ &\propto \left|\frac{f^{-1}(\lambda)}{d\lambda}\right|. \end{aligned}$$

If the transformation is linear, the Jacobian is a constant, so it follows that the density of λ is uniform as well. On the other hand, if the transformation is nonlinear, the density varies with λ. This implies that if one claims ignorance with respect to θ, the same cannot be said about λ. This is a severe inconsistency.

Example 7.13 *The improper uniform prior as a limiting case of the normal distribution*

Let the prior be $\theta \sim N\left(\mu_\theta, \sigma_\theta^2\right)$. If σ_θ^2 is very small, this implies that the values are concentrated around μ_θ or, equivalently, sharp prior knowledge. On the other hand, if there is large prior uncertainty, as measured by a large variance, the distribution becomes dispersed. As σ_θ^2 increases, the normal distribution gets flatter and flatter and, in the limit, it degenerates to a uniform distribution between $-\infty$ and ∞. This is an improper distribution, as the integral of the density is not finite. However, a posterior distribution can be proper even if the prior is improper. For example, take $y_i \sim N\left(\mu, \sigma^2\right)$, $(i = 1, 2, \ldots, n)$ to be a sample of i.i.d. random variables with known variance. Adopting as prior

$$p(\mu) \propto \text{constant},$$

generates an improper distribution, unless finite boundaries are assigned to the values of μ. However, it is easy to verify that the posterior density is always proper, since

$$p(\mu|y_1, y_2, \ldots, y_n) \propto \exp\left[-\frac{n}{2\sigma^2}(\mu - \overline{y})^2\right],$$

integrates to $\sqrt{2\pi\sigma^2/n}$. ■

Example 7.14 *A uniform prior distribution for heritability*

Consider the following example, in the same spirit as Examples 2.20 and 2.21 of Chapter 2. Suppose that in a linear model the residual variance is known. Heritability, h^2, is defined as usual, but observations have been rescaled (by dividing by the residual standard deviation) so that one can

write

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} = \frac{\sigma_a^{2*}}{\sigma_a^{2*} + 1},$$

where $\sigma_a^{2*} = \sigma_a^2/\sigma_e^2$. Suppose that all we know, a priori, is that h^2 is between 0 and 1. A uniform prior is assigned to this parameter, following the principle of insufficient reason. Hence, the prior probability that h^2 is smaller than or equal to $\frac{1}{4}$ is $\frac{1}{4}$, and so is the prior probability that it is larger than or equal to $\frac{3}{4}$.

What is the implied prior distribution of the ratio of variances σ_a^{2*}? Note that $\sigma_a^{2*} = h^2/\left(1 - h^2\right)$, so that the parameter space is $(0, \infty)$. After calculating the Jacobian of the transformation, the density of the prior distribution is

$$p\left(\sigma_a^{2*}\right) = \frac{1}{\left(\sigma_a^{2*} + 1\right)^2}.$$

Now, the statement "heritability is smaller than or equal to $\frac{1}{5}$" is equivalent to the statement that σ_a^{2*} is smaller than or equal to $\frac{1}{4}$. The resulting probability is

$$\begin{aligned}\Pr\left(\sigma_a^{2*} \leq \frac{1}{4}\right) &= \int_0^{\frac{1}{4}} \frac{1}{\left(\sigma_a^{2*} + 1\right)^2} d\sigma_a^{2*} \\ &= \left.\frac{-1}{\sigma_a^{2*} + 1}\right|_0^{\frac{1}{4}} = \frac{1}{5}.\end{aligned}$$

Also, the probability of the statement "heritability is smaller than or equal to $\frac{2}{5}$" is equivalent to

$$\Pr\left(\sigma_a^{2*} \leq \frac{2}{3}\right) = \left.\frac{-1}{\sigma_a^{2*} + 1}\right|_{\frac{2}{3}}^{\infty} = \frac{2}{5}.$$

Hence, while the prior distribution of heritability assigns equal probability to intervals of equal length, this is not so for the induced prior distribution of σ_a^{2*}. This implies that prior indifference about heritability (as reflected by the uniform prior) does not translate into prior indifference about the scaled additive genetic variance. ■

7.5.2 Other Vague Priors

Although the uniform distribution is probably the most widely employed "vague" prior, other distributions that supposedly convey vague prior knowledge have been suggested. For example, imagine one seeks to infer the probability of success in a binomial distribution (θ) and that a $Be\left(\theta|a, b\right)$

distribution is used as prior for θ. If x is the number of successes after n independent Bernoulli trials, the posterior density of θ is

$$p(\theta|n,x,a,b) \propto \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1}.$$

Since $a+b$ can be interpreted as "the size of a prior sample", one can take $a=b=0$, yielding an improper prior distribution. However, one obtains as posterior density

$$p(\theta|n,x,a=0,b=0) \propto \theta^{x-1} (1-\theta)^{n-x-1},$$

which is a $Be(\theta|x,n-x)$ density function. For this distribution to be proper, the two parameters must be positive. Hence, if either $x=0$ or $x=n$, the posterior distribution is improper. For example, if θ is small, it is not unlikely that the posterior distribution turns out to be improper unless n is large. The preceding illustrates that an improper prior can lead to an improper posterior, and that caution must be exercised when using this form of prior assignment.

Example 7.15 *Haldane's analysis of mutation rate*
Haldane (1948) studied the problem of inferring mutation rates in a population. He noted that since a mutation is a rare event, the sampling distribution of the relative frequencies (number of mutants/number of individuals scored) is skewed. Haldane considered using a prior distribution for the mutation rate θ. If $p(\theta|H)$ is some prior density (where H denotes hyperparameters) and x mutants are observed out of n individuals, the typical Bernoulli sampling model produces as posterior density of θ,

$$p(\theta|n,x,H) = \frac{\theta^x (1-\theta)^{n-x} p(\theta|H)}{\int_0^1 \theta^x (1-\theta)^{n-x} p(\theta|H)\, d\theta}.$$

Observe that the uniform prior corresponds to a $Be(1,1)$ distribution, in which case the posterior is always proper. Haldane noted that if the uniform prior is adopted for θ, the posterior density is in the form

$$Be(\theta|x+1,n-x+1),$$

corresponding to a Beta distribution, with mean $(x+1)/(n+2)$. If the posterior mean is used as a point estimator, the statistic is biased in the frequentist sense, a fact found objectionable by Haldane. He argued that assuming a uniform prior does not make sense because θ should be greater than 10^{-20} and lower than 10^{-3}; in some particular cases, θ would be as likely to be between 10^{-6} and 10^{-7} as between 10^{-6} and 10^{-5}. He suggested the prior

$$p(\theta) \propto \frac{1}{\theta(1-\theta)},$$

which is the improper $Be(0,0)$ distribution. The posterior distribution of the mutation rate is then $Be(x, n-x)$, with mean x/n, satisfying Haldane's requirement of unbiasedness. As noted above, this posterior can be improper, but Haldane (without elaboration) cautioned that his prior would "work" only if $x > 0$ (at least one mutant is observed) and if $x < n$ (which would be almost certain, unless n is extremely small and by sheer accident all individuals scored are mutant). He stated that for small $x/\sqrt{n}$ the posterior distribution would be well-approximated by a gamma process with density

$$p(\theta|n,x) \propto \theta^{x-1} \exp(-n\theta),$$

which also has mean x/n. Haldane (1948) showed then that a cube root transformation of θ would have a nearly normal distribution. Note that Haldane's "vague" prior for θ implies an uniform prior for the logit transformation. Let

$$\lambda = \log \frac{\theta}{1-\theta},$$

with inverse transformation

$$\theta = \frac{\exp(\lambda)}{1+\exp(\lambda)}$$

and Jacobian

$$\frac{d\theta}{d\lambda} = \frac{\exp(\lambda)}{[1+\exp(\lambda)]^2}.$$

Observe that the logit can take any value in the real line. Then the induced prior density of λ is

$$p(\lambda) \propto \left[\frac{\exp(\lambda)}{1+\exp(\lambda)}\right]^{-1} \left[\frac{1}{1+\exp(\lambda)}\right]^{-1} \frac{\exp(\lambda)}{[1+\exp(\lambda)]^2} = 1.$$

This uniform distribution is improper because the integral of the density over the entire parameter space is not finite. ■

A Single Parameter

Although Jeffreys (1961) proposed a class of improper priors for representing ignorance, it is not entirely transparent why these should be considered "noninformative" in some precise sense. Zellner (1971) and Box and Tiao (1973) provide some heuristic justifications, based on the two rules suggested by Jeffreys:
(1) If a parameter θ can take any value in a finite range or between $-\infty$ and ∞, it should be taken as distributed uniformly, a priori.
(2) If it takes values between 0 and ∞, then its logarithm should have a uniform distribution.

The justification for the first rule (Zellner, 1971) is that since the range covers the entire real line, the probability that θ takes any value drawn from a uniform distribution is

$$\Pr(-\infty < \theta < \infty) = \int_{-\infty}^{\infty} d\theta = \infty,$$

and Jeffreys interprets this as the probability of the certain event (1 becomes ∞!). Then the probability that it falls in any interval (a, b) is 0, and similarly for another interval (c, d). Since the ratio of probabilities is indeterminate, Jeffreys argues that this constitutes a formal representation of ignorance, as one cannot favor an interval when choosing over any pair of finite intervals. If the uniform prior is bounded (a typical device used by quantitative geneticists to avoid improper posteriors), this implies that one knows something about the range of values and the ratios between probabilities are then determinate. However, if the boundaries are stretched, the ratio of probabilities becomes indeterminate in the limit, thus satisfying Jeffreys' requirements. Concerning the second rule, note that if $\log\theta$ is taken as uniformly distributed, the prior density of θ should be proportional to $1/\theta$. In this setting (Zellner, 1971), one has

$$\int_0^{\infty} \frac{1}{\theta} d\theta = \infty, \quad \int_0^{a} \frac{1}{\theta} d\theta = \infty, \quad \int_a^{\infty} \frac{1}{\theta} d\theta = \infty.$$

It ∞ represents certainty, then the ratio between the last two integrals is indeterminate, leading to a representation of ignorance: one cannot pick the interval in which θ is more likely to reside.

An important step toward the development of noninformative priors was the introduction of invariance requirements by Jeffreys (1961). In a nutshell, the central point is to recognize that ignorance about θ implies ignorance about the monotone transformation $\lambda = f(\theta)$. Hence, if the prior density of θ is $p(\theta)$, the requirement is that the probability contents must be preserved, that is, $p(\theta)\, d\theta = p(\lambda)\, d\lambda$, which leads directly to the usual formula for the density of a transformed random variable. Jeffreys' idea was to use, as prior density of θ, the square root of Fisher's information measure

$$\begin{aligned} \sqrt{I(\theta)} &= \sqrt{E\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]^2} \\ &= \sqrt{-E\left[\frac{d^2 l(\theta|\mathbf{y})}{(d\theta)^2}\right]}, \end{aligned} \tag{7.45}$$

where $l(\theta|\mathbf{y})$ is the log-likelihood function. Making a change of variables, consider the prior density of λ induced by Jeffreys' rule

$$\begin{aligned} p(\lambda) &\propto \sqrt{E\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]^2}\frac{d\theta}{d\lambda} \\ &\propto \sqrt{E\left[\frac{dl(f^{-1}(\lambda)|\mathbf{y})}{d\theta}\frac{d\theta}{d\lambda}\right]^2} \\ &\propto \sqrt{E\left[\frac{dl(f^{-1}(\lambda)|\mathbf{y})}{d\lambda}\right]^2} = \sqrt{I(\lambda)}, \end{aligned}$$

recalling that the likelihood is invariant under a change in parameterization. One can transform back to θ, and retrieve $\sqrt{I(\theta)}$ as prior distribution. Hence, Jeffreys' prior is invariant under transformation. The posterior density of θ has the form

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto L(\theta|\mathbf{y})\sqrt{I(\theta)} \\ &\propto L(\theta|\mathbf{y})\sqrt{E\left[\frac{dl(\theta|\mathbf{y})}{d\theta}\right]^2}, \end{aligned} \tag{7.46}$$

where $L(\theta|\mathbf{y})$ is the likelihood. The posterior density of λ is then

$$\begin{aligned} p(\lambda|\mathbf{y}) &\propto L[\lambda|\mathbf{y}]\sqrt{E\left[\frac{dl(f^{-1}(\lambda)|\mathbf{y})}{d\theta}\right]^2}\frac{d\theta}{d\lambda} \\ &\propto L(\lambda|\mathbf{y})\sqrt{I(\lambda)}. \end{aligned} \tag{7.47}$$

One can now go back and forth between parameterizations, and preserve the probability contents. Some examples of Jeffreys' rule are presented subsequently.

At first sight, Jeffreys' prior seems to depend on the data, since it involves the likelihood function. Actually, it does not; rather, it depends on the form of the "experiment". Note that in the process of taking expectations, the dependence on the observed data is removed. However, the prior depends on how the data are to be collected. From an orthodox Bayesian point of view, prior information should not be affected by the form of the experiment or by how much data is to be collected, but this is not the case with Jeffreys' prior. Box and Tiao (1973) argue that a noninformative prior does not necessarily represent a prior opinion and that "knowing little" is meaningful only relative to a specific experiment. Thus, the form of a noninformative prior would depend on the experiment, and two different experiments would lead to different noninformative priors.

Example 7.16 *Normal distribution with unknown mean*
Let there be n samples from the same normal distribution with unknown

mean but known variance. The log-likelihood is

$$l\left(\mu|\sigma^2,\mathbf{y}\right) = \text{constant} - \frac{n\left(\overline{y}-\mu\right)^2}{2\sigma^2}.$$

The square of the score with respect to μ is

$$\left[\frac{dl\left(\mu|\sigma^2,\mathbf{y}\right)}{d\mu}\right]^2 = \left[\frac{n\left(\overline{y}-\mu\right)}{\sigma^2}\right]^2,$$

and its expectation is

$$I\left(\mu\right) = \frac{n}{\sigma^2}.$$

Jeffreys' prior is the square root of this expression. Since it does not involve μ, it follows that the noninformative prior is flat and improper. ■

Example 7.17 *Normal distribution with unknown variance*
The setting is as in the previous example, but now we seek Jeffreys' prior for the variance. The log-likelihood is now

$$l\left(\sigma^2|\mu,\mathbf{y}\right) = \text{constant} - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}.$$

Differentiating twice with respect to σ^2 and multiplying by -1 to obtain the observed Fisher's information gives

$$-\frac{n}{2\sigma^4} + \frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{\sigma^6}.$$

Taking expectations and then the square root yields as Jeffreys' prior

$$p\left(\sigma^2\right) \propto \sqrt{\frac{n}{2\sigma^4}} \propto \frac{1}{\sigma^2},$$

which is in conformity with the second rule discussed above. Since Jeffreys' prior is invariant under reparameterization one obtains for the standard deviation,

$$p\left(\sigma\right) \propto \frac{1}{\sigma^2}\frac{d\sigma^2}{d\sigma} \propto \frac{1}{\sigma},$$

whereas for the logarithm of the standard deviation, the prior is

$$\begin{aligned} p\left[\log\left(\sigma\right)\right] &\propto \frac{1}{\sigma}\frac{d\left(\exp^{\log(\sigma)}\right)}{d\log\left(\sigma\right)} \\ &\propto \frac{1}{\sigma}\exp^{\log(\sigma)} \propto 1, \end{aligned}$$

which is the improper uniform prior. ■

Example 7.18 *Exponential distribution*
As in Leonard and Hsu (1999), draw a random sample of size n from an exponential distribution and take its parameter θ to be unknown. The likelihood function is

$$L(\theta|\mathbf{y}) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} y_i\right).$$

The second derivative of the log-likelihood with respect to θ is $-n/\sigma^2$. Since this does not depend on the observations, Jeffreys' prior is

$$p(\theta) \propto \sqrt{\frac{n}{\theta^2}} \propto \frac{1}{\theta}.$$

The posterior distribution is then

$$p(\theta|\mathbf{y}) \propto \theta^{n-1} \exp\left(-\theta \sum_{i=1}^{n} y_i\right),$$

which is a $Ga\left(n, \sum_{i=1}^{n} y_i\right)$ process. ■

Many Parameters

Consider now a multi-parameter situation. Jeffreys' rule generalizes to:

$$p(\boldsymbol{\theta}) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta})|}, \tag{7.48}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher's information matrix about the $p \times 1$ parameter vector $\boldsymbol{\theta}$. The multiparameter rule has often been criticized in the Bayesian literature because of "inconsistencies" (e.g., O'Hagan, 1994), or due to "intuitively unappealing implications" (Bernardo and Smith, 1994). The latter objection stems from the fact that when the rule is applied to certain problems, it does not yield results that are equivalent to their frequentist counterparts, or because no account is taken of degrees of freedom lost. We give an example where the rule leads to an objectionable result.

Example 7.19 *Jeffreys' rule in a regression model*
Consider the usual linear regression model under normality assumptions. The unknown parameters are the regression coefficients $\boldsymbol{\beta}$ $(p \times 1)$ and the residual variance σ^2. As shown in Chapter 3, the expected information matrix is

$$I\left(\boldsymbol{\beta}, \sigma^2\right) = \begin{bmatrix} \dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{2\sigma^4} \end{bmatrix}.$$

According to Jeffreys' rule (7.48), the joint prior density of the parameters is

$$p\left(\boldsymbol{\beta}, \sigma^2\right) \propto \left| \begin{bmatrix} \dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{2\sigma^4} \end{bmatrix} \right|^{\frac{1}{2}} \propto \left(\sigma^2\right)^{-\frac{p+2}{2}}. \tag{7.49}$$

When $\boldsymbol{\beta}$ contains a sole component (a mean, μ, say), the prior reduces to σ^{-3}. We now further develop this simpler situation and examine the marginal posterior distribution of σ^2. For $p = 1$, the joint posterior based on a sample of size n is

$$\begin{aligned} p\left(\mu, \sigma^2|\mathbf{y}\right) &\propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}\right]\left(\sigma^2\right)^{-\frac{3}{2}} \\ &\propto \left(\sigma^2\right)^{-\frac{n+3}{2}} \exp\left[\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{2\sigma^2}\right] \exp\left[-\frac{n\left(\overline{y}-\mu\right)^2}{2\sigma^2}\right]. \end{aligned} \tag{7.50}$$

Integration over the mean, μ, leads to the following marginal density of σ^2 (after rearrangement):

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\left(\frac{n}{2}+1\right)} \exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{2\sigma^2}\right]. \tag{7.51}$$

Change now variables to

$$\chi^2 = \frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{\sigma^2}.$$

In a frequentist setting, this is a central chi-square random variable on $n-1$ degrees of freedom. From a Bayesian point of view, we need to consider its posterior distribution. Noting that the absolute value of the Jacobian of the transformation is $\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\chi^{-4}$, the posterior density of χ^2 is

$$p\left(\chi^2|\mathbf{y}\right) \propto \left(\chi^2\right)^{\frac{n}{2}-1} \exp\left(-\frac{\chi^2}{2}\right),$$

indicating a chi-square distribution on n degrees of freedom. Since "everybody knows" that estimating the mean consumes one degree of freedom, it becomes apparent that Jeffreys' multiparameter prior does not take into account the loss of information incurred.

On the other hand, suppose one employs Jeffreys' rule for each of μ and σ^2 separately (assuming that the other parameter is known in each case), and takes as joint prior

$$p\left(\mu,\sigma^2\right) \propto p\left(\mu|\sigma^2\right) p\left(\sigma^2|\mu\right) \propto \frac{1}{\sigma^2}.$$

Replacing $\left(\sigma^2\right)^{-\frac{3}{2}}$ in (7.50) by $\left(\sigma^2\right)^{-1}$ and integrating over μ yields an expression that differs from (7.51) in the exponent of the first term, which is instead $-[(n+1)/2]$. After transforming as before, the posterior density of χ^2 is now

$$p\left(\chi^2|\mathbf{y}\right) \propto \left(\chi^2\right)^{\frac{n-1}{2}-1} \exp\left(-\frac{\chi^2}{2}\right).$$

Here, one arrives at the "correct" posterior distribution of chi-square, i.e., one on $n-1$ degrees of freedom. There is no formal justification for such prior, but "it works", at least in the sense of coinciding with the frequentist treatment, which was one of Jeffreys' objectives. Box and Tiao (1973) recommend exercising "special care" when choosing noninformative priors for location and scale parameters simultaneously.

The problems of the multiparameter rule become even more serious when the model is even more richly parameterized. Suppose now that $\boldsymbol{\beta}$ contains two means (μ_1 and μ_2). Using (7.49), the joint prior $p\left(\mu_1,\mu_2,\sigma^2\right)$ turns out to be proportional to σ^{-4}. The joint posterior density can be written as

$$p\left(\mu_1,\mu_2,\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n_1+n_2+4}{2}} \exp\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2}{2\sigma^2}\right]$$
$$\times \prod_{i=1}^{2} \exp\left[-\frac{n_i\left(\overline{y}_i-\mu_i\right)^2}{2\sigma^2}\right],$$

where n_i is the number of observations associated with mean i; the rest of the notation is clear from the context. Integrating over the two means yields as marginal posterior density of σ^2,

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-\frac{n_1+n_2+2}{2}} \exp\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2}{2\sigma^2}\right].$$

Now put

$$\chi^2 = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2}{\sigma^2}.$$

This variable, in a frequentist setting, should have a chi-squared distribution on $n_1 + n_2 - 2$ degrees of freedom. Changing variables, the posterior density of χ^2 becomes

$$p\left(\chi^2|\mathbf{y}\right) \propto \left(\chi^2\right)^{\frac{n_1+n_2}{2}-1} \exp\left(-\frac{\chi^2}{2}\right).$$

Again, the multiparameter rule does not take into account the fact that several means appear as nuisance parameters: the degrees of freedom of the chi-square distribution are left intact at $n = n_1 + n_2$. Jeffreys (1961), in the context of testing location parameters, writes: "The index in the corresponding t distribution would always be $(n+1)/2$ however many true values were estimated. This is unacceptable". We agree. ■

7.5.3 Maximum Entropy Prior Distributions

Using the principle of maximum entropy as a means of allocating probabilities in a prior distribution was suggested by Jaynes, a physicist. A comprehensive account of the idea is in his unfinished book (Jaynes, 1994).

Suppose one wishes to assign a prior distribution to some unknown quantity. Naturally, this prior should take into account whatever information is available, but not more. For example, knowledge of average values (or of other aspects of the distribution) will give a reason for favoring some possibilities over others but, beyond this, the distribution should be as uncommitted as possible. Further, no possibilities should be ruled out, unless dictated by prior knowledge. The information available defines constraints that fix some properties of the prior distribution, but not all of them. Jaynes formulates the problem as follows:

> "To cast it into mathematical form, the aim of avoiding unwarranted conclusions leads us to ask whether there is some reasonable numerical measure of how uniform a probability distribution is, which the robot could maximize subject to constraints which represent its available information."

Jaynes used Shannon's (1948) theorem, which states that the only measure of the uncertainty represented by a probability distribution is entropy. Then he argued that a distribution maximizing entropy, subject to the constraints imposed by the information available, would represent the "most honest" description of what is known about a set of propositions. He noted that the only source of arbitrariness is the base of the logarithms employed in entropy. However, since this operates as a multiplicative constant in the expression for entropy, it has no effect on the values of the probabilities that maximize H; see, e.g., the form of entropy in (7.9). A derivation of the principle, as in Jaynes (1994) and in Sivia (1996), follows.

Discrete Case

Let I represent the information to be used for assigning probabilities

$$\{p_1, p_2, ..., p_K\}$$

to K different possibilities. Suppose there are $n >> K$ small "quanta" of probability (so that n is very large) to distribute in any way one sees fit. If the quanta are tossed completely at random (reflecting lack of information in the process of allocating quanta), so that each of the options has an equal probability $1/K$ of getting a quanta, the allocation would be viewed as a fair one. At the end of the experiment, it is observed that option i has received n_i quanta, and so on, so that the experiment has generated the probability assignment

$$p_i = \frac{n_i}{n}; \quad i = 1, 2, \ldots, K.$$

The probability that this particular allocation of quanta will be observed is given by the multinomial distribution

$$\begin{aligned} W &= \frac{n!}{n_1!\, n_2! \ldots n_K!} \left(\frac{1}{K}\right)^{n_1} \left(\frac{1}{K}\right)^{n_2} \ldots \left(\frac{1}{K}\right)^{n_K} \\ &= \frac{n!}{n_1!\, n_2! \ldots n_K!} K^{-\sum_{i=1}^{K} n_K} = \frac{n!}{n_1!\, n_2! \ldots n_K!} K^{-n}, \end{aligned} \tag{7.52}$$

where $n = \sum_{i=1}^{K} n_i$. The experiment is repeated over and over, and the probability assignment is examined to see if it is consistent with the information I; if not, the assignment is rejected. What is the most likely probability distribution resulting from the experiment? In order to find it, use will be made of Stirling's approximation to factorials (e.g., Abramowitz and Stegun, 1972) for large n:

$$\log n! \approx n \log n - n.$$

Employing this in (7.52), one can write

$$\begin{aligned} \frac{1}{n} \log W &= \frac{1}{n} \left[n \log n - n - \sum_{i=1}^{K} (n_i \log n_i - n_i) - n \log K \right] \\ &= \log n - \sum_{i=1}^{K} \left(\frac{n_i}{n} \log n_i \right) - \log K. \end{aligned}$$

Since $p_i = n_i/n$,

$$\begin{aligned}\frac{1}{n}\log W &= \log n - \sum_{i=1}^{K}(p_i \log np_i) - \log K \\ &= \text{constant} - \sum_{i=1}^{K}(p_i \log p_i) \\ &= \text{constant} + H(p_1, p_2, ..., p_K). \end{aligned} \tag{7.53}$$

Maximizing W with respect to p_i is equivalent to maximizing $(1/n)\log W$, and the preceding expression indicates that this is achieved when the entropy $H(p_1, p_2, ..., p_K)$ is extremized with respect to the probabilities, subject to any constraints imposed by the available information I.

Now consider finding an explicit representation of the maximum entropy prior distribution of a discrete random variable θ with K mutually exclusive and exhaustive states θ_i. The information available is that the sum of the probabilities must be equal to 1 and assume, further, that the mean of the prior distribution to be specified, i.e., $\overline{\theta} = \sum_{i=1}^{K}\theta_i p_i$, is taken to be known. This knowledge must be incorporated in the maximization problem as a Lagrangian condition. Hence, the objective function to be extremized is:

$$\begin{aligned}&H^*(p_1, p_2, \ldots, p_K, \lambda_0, \lambda_1) \\ &= -\sum_{i=1}^{K}(p_i \log p_i) + \lambda_0\left(\sum_{i=1}^{K} p_i - 1\right) + \lambda_1\left(\sum_{i=1}^{K}\theta_i p_i - \overline{\theta}\right), \end{aligned} \tag{7.54}$$

where λ_0 and λ_1 are Lagrange multipliers ensuring that the two information constraints are observed. The multipliers ensure propriety and knowledge of the mean of the prior distribution, respectively. Differentiation with respect to the $K+2$ unknowns yields

$$\frac{\partial H^*(p_1, p_2, ..., p_K, \lambda_0, \lambda_1)}{\partial p_i} = -\log p_i - 1 + \lambda_0 + \lambda_1\theta_i, \quad i = 1, 2, ..., K,$$

$$\frac{\partial H^*(p_1, p_2, ..., p_K, \lambda_0, \lambda_1)}{\partial \lambda_0} = \sum_{i=1}^{K} p_i - 1,$$

$$\frac{\partial H^*(p_1, p_2, ..., p_K, \lambda_0, \lambda_1)}{\partial \lambda_1} = \sum_{i=1}^{K}\theta_i p_i - \overline{\theta}.$$

The differentials are set to 0 simultaneously and solved for the unknowns. The p_i equation yields

$$p_i = \exp(\lambda_1\theta_i)\exp(\lambda_0 - 1).$$

Since the first Lagrangian condition dictates that the sum of the probabilities must add up to 1, the preceding must be equal to

$$p_i = \frac{\exp(\lambda_1\theta_i)\exp(\lambda_0 - 1)}{\sum_{i=1}^{K}\exp(\lambda_1\theta_i)\exp(\lambda_0 - 1)} = \frac{\exp(\lambda_1\theta_i)}{\sum_{i=1}^{K}\exp(\lambda_1\theta_i)}, \quad i = 1, 2, ..., K. \tag{7.55}$$

This is called the Gibbs distribution, and the denominator is known as the partition function (Applebaum, 1996). Expression (7.55) gives the maximum entropy prior distribution of a univariate discrete random variable with known mean. If the mean of the distribution is left unspecified, this is equivalent to removing the second Lagrangian condition in (7.54), or setting λ_1 to 0. Doing this in expression (7.55), one obtains $p_i = 1/K$, $(i = 1, 2, ..., K)$ as maximum entropy distribution. In general, as additional side conditions are introduced (as part of the prior information I), the system of equations that needs to be solved is non-linear. Zellner and Highfield (1988) discuss one of the possible numerical solutions.

Continuous Case via Discretization

When θ is continuous, the technical arguments are more involved. Further, it must be recalled that entropy is not well defined in such a setting (for an illustration, see Example 7.6). The problem now consists of finding the prior density or distribution that maximizes the entropy

$$H[p(\theta)] = -\int (\log\theta)\, p(\theta)\, d\theta,$$

subject to the constraints imposed by the available information I. The integral above is a functional, that is, a function that depends on another function, this being the density $p(\theta)$. Hence, one must find the function that extremizes the integral. First, we present a solution based on discretizing the prior distribution, and a formal calculus of variations solution is given later on.

First, return to the derivation of the principle of maximum entropy given in (7.52) and (7.53), but assume now that the chance of a quanta falling into a specific one of the K options is m_i, instead of being equal for all options (Sivia, 1996). Then the experiment generates the probability assignment

$$W^* = \frac{n!}{n_1!\, n_2! \ldots n_K!}\,(m_1)^{n_1}(m_2)^{n_2}\ldots(m_k)^{n_K},$$

which is the multinomial distribution. Next, as before, put

$$\frac{1}{n}\log W^* = \frac{1}{n}\left(\log n! - \sum_{i=1}^{K}\log n_i! + \sum_{i=1}^{K} n_i \log m_i\right),$$

and make use of Stirling's approximation, and of $p_i = n_i/n$, to arrive at

$$\frac{1}{n}\log W^* = \log n - \sum_{i=1}^{K}(p_i \log np_i) + \sum_{i=1}^{K} p_i \log m_i$$
$$= -\sum_{i=1}^{K}\left(p_i \log \frac{p_i}{m_i}\right). \tag{7.56}$$

This is a generalization of (7.53), where allowance is made for the options in the random experiment to have unequal probability. Note that (7.56) is a relative entropy, which has the advantage (in the continuous case) of being invariant under a change of variables, as noted earlier. Now the continuous counterpart of (7.56) is

$$H^*(\theta) = -\int \left[\log \frac{p(\theta)}{m(\theta)}\right] p(\theta)\, d\theta. \tag{7.57}$$

This is in the same mathematical form as Kullback's expected information for discrimination, and is also known as the Shannon-Jaynes entropy (Sivia, 1996). Here, $m(\theta)$ plays the role of a "reference" distribution, often taken to be the uniform one. In this case, (7.57) reduces to the standard entropy.

Now consider a discretization approach to finding the maximum relative entropy distribution. Suppose that θ takes appreciable density in the interval (a, b). Following Applebaum (1996), one can define a partition of (a, b) such that:

$$a = \theta_0 < \theta_1 < \cdots < \theta_{K-1} < \theta_K = b.$$

Define the event $\xi_j \in (\theta_{j-1}, \theta_j)$ for $1 \le j \le K$, and the discrete random variable θ^* with K mutually exclusive and exhaustive states and probability distribution

$$\Pr\left(\theta^* = \xi_j\right) = \int_{\theta_{j-1}}^{\theta_j} p(\theta)\, d\theta$$
$$= F(\theta_j) - F(\theta_{j-1}) = p_j, \tag{7.58}$$

where $F(\cdot)$ is the c.d.f.. The relative entropy of the discretized distribution is then

$$H(\theta^*) = -\sum_{j=1}^{K} p_j \log \frac{p_j}{m_j}.$$

Proceed now to maximize $H(\theta^*)$ with respect to p_j, subject to the Lagrangian condition that $\sum_{j=1}^{K} p_j = 1$. Then

$$\begin{aligned}\frac{\partial H(\theta^*)}{\partial p_j} &= -1 - \log\frac{p_j}{m_j} - \frac{\partial}{\partial p_j}\lambda\left(\sum_{i=1}^{K} p_i - 1\right) \\ &= -1 - \log\frac{p_j}{m_j} - \lambda,\end{aligned}$$

with the derivative with respect to the multiplier λ leading directly to the condition that the sum of the probabilities must be equal to 1. Setting the preceding differential to 0 produces

$$p_j = m_j \exp\left[-(1+\lambda)\right]. \tag{7.59}$$

Summing now over the K states, and assuming that the discretized "reference" distribution is proper, gives $\exp\left[-(1+\lambda)\right] = 1$, so that $\lambda = -1$. It follows that $p_j = m_j$. Further, the distribution with probabilities m_j is the one representing complete randomness of the experiment, since the probability "quanta" are allocated completely at random, with $m_j = 1/n$ for all j when the options are equally likely. Now, the definition of the derivative implies that

$$\lim_{\Delta\theta\to 0} p_j = \lim_{\Delta\theta\to 0} \frac{F(\theta_j) - F(\theta_{j-1})}{\theta_j - \theta_{j-1}} = F'(\theta_j) = p(\theta).$$

Hence, in the limit, it follows that the density of the maximum entropy distribution $p(\theta)$ is the uniform density distribution $m(\theta)$, since all the options are of equal size and infinitesimally small. Before presenting the variational argument, two examples are given, using the discretization procedure given above. Note that in the continuous case the maximum relative entropy distribution is not invariant with respect to the reference distribution chosen.

Example 7.20 *Maximum entropy prior distribution when the mean is known*

Suppose the mean of the prior distribution is given. The setting then is as in (7.54). However, the objective function to be optimized now involves entropy relative to the reference uniform distribution. The objective function takes the form:

$$\begin{aligned}&H^*(p_1, p_2, \ldots, p_K, \lambda_0, \lambda_1) \\ &= -\sum_{i=1}^{K}\left(p_i \log\frac{p_i}{m_i}\right) + \lambda_0\left(\sum_{i=1}^{K} p_i - 1\right) + \lambda_1\left(\sum_{i=1}^{K}\theta_i p_i - \overline{\theta}\right).\end{aligned}$$

Setting the derivatives to 0 leads to

$$p_i = m_i \exp(\lambda_1 \theta_i) \exp(\lambda_0 - 1).$$

The continuous counterpart gives, as density of the maximum relative entropy distribution,

$$p(\theta) = m(\theta) \exp(\lambda_1 \theta) \exp(\lambda_0 - 1).$$

If $m(\theta)$ (the reference density) is uniform, it follows that

$$p(\theta) \propto \exp(-\lambda_1^* \theta),$$

where $\lambda_1^* = -\lambda_1$. If θ is a strictly positive parameter, it follows that the maximum entropy distribution relative to a uniform measure is exponential. Since the mean of an exponential distribution is $\overline{\theta} = 1/\lambda_1^*$, the λ_1^* parameter can be assessed readily once the prior mean is specified. ■

Example 7.21 *Maximum entropy prior distribution when the mean, variance, and higher-order moments are given*

Knowledge of the mean and variance of the prior distribution imposes three constraints in the optimization procedure. The first constraint ensures propriety of the distribution, and the other two give the conditions stating knowledge of the first and second moments. Using the discretization procedure, the objective function to be maximized is

$$H^*(p_1, p_2, ..., p_K, \lambda_0, \lambda_1, \lambda_2) = -\sum_{i=1}^{K} \left(p_i \log \frac{p_i}{m_i} \right) + \lambda_0 \left(\sum_{i=1}^{K} p_i - 1 \right) + \lambda_1 \left(\sum_{i=1}^{K} \theta_i p_i - \overline{\theta} \right) + \lambda_2 \left(\sum_{i=1}^{K} \theta_i^2 p_i - \overline{\theta^2} \right),$$

where $\overline{\theta^2} = \sum_{i=1}^{K} \theta_i^2 p_i$ and, therefore, $\sigma^2 = \overline{\theta^2} - \overline{\theta}^2$ is the variance of the prior distribution. Differentiation with respect to p_i, and setting to 0, gives

$$\log \frac{p_i}{m_i} = 1 - \lambda_0 - \lambda_1 \theta_i - \lambda_2 \theta_i^2.$$

The continuous counterpart, after solving for the maximum relative entropy density $p(\theta)$, is

$$p(\theta) \propto m(\theta) \exp\left(1 - \lambda_0 - \lambda_1 \theta - \lambda_2 \theta^2\right). \tag{7.60}$$

Note that by restating the Lagrangian condition

$$\lambda_0 \left(\sum_{i=1}^{K} p_i - 1 \right)$$

as

$$\lambda_0 \left(1 - \sum_{i=1}^{K} p_i\right),$$

and so on, one can put, without loss of generality,

$$p(\theta) \propto m(\theta) \exp\left(1 + \lambda_0 + \lambda_1\theta + \lambda_2\theta^2\right). \tag{7.61}$$

If the reference density $m(\theta)$ is taken to be uniform (and proper, so that the maximum relative entropy density is guaranteed to be proper as well) one obtains

$$p(\theta) \propto \exp\left(\lambda_1\theta + \lambda_2\theta^2\right),$$

and the maximum entropy density is then

$$\begin{aligned} p(\theta) &= \frac{\exp\left(\lambda_1\theta + \lambda_2\theta^2\right)}{\int \exp\left(\lambda_1\theta + \lambda_2\theta^2\right) d\theta} \\ &= \frac{\exp\left(1 + \lambda_0 + \lambda_1\theta + \lambda_2\theta^2\right)}{\int \exp\left(1 + \lambda_0 + \lambda_1\theta + \lambda_2\theta^2\right) d\theta}. \end{aligned} \tag{7.62}$$

If the first M moments of the prior distribution are specified, the preceding generalizes to

$$p(\theta) = \frac{\exp\left(1 + \sum_{i=0}^{M} \lambda_i\theta^i\right)}{\int \exp\left(1 + \sum_{i=0}^{M} \lambda_i\theta^i\right) d\theta}. \tag{7.63}$$

■

Example 7.22 *The special case of known mean and variance*
Return now to a setting of known mean and variance of the prior distribution and consider the Kullback–Leibler discrepancy in (7.41). Let $g(\theta)$ denote any other density with mean $\overline{\theta}$ and variance σ^2 and let $p(\theta)$ be the maximum relative entropy distribution sought. Since the discrepancy is at least null

$$D[g(\theta), p(\theta)] = \int \left[\log \frac{p(\theta)}{g(\theta)}\right] p(\theta)\, d\theta \geq 0,$$

which implies that

$$H[p(\theta)] \leq -\int \{\log[g(\theta)]\}\, p(\theta)\, d\theta, \tag{7.64}$$

with the equality holding if and only if $p(\theta) = g(\theta)$ (Applebaum, 1996). Now take $g(\theta)$ to be the density of the normal distribution $N\left(\overline{\theta}, \sigma^2\right)$. Then

$$-\int \{\log[g(\theta)]\} p(\theta)\, d\theta = -\int \left\{ \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(\theta - \overline{\theta}\right)^2}{2\sigma^2} \right] \right) \right\} p(\theta)\, d\theta$$
$$= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \frac{1}{2\sigma^2} \int \left(\theta - \overline{\theta}\right)^2 p(\theta)\, d\theta.$$

Note that the integral in the preceding expression is σ^2, the variance of the prior distribution, by definition. Then, using (7.64),

$$H[p(\theta)] \leq -\int \{\log[g(\theta)]\} p(\theta)\, d\theta = \frac{1}{2}\left[1 + \log\left(2\pi\sigma^2\right)\right].$$

It follows that the entropy of the maximum entropy distribution cannot exceed $\left[1 + \log\left(2\pi\sigma^2\right)\right]/2$. Now, from Example 7.7, this quantity is precisely the entropy of a normal distribution. Hence, the maximum entropy prior distribution when the mean and variance are given must be a normal distribution with mean $\overline{\theta}$ and variance σ^2. ■

Continuous Case via Variational Arguments

As noted, the search for a maximum relative entropy distribution involves finding the function $p(\theta)$ that minimizes the integral

$$Int[p(\theta)] = \int \left[\log \frac{p(\theta)}{m(\theta)} \right] p(\theta)\, d\theta. \tag{7.65}$$

This is called a "variational" problem and its solution requires employing the advanced techniques of the calculus of variations (e.g., Weinstock, 1974; Fox, 1987). Since few biologists are exposed to the basic ideas (referred to as "standard" in some statistical texts) we provide a cursory introduction. Subsequently, a specific result is applied to the problem of finding a continuous maximum entropy distribution. We follow Weinstock (1974) closely.

In general, consider the integral

$$Int[y(x)] = \int_{x_1}^{x_2} f[x, y(x), y'(x)]\, dx, \tag{7.66}$$

where $y(x)$ is a twice differentiable function and $y'(x)$ is its first derivative with respect to x. We seek the function $y(x)$ rendering the above integral minimum (or maximum). The problem resides in finding such a function or, equivalently, in arriving at conditions that the function must obey. First, the function must satisfy the boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$, with x_1, x_2, y_1, and y_2 given. Now, when going from the point (x_1, y_1)

to the point (x_2, y_2), one can draw a number of curves, each relating the dependent variable to x, and with each curve defining a specific integration path. Next, define the "family" of functions

$$Y(x) = y(x) + \varepsilon\eta(x), \tag{7.67}$$

where ε is a parameter of the family and $\eta(x)$ is differentiable with $\eta(x_1) = \eta(x_2) = 0$. This ensures that $Y(x_1) = y(x_1) = y_1$ and $Y(x_2) = y(x_2) = y_2$, so that all members of the family possess the required ends. Also, note that the form of (7.67) indicates that no matter what $\eta(x)$ is chosen, the minimizing function $y(x)$ will be a member of the family for $\varepsilon = 0$. Now, return to (7.66), and replace y and y' by Y and Y', respectively. Then, write

$$Int(\varepsilon) = \int_{x_1}^{x_2} f[x, Y(x), Y'(x)]\, dx, \tag{7.68}$$

where, for a given $\eta(x)$, the integral is a function of ε. Observe from (7.67) that

$$Y'(x) = y'(x) + \varepsilon\eta'(x),$$

so, in conjunction with (7.67), it becomes clear that setting $\varepsilon = 0$ is equivalent to replacing Y and Y' by y and y', respectively. Hence (7.68) is minimum with respect to ε at the value $\varepsilon = 0$. The problem of minimizing (7.68) then reduces to one of standard calculus with respect to ε, except that we know in advance that the minimizing value is $\varepsilon = 0$! Hence, it must be true that

$$\left.\frac{dInt(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = 0.$$

Now, since the limits of integration do not involve ε:

$$\frac{dInt(\varepsilon)}{d\varepsilon} = \int_{x_1}^{x_2} \frac{d}{d\varepsilon} f[x, Y(x), Y'(x)]\, dx. \tag{7.69}$$

Employing standard results for the derivative of a function of several variables (Kaplan, 1993):

$$\begin{aligned}\frac{d}{d\varepsilon} f[x, Y(x), Y'(x)] &= \frac{\partial f}{\partial Y}\frac{\partial Y}{\partial \varepsilon} + \frac{\partial f}{\partial Y'}\frac{\partial Y'}{\partial \varepsilon} \\ &= \frac{\partial f}{\partial Y}\eta(x) + \frac{\partial f}{\partial Y'}\eta'(x),\end{aligned}$$

the integral in (7.69) becomes

$$\frac{dInt(\varepsilon)}{d\varepsilon} = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial Y}\eta(x) + \frac{\partial f}{\partial Y'}\eta'(x)\right] dx.$$

Further, since setting $\varepsilon = 0$ is equivalent to replacing (Y, Y') by (y, y'), use of the preceding in (7.69) gives

$$\left.\frac{d\,Int\,(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = \int_{x_1}^{x_2} \frac{\partial f}{\partial y}\eta\,(x)\,dx + \int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\eta'\,(x)\,dx = 0. \tag{7.70}$$

Integrating the second term of (7.70) by parts gives

$$\left.\frac{d\,Int\,(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$$
$$= \int_{x_1}^{x_2} \frac{\partial f}{\partial y}\eta\,(x)\,dx + \left.\frac{\partial f}{\partial y'}\eta\,(x)\right|_{x_1}^{x_2} - \int_{x_1}^{x_2} \left[\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right]\eta\,(x)\,dx = 0.$$

Now recall that at the boundary points, $\eta\,(x_1) = \eta\,(x_2) = 0$, so the second term in the preceding three-term expression vanishes. Rearranging,

$$\left.\frac{d\,Int\,(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right]\eta\,(x)\,dx = 0. \tag{7.71}$$

This condition must hold true for all $\eta\,(x)$, in view of the requirements imposed above. Hence, a condition that the function must obey in order to extremize the integral (7.66) is that

$$\frac{\partial f\,[x, y\,(x), y'\,(x)]}{\partial y} - \left\{\frac{d}{dx}\frac{\partial f\,[x, y\,(x), y'\,(x)]}{\partial y'}\right\} = 0. \tag{7.72}$$

This differential equation is called the Eulerian or Euler–Lagrange condition (Weinstock, 1974; Fox, 1987). Solving the equation for $y\,(x)$ provides the function that minimizes the integral, provided that a minimum exists.

Return now to the problem of finding the prior density $p\,(\theta)$ that minimizes the integral (7.65) subject to the constraints imposed by the information available. As seen before, the constraints may result from specifying moments of the prior distribution or some features thereof; for example, one of the constraints is that the prior must be proper. Suppose there are M constraints having the form

$$E\,[q_i\,(\theta)] = m_i, \quad i = 0, 1, \ldots, M-1,$$

where $q\,(\cdot)$ denotes some function of θ. For instance, if one specifies the first and second moments of the prior distribution, the constraints would be

$$\begin{aligned} \int p\,(\theta)\,d\theta &= 1 = m_0, \\ \int \theta p\,(\theta)\,d\theta &= E\,(\theta) = m_1, \\ \int \theta^2 p\,(\theta)\,d\theta &= E\left(\theta^2\right) = m_2. \end{aligned}$$

In general, we seek to minimize the objective function

$$I\left[p\left(\theta\right)\right]=\int\left[\log\frac{p\left(\theta\right)}{m\left(\theta\right)}\right]p\left(\theta\right)d\theta+\sum_{i=0}^{M-1}\lambda_{i}\left[\int q_{i}\left(\theta\right)p\left(\theta\right)d\theta-m_{i}\right], \quad (7.73)$$

with respect to $p\left(\cdot\right)$; as usual, the λ's are Lagrange multipliers. We now apply the variational argument leading to the derivative (7.70) and, for simplicity, set $m\left(\theta\right)=1$; that is, the reference density is taken to be uniform. Then, write:

$$\begin{aligned} p_{\varepsilon}\left(\theta\right) &= p\left(\theta\right)+\varepsilon\eta\left(\theta\right), \\ p_{\varepsilon}'\left(\theta\right) &= \frac{d}{d\varepsilon}\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]=\eta\left(\theta\right). \end{aligned}$$

Using this in (7.73), with $m\left(\theta\right)=1$ gives:

$$I\left[p_{\varepsilon}\left(\theta\right)\right]=\int\left\{\log\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]\right\}\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]d\theta+$$
$$\sum_{i=0}^{M-1}\lambda_{i}\left[\int q_{i}\left(\theta\right)\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]d\theta-m_{i}\right].$$

Hence:

$$\frac{dI\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]}{d\varepsilon}=\int\left\{\log\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]\right\}\eta\left(\theta\right)d\theta+\int\eta\left(\theta\right)d\theta$$
$$+\sum_{i=0}^{M-1}\lambda_{i}\int q_{i}\left(\theta\right)\eta\left(\theta\right)d\theta.$$

Rearranging and setting the derivative to 0 as required by the variational argument, yields:

$$\left.\frac{dI\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]}{d\varepsilon}\right|_{\varepsilon=0}$$
$$=\int\left\{\log\left[p\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]+\sum_{i=0}^{M-1}\lambda_{i}q_{i}\left(\theta\right)\right\}\eta\left(\theta\right)d\theta\Bigg|_{\varepsilon=0}=0.$$

Since this must be true for all integration paths, the extremizing density satisfies the equation:

$$\log\left[p\left(\theta\right)\right]+\sum_{i=0}^{M-1}\lambda_{i}q_{i}\left(\theta\right)=0.$$

Solving for the maximum entropy density gives

$$p\left(\theta\right)=\exp\left[-\sum_{i=0}^{M-1}\lambda_{i}q_{i}\left(\theta\right)\right].$$

After normalization, and noting that the sign of the Lagrange multipliers is unimportant (one can write $\lambda_i^* = -\lambda_i$), this becomes

$$p(\theta) = \frac{\exp\left[\sum_{i=0}^{M-1} \lambda_i^* q_i(\theta)\right]}{\int \exp\left[\sum_{i=0}^{M-1} \lambda_i^* q_i(\theta)\right] d\theta}. \tag{7.74}$$

This is precisely in the form of (7.63), with the latter being a particular case of (7.74) when the $q_i(\theta)$ functions are the moments of the prior distribution.

It is important to note that in all cases the maximum entropy distribution, must be defined relative to a reference distribution $m(\theta)$; otherwise, the concept of entropy does not carry to the continuous case. Here, a uniform reference distribution has been adopted arbitrarily. This illustrates that the concept of maximum entropy (in the continuous case) does not help to answer completely the question of how a non-informative prior should be constructed (Bernardo, 1979; Bernardo and Smith, 1994). On the one hand, information is introduced (and perhaps legitimately so) via the constraints, these stating features of the prior that are known. On the other hand, the reference measure must be a representation of ignorance itself, so the problem relays to one of how to construct a truly noninformative (in some sense) reference distribution.

7.5.4 Reference Prior Distributions

Reference analysis (Bernardo, 1979) can be viewed as a way of rendering inferences as "objective" as possible, in the sense that the prior should have a minimal effect, relative to the data, on posterior inference. As discussed below, the notion of "minimal effect" is defined in a precise manner. Statisticians often use the term reference priors rather than reference analysis. The theory is technically involved, and the area is still the subject of much research. Hence, only a sketch of the main ideas is presented following Bernardo and Smith (1994) closely.

Single Parameter Model

From (7.29), the amount of information about a parameter that an experiment is expected to provide, can be written as

$$\begin{aligned} I[e, h(\theta)] &= \int \left\{ \int \log\left[\frac{h(\theta|\mathbf{y})}{h(\theta)}\right] h(\theta|\mathbf{y})\, d\theta \right\} \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &= E\left\{ \int \log\left[\frac{h(\theta|\mathbf{y})}{h(\theta)}\right] h(\theta|\mathbf{y})\, d\theta \right\}, \end{aligned}$$

where the expectation is taken with respect to the marginal (in the Bayesian sense) distribution of the observations, but conditionally on the experimen-

tal design adopted, as usual. The notation $I[e, h(\theta)]$ makes explicit that this measure of information is actually a functional of the prior density $h(\theta)$; e denotes a specific experiment. Note that the information measure is the expectation of the Kullback–Leibler distance between the posterior and prior distributions taken over all data that can result from this experiment, given a certain probability model. Hence, this expectation must be nonnegative. Also, it can be verified readily that $I[e, h(\theta)]$ is invariant under one-to-one transformations.

Now suppose that the experiment is replicated K times, yielding the hypothetical data vector

$$\mathbf{y}_K^* = \left[\begin{array}{cccccc} \mathbf{y}_1' & \mathbf{y}_2' & . & . & . & \mathbf{y}_K' \end{array} \right]'.$$

The sampling model would then be

$$p(\mathbf{y}_K^* | \theta) = \prod_{i=1}^{K} p(\mathbf{y}_i | \theta).$$

The expected information from such an experiment is

$$I[e(K), h(\theta)] = \int \left\{ \int \log \left[\frac{h(\theta | \mathbf{y}_K^*)}{h(\theta)} \right] h(\theta | \mathbf{y}_K^*) \, d\theta \right\} \mathbf{g}(\mathbf{y}_{\mathbf{K}}^*) d\mathbf{y}_K^*. \quad (7.75)$$

If the experiment could be replicated an infinite number of times, one would be in a situation of perfect or complete information about the parameter. Hence, the quantity

$$I[e(\infty), h(\theta)] = \lim_{K \to \infty} I[e(K), h(\theta)]$$

measures, in some sense, the missing information about θ expressed as a function of the prior density $h(\theta)$. As more information is contained in the prior, less is expected to be gained from exhaustive data. On the other hand, if the prior contains little information, more would be expected to be gained from valuable experimentation. A "noninformative" prior would then be that maximizing the missing information.

The reference prior, denoted as $\pi(\theta)$, is defined formally to be the prior that maximizes the missing information functional given above. If $\pi_K(\theta)$ denotes the prior density that maximizes $I[e(K), h(\theta)]$ for a certain amount of replication K, then $\pi(\theta)$ is the limiting value as $K \to \infty$ of the sequence of priors $\{\pi_K(\theta), \ K = 1, 2, \ldots\}$ that ensues as replication increases. Associated with each $\pi_K(\theta)$ there is the corresponding posterior

$$\pi_K(\theta | \mathbf{y}) \propto p(\mathbf{y} | \theta) \pi_K(\theta), \quad (7.76)$$

whose limit, as $K \to \infty$, is defined as the reference posterior distribution

$$\pi(\theta | \mathbf{y}) = \lim_{K \to \infty} \pi_K(\theta | \mathbf{y}).$$

The reference prior is defined as any positive function $\pi(\theta)$, such that

$$\pi(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\,\pi(\theta).$$

The definition implies that the reference posterior distribution depends only on the asymptotic behavior of the model since the amount of replication is allowed to go to infinity.

The technical details of the procedure for obtaining the reference prior, that culminates in expressions (7.82) and (7.83), are given below. Consider (7.75) and rewrite it as

$$I[e(K), h(\theta)] = \int \left\{ \int \log \left[\frac{h(\theta|\mathbf{y}_K^*)}{h(\theta)} \right] p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^* \right\} h(\theta)\, d\theta. \quad (7.77)$$

Since $p(\mathbf{y}_{\mathbf{K}}^*|\boldsymbol{\theta})$ integrates to 1, the inner integral can be rearranged as

$$\int [\log h(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^* - \log h(\theta).$$

Thus

$$I[e(K), h(\theta)] = \int \log \left[\frac{\exp\left\{ \int [\log h(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^* \right\}}{h(\theta)} \right] h(\theta)\, d\theta.$$

Define

$$f_K(\theta) = \exp\left\{ \int [\log h(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^* \right\}, \quad (7.78)$$

and note that $f_K(\theta)$ depends implicitly on $h(\theta)$ through the posterior density $h(\theta|\mathbf{y}_K^*)$. Then one has

$$I[e(K), h(\theta)] = \int \log \left[\frac{f_K(\theta)}{h(\theta)} \right] h(\theta)\, d\theta.$$

Imposing the constraint that the prior density integrates to 1, the prior $\pi_K(\theta)$ that maximizes $I[e(K), h(\theta)]$ must be an extremal of the functional

$$F\{h(\theta)\} = \int \log \left[\frac{f_K(\theta)}{h(\theta)} \right] h(\theta)\, d\theta + \lambda \left[\int h(\theta)\, d\theta - 1 \right], \quad (7.79)$$

where λ is a Lagrange multiplier.

Define

$$h_\varepsilon(\theta) = h(\theta) + \varepsilon\eta(\theta),$$

so

$$h'_\varepsilon(\theta) = \frac{d}{d\varepsilon}[h(\theta) + \varepsilon\eta(\theta)] = \eta(\theta).$$

Now we make use of the variational argument employed in connection with maximum entropy priors and consider the function

$$F\left\{h_{\varepsilon}\left(\theta\right)\right\}=\int\log\left[\frac{f_{\varepsilon K}\left(\theta\right)}{h_{\varepsilon}\left(\theta\right)}\right]h_{\varepsilon}\left(\theta\right)d\theta+\lambda\left[\int h_{\varepsilon}\left(\theta\right)d\theta-1\right].$$

Hence

$$\left.\frac{dF\left[h\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]}{d\varepsilon}\right|_{\varepsilon=0}$$
$$=\int\left\{\frac{d}{d\varepsilon}\log\left[\frac{f_{\varepsilon K}\left(\theta\right)}{h_{\varepsilon}\left(\theta\right)}\right]h_{\varepsilon}\left(\theta\right)\right\}d\theta+\lambda\int\left\{\frac{d}{d\varepsilon}h_{\varepsilon}\left(\theta\right)\right\}d\theta=0.$$

Further,

$$\left.\frac{dF\left[h\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]}{d\varepsilon}\right|_{\varepsilon=0}=\int h_{\varepsilon}'\left(\theta\right)\log\left[\frac{f_{\varepsilon K}\left(\theta\right)}{h_{\varepsilon}\left(\theta\right)}\right]d\theta$$
$$+\int h_{\varepsilon}\left(\theta\right)\frac{h_{\varepsilon}\left(\theta\right)}{f_{\varepsilon K}\left(\theta\right)}\left[\frac{f_{\varepsilon K}'\left(\theta\right)h_{\varepsilon}\left(\theta\right)-f_{\varepsilon K}\left(\theta\right)h_{\varepsilon}'\left(\theta\right)}{h_{\varepsilon}^{2}\left(\theta\right)}\right]d\theta$$
$$+\lambda\int h_{\varepsilon}'\left(\theta\right)d\theta$$
$$=\int\eta\left(\theta\right)\log\left[\frac{f_{\varepsilon K}\left(\theta\right)}{h_{\varepsilon}\left(\theta\right)}\right]d\theta+\int\frac{f_{\varepsilon K}'\left(\theta\right)h_{\varepsilon}\left(\theta\right)}{f_{\varepsilon K}\left(\theta\right)}d\theta$$
$$-\int\eta\left(\theta\right)d\theta+\lambda\int\eta\left(\theta\right)d\theta=0.$$

Hence, as in Bernardo and Smith (1994)

$$\left.\frac{dF\left[h\left(\theta\right)+\varepsilon\eta\left(\theta\right)\right]}{d\varepsilon}\right|_{\varepsilon=0}$$
$$=\int\{\left[\log f_{K}\left(\theta\right)\right]\eta(\theta)+\frac{h\left(\theta\right)}{f_{K}\left(\theta\right)}f_{K}'\left(\theta\right)$$
$$-\left[\log h\left(\theta\right)+1\right]\eta(\theta)+\lambda\eta\left(\theta\right)\}d\theta=0, \tag{7.80}$$

where $f_{K}'\left(\theta\right)$ is the derivative of $f_{\varepsilon K}\left(\theta\right)$ with respect to ε evaluated at $\varepsilon=0$.

Given the form of $f_{K}\left(\theta\right)$ given in (7.78),

$$f_{K}'\left(\theta\right)=\left.\frac{d}{d\varepsilon}f_{\varepsilon K}\left(\theta\right)\right|_{\varepsilon=0}$$
$$=\left.\frac{d}{d\varepsilon}\exp\left\{\int\left[\log h_{\varepsilon}\left(\theta|\mathbf{y}_{K}^{*}\right)\right]p\left(\mathbf{y}_{K}^{*}|\theta\right)d\mathbf{y}_{K}^{*}\right\}\right|_{\varepsilon=0}.$$

Recalling that the posterior is proportional to the product of the prior and of the sampling model, the preceding can be expressed as

$$f'_K(\theta) = \frac{d}{d\varepsilon} \exp\left\{ \int \left[\log \frac{p(\mathbf{y}^*_K|\theta)[h(\theta)+\varepsilon\eta(\theta)]}{\int p(\mathbf{y}^*_K|\theta)[h(\theta)+\varepsilon\eta(\theta)]d\theta} \right] p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}.$$

Carrying out the differentiation

$$f'_K(\theta) = f_K(\theta) \left\{ \frac{d}{d\varepsilon} \int \log\left(p(\mathbf{y}^*_K|\theta)\left[h(\theta)+\varepsilon\eta(\theta)\right]\right) p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}$$
$$- f_K(\theta) \left\{ \frac{d}{d\varepsilon} \int \left[\log\left(\int p(\mathbf{y}^*_K|\theta)\left[h(\theta)+\varepsilon\eta(\theta)\right] d\theta \right) \right] p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}.$$

Proceeding with the algebra,

$$f'_K(\theta) = f_K(\theta) \left\{ \int \frac{d}{d\varepsilon} \left(\log\left[h(\theta)+\varepsilon\eta(\theta)\right]\right) p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0} - f_K(\theta)$$
$$\times \left\{ \int \frac{d}{d\varepsilon} \left[\log\left(\int p(\mathbf{y}^*_K|\theta)\left[h(\theta)+\varepsilon\eta(\theta)\right] d\theta \right) \right] p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}.$$

Further,

$$f'_K(\theta) = f_K(\theta) \left\{ \frac{\eta(\theta)}{h(\theta)+\varepsilon\eta(\theta)} \int p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}$$
$$- f_K(\theta) \left\{ \int \frac{\int p(\mathbf{y}^*_K|\theta)\eta(\theta)d\theta}{\int p(\mathbf{y}^*_K|\theta)[h(\theta)+\varepsilon\eta(\theta)]d\theta} p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}\Bigg|_{\varepsilon=0}.$$

Evaluating appropriate terms at $\varepsilon = 0$, noting that $\int p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K = 1$, and that $\int p(\mathbf{y}^*_K|\theta)\, h(\theta)\, d\mathbf{y}^*_K = p(\mathbf{y}^*_{\mathbf{K}})$, gives

$$f'_K(\theta) = f_K(\theta)\frac{\eta(\theta)}{h(\theta)} - f_K(\theta) \left\{ \int \frac{\int p(\mathbf{y}^*_{\mathbf{K}}|\boldsymbol{\theta})\eta(\theta)d\theta}{p(\mathbf{y}^*_{\mathbf{K}})} p(\mathbf{y}^*_K|\theta)\, d\mathbf{y}^*_K \right\}. \quad (7.81)$$

Note now that $h_\varepsilon(\theta) = h(\theta) + \varepsilon\eta(\theta)$ implies that

$$\int h_\varepsilon(\theta)\, d\theta = \int h(\theta)\, d\theta + \varepsilon \int \eta(\theta)\, d\theta.$$

Hence, $\int \eta(\theta)\, d\theta = 0$ is a necessary condition for $h(\cdot)$ being a p.d.f.. This follows because, then,

$$\int h_\varepsilon(\theta)\, d\theta = \int h(\theta)\, d\theta = 1.$$

Now, if $h(\cdot)$ is a proper density function, the posterior is also a proper density. Therefore,

$$\begin{aligned} 1 &= \int \frac{p(\mathbf{y}_K^*|\theta)\, h_\varepsilon(\theta)\, d\theta}{p(\mathbf{y}_\mathbf{K}^*)} \\ &= \int \frac{p(\mathbf{y}_K^*|\theta)\, h(\theta)\, d\theta}{p(\mathbf{y}_\mathbf{K}^*)} + \varepsilon \int \frac{p(\mathbf{y}_K^*|\theta)\, \eta(\theta)\, d\theta}{p(\mathbf{y}_\mathbf{K}^*)} \\ &= 1 + \varepsilon \int \frac{p(\mathbf{y}_K^*|\theta)\, \eta(\theta)\, d\theta}{p(\mathbf{y}_\mathbf{K}^*)} = 1. \end{aligned}$$

Hence $\int p(\mathbf{y}_K^*|\theta)\, \eta(\theta)\, d\theta = 0$. Using this condition in (7.81), we get

$$f_K'(\theta) = f_K(\theta) \frac{\eta(\theta)}{h(\theta)}.$$

Employing this in (7.80), the condition that the extremal must satisfy is

$$\begin{aligned} &\int \{\log[f_K(\theta)] + 1 - \log h(\theta) - 1 + \lambda\}\, \eta(\theta)\, d\theta \\ = &\int \{\log[f_K(\theta)] - \log h(\theta) + \lambda\}\, \eta(\theta) d\theta = 0. \end{aligned}$$

Since this must hold for all $\eta(\theta)$, the extremizing density can be found by solving

$$\log[f_K(\theta)] - \log h(\theta) + \lambda = 0.$$

Upon retaining only the terms that vary with θ:

$$\begin{aligned} h(\theta) &\propto \exp\{\log[f_K(\theta)] + \lambda\} \\ &\propto f_K(\theta). \end{aligned} \tag{7.82}$$

Finally, in view of the form of $f_K(\theta)$ given in (7.78),

$$h(\theta) \propto \exp\left\{\int [\log h(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^*\right\}. \tag{7.83}$$

This prior $h(\theta)$, which maximized $I[e(K), h(\theta)]$ for each K, was denoted before $\pi_K(\theta)$. Expression (7.83) gives an implicit solution because $f_K(\theta)$ depends on the prior through the posterior density $h(\theta|\mathbf{y}_K^*)$. However, as $K \to \infty$ the posterior density $h(\theta|\mathbf{y}_K^*)$ will approach an asymptotic form with density $h^*(\theta|\mathbf{y}_K^*)$ say, that does not depend on the prior at all. Such an approximation is given, for example, by the asymptotic process (7.6), either with $\mathbf{I}(\theta)$ or $\widehat{\mathbf{H}}$ as precision matrix. Let the asymptotic posterior have density $h^*(\theta|\mathbf{y}_K^*)$. Then, the sequence of positive functions

$$h_K^*(\theta) \propto \exp\left\{\int [\log h^*(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^*\right\}, \quad K = 1, 2, \ldots, \tag{7.84}$$

derived from such an asymptotic posterior, will induce a sequence of posterior distributions

$$\pi_K(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\, h_K^*(\theta), \quad K = 1, 2, \ldots,$$

as in (7.76). Note that $p(\mathbf{y}|\theta)$ is the density of the actual observations resulting from the experiment. Then, as in (7.76), the reference posterior distribution of θ, $\pi(\theta|\mathbf{y})$, is defined as the limiting distribution resulting from this K-fold replicated "conceptual experiment"

$$\pi(\theta|\mathbf{y}) = \lim_{K\to\infty} \pi_K(\theta|\mathbf{y}), \tag{7.85}$$

where the limit is understood in the information-entropy sense

$$\lim_{K\to\infty} \int \log\left[\frac{\pi_K(\theta|\mathbf{y})}{\pi(\theta|\mathbf{y})}\right] \pi_K(\theta|\mathbf{y})\, d\theta = 0.$$

The reference prior is a function retrieving the reference posterior $\pi(\theta|\mathbf{y})$ by formal use of Bayes theorem, i.e., a positive function $\pi(\theta)$, such that, for all $\mathbf{y}$,

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)\,\pi(\theta)}{\int p(\mathbf{y}|\theta)\,\pi(\theta)\, d\theta}.$$

The limiting posterior distribution (7.85) is the same as the one that could be obtained from the sequence of priors $\pi_K(\theta)$ maximizing the expected information $I[e(K), h(\theta)]$, as obtained from (7.76).

Although the construction guarantees that $\pi_K(\theta)$ is proper for each K, this is not so for $\pi(\theta)$, the limiting reference prior, as will be shown in some examples. Hence, only the reference posterior distribution is amenable to probabilistic interpretation. Also, observe that each of the terms in the sequence is the expected value of the logarithm of the density of the asymptotic approximation of the posterior taken with respect to the sampling distribution of the observations generated in the appropriate K-fold replicated experiment. This defines an algorithm for arriving at the form of the reference function $h_K^*(\theta)$, with this leading to the reference prior when $K \to \infty$. Reference priors are, thus, limiting forms.

Invariance under Transformation

The invariance under one-to-one transformations is shown in Bernardo (1979). Let $\xi = g(\theta)$ be such a transformation. The reference prior of ξ should have the form

$$\pi(\xi) \propto \pi\left[g^{-1}(\xi)\right] \left|\frac{dg^{-1}(\xi)}{d\xi}\right| \propto \pi\left[g^{-1}(\xi)\right] |J_\xi|,$$

where $|J_\xi|$ is the absolute value of the Jacobian of the transformation. The sequence of functions approaching the reference prior, using (7.84), is

$$\begin{aligned} h_K^*(\xi) &\propto \exp\left\{\int [\log h^*(\xi|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\xi)\, d\mathbf{y}_K^*\right\} \\ &\propto \exp\left\{\int [\log h^*(\theta|\mathbf{y}_K^*)\,|J_\xi|]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^*\right\} \\ &\propto \exp(\log|J_\xi|)\exp\left\{\int [\log h^*(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^*\right\} \\ &\propto \exp\left\{\int [\log h^*(\theta|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta)\, d\mathbf{y}_K^*\right\}|J_\xi| \\ &\propto h_K^*(\theta)\,|J_\xi|\,. \end{aligned}$$

Hence, the reference functions follow the usual rules for change of variables, that is, the reference function for ξ is proportional to the product of the reference function for θ times the absolute value of the Jacobian of the transformation.

Reference Prior under Consistency and Asymptotic Normality

It is now shown that the reference prior has an explicit form when a consistent estimator of θ exists, and when the asymptotic posterior distribution of θ, given the hypothetical data $\mathbf{y}_K^*$ from a K-fold replicated experiment, is normal. As mentioned at the beginning of this chapter, important regularity conditions must be satisfied to justify these asymptotic assumptions.

Consider an experiment based on n observations. As before, let $\mathbf{y}_K^*$ be a hypothetical data vector of order $kn \times 1$ resulting from a K-fold replicate of the said experiment. Also, let $\widehat{\theta}_{kn}$ be a sufficient statistic or estimator (a function of $\mathbf{y}_K^*$) such that, with complete certainty,

$$\lim_{K\to\infty} \widehat{\theta}_{kn} = \theta.$$

Since $\widehat{\theta}_{kn}$ is sufficient for the parameter, the posterior density of θ, given the data, is the same as the posterior density, given $\widehat{\theta}_{kn}$. This is so because the likelihood function can be written as the product of a function of the data only (which gets absorbed in the integration constant), times another part involving both θ and $\widehat{\theta}_{kn}$. Next, let the asymptotic approximation to the posterior density be

$$h^*(\theta|\mathbf{y}_K^*) = h^*\left(\theta|\widehat{\theta}_{kn}\right).$$

The counterpart of (7.84) can be written as

$$h_K^*(\theta) \propto \exp\left\{\int \left[\log h^*\left(\theta|\widehat{\theta}_{kn}\right)\right] p\left(\widehat{\theta}_{kn}|\theta\right) d\widehat{\theta}_{kn}\right\},$$

where $p\left(\widehat{\theta}_{kn}|\theta\right)$ is the density of the sampling distribution of the sufficient statistic or consistent estimator. Now evaluate the asymptotic approximation at $\widehat{\theta}_{kn} = \theta$, and denote this as:

$$h^*\left(\theta|\widehat{\theta}_{kn}\right)\Big|_{\widehat{\theta}_{kn}=\theta}.$$

This should be very "close" to $h^*\left(\theta|\widehat{\theta}_{kn}\right)$ (in the Kullback–Leibler sense). Hence, one can write

$$\begin{aligned} h_K^*(\theta) &\propto \exp\left\{\int\left[\log h^*\left(\theta|\widehat{\theta}_{kn}\right)\Big|_{\widehat{\theta}_{kn}=\theta}\right] p\left(\widehat{\theta}_{kn}|\theta\right) d\widehat{\theta}_{kn}\right\} \\ &\propto \exp\left\{\left[\log h^*\left(\theta|\widehat{\theta}_{kn}\right)\Big|_{\widehat{\theta}_{kn}=\theta}\right]\int p\left(\widehat{\theta}_{kn}|\theta\right) d\widehat{\theta}_{kn}\right\} \\ &\propto h^*\left(\theta|\widehat{\theta}_{kn}\right)\Big|_{\widehat{\theta}_{kn}=\theta}. \end{aligned} \tag{7.86}$$

This implies that $h_K^*(\theta)$ is proportional to the density of any asymptotic approximation to the posterior in which the consistent estimator $\widehat{\theta}_{kn}$ is replaced by the unknown parameter θ.

Suppose now that the asymptotic posterior distribution of θ is normal; as seen in Section 7.3, the assumption is valid under regularity conditions. Hence, for a K-fold replicated experiment, write

$$\theta|\widehat{\theta}_{kn} \sim N\left(\widehat{\theta}_{kn}, \left[knI_1\left(\widehat{\theta}_{kn}\right)\right]^{-1}\right),$$

where $I_1\left(\widehat{\theta}_{kn}\right)$ is Fisher's information measure for a single observation evaluated at $\widehat{\theta}_{kn}$. Then, by virtue of (7.86), the reference function must be

$$\begin{aligned} h_K^*(\theta) &\propto N\left(\theta, \left[knI_1\left(\widehat{\theta}_{kn}\right)\right]^{-1}\right)\Big|_{\widehat{\theta}_{kn}=\theta} \\ &\propto \frac{1}{\left[knI_1\left(\widehat{\theta}_{kn}\right)\right]^{-\frac{1}{2}}} \exp\left[-\frac{knI_1\left(\widehat{\theta}_{kn}\right)}{2}\left(\theta-\widehat{\theta}_{kn}\right)^2\right]\Bigg|_{\widehat{\theta}_{kn}=\theta} \\ &\propto \sqrt{I_1(\theta)}. \end{aligned} \tag{7.87}$$

The limit of this sequence, which is the reference prior $\pi(\theta)$, is also $\sqrt{I_1(\theta)}$, the square root of Fisher's information measure for a single observation. Hence,

$$\pi(\theta) \propto \sqrt{I_1(\theta)}.$$

It follows that in the special case of a single continuous parameter and when the posterior distribution is asymptotically normal, the reference prior algorithm yields Jeffreys' invariance prior (see Section 7.5.2).

Example 7.23 *Reference prior for the mean of a normal distribution*
The experiment consists of n samples from a normal distribution with unknown mean μ and known variance σ^2. Fisher's information measure for a single observation is

$$E_{y|\mu,\sigma^2}\left[-\frac{d^2}{(d\mu)^2}\left\{\log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]\right\}\right]=\frac{1}{\sigma^2}.$$

Hence, the reference prior $\pi(\mu)$ is a constant, since the information measure does not involve the parameter of interest. Note that the reference prior is improper. Although, by construction, the reference functions obtained by maximizing the information measure $I[e(K), h(\theta)]$ are proper for each K, as shown in (7.79), the reference prior obtained by taking limits may be improper. ■

Example 7.24 *Reference prior for the probability of success in the binomial distribution*
The experiment consists of n Bernoulli trials with success probability θ. The distribution of a single observation in a Bernoulli trial is

$$p(y|\theta)=\theta^y(1-\theta)^{1-y},$$

where y takes the value 1 with probability θ or the value 0 with probability $1-\theta$. The information measure from a single draw is

$$E_{y|\theta}\left\{-\frac{d^2}{(d\theta)^2}\left[y\log\theta+(1-y)\log(1-\theta)\right]\right\}$$
$$=E_{y|\theta}\left[\frac{y}{\theta^2}+\frac{(1-y)}{(1-\theta)^2}\right]=\frac{1}{\theta(1-\theta)}.$$

The reference prior is then

$$\pi(\theta)\propto\sqrt{\frac{1}{\theta(1-\theta)}},$$

which is a $Be\left(\frac{1}{2},\frac{1}{2}\right)$ distribution. The reference prior is proper in this case, and the reference posterior distribution is

$$\begin{aligned}\pi(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}\\ &\propto \theta^{y+\frac{1}{2}-1}(1-\theta)^{n-y+\frac{1}{2}-1},\end{aligned}$$

where y is the number of successes observed out of n Bernoulli trials. The posterior is then a $Be\left(y+\frac{1}{2}, n-y+\frac{1}{2}\right)$ distribution, which is always proper (even if all trials are all successes or all failures). If all observations

are failures, the mean of the posterior distribution is equal to $1/(2n+2)$. For example, if $n = 1000$ flies are screened in search of a specific mutant and all are found to be normal, the posterior mean estimate of the mutation rate is 1/2002. The Bayesian estimate admits the possibility that the population is liable to at least some mutation. On the other hand, the ML estimator of the mutation rate would be 0 in this case. ■

Presence of a Single Nuisance Parameter

Suppose that the model has two unknown parameters, so $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. Parameter θ_1 is of primary inferential interest and θ_2 acts as a nuisance parameter. This will be called an ordered parameterization, denoted as (θ_1, θ_2). The problem is to develop a reference prior for θ_1. As usual, the joint prior density can be written as

$$h(\boldsymbol{\theta}) = h(\theta_1)\, h(\theta_2|\theta_1),$$

where $h(\theta_2|\theta_1)$ is the density of the conditional distribution of the nuisance parameter, given θ_1. Correspondingly, the density of the θ_1-reference prior distribution is expressible as

$$\pi_{\theta_1}(\theta_1, \theta_2) = \pi_{\theta_1}(\theta_1)\, \pi_{\theta_1}(\theta_2|\theta_1).$$

It is important to note that the reference prior may depend on the order of the parameterization. That is, the θ_1-reference prior distribution will be different, in general, from the θ_2-reference prior distribution with density

$$\pi_{\theta_2}(\theta_2, \theta_1) = \pi_{\theta_2}(\theta_2)\, \pi_{\theta_2}(\theta_1|\theta_2).$$

This is perplexing at first sight, but it can be explained on the grounds that the information-theoretic measures that are maximized involve logarithmic divergences between distributions, and these depend on the specific distributions intervening.

Note that if the conditional reference prior density $\pi(\theta_2|\theta_1)$ were known, it could be used to integrate the nuisance parameter θ_2 out of the likelihood to obtain the one-parameter model

$$p(\mathbf{y}|\theta_1) = \int p(\mathbf{y}|\theta_1, \theta_2)\, \pi(\theta_2|\theta_1)\, d\theta_2,$$

from which $\pi(\theta_1)$ can be deduced as before. Then, the reference posterior for θ_1 would be

$$p(\theta_1|\mathbf{y}) \propto p(\mathbf{y}|\theta_1)\, \pi(\theta_1).$$

We now describe the algorithm for obtaining the marginal reference posterior density of θ_1. There are essentially two steps. First, the conditional reference prior for the nuisance parameter, $\pi(\theta_2|\theta_1)$ (dropping the subscripts indexing π from now on), can be arrived at by applying the algorithm presented before for the single parameter situation. This is because,

given θ_1, $p(\mathbf{y}|\theta_1, \theta_2)$ only depends on the nuisance parameter θ_2. The reference function is now

$$h_K^*(\theta_2|\theta_1) \propto \exp\left\{\int [\log h^*(\theta_2|\theta_1, \mathbf{y}_K^*)]\, p(\mathbf{y}_\mathbf{K}^*|\boldsymbol{\theta}) d\mathbf{y}_K^*\right\}, \quad K = 1, 2, \ldots,$$

where $h^*(\theta_2|\theta_1, \mathbf{y}_K^*)$ is the density of any asymptotic approximation to the conditional posterior distribution of θ_2, given θ_1, that does not depend on the prior. The conditional reference prior, $\pi(\theta_2|\theta_1)$ is obtained as the limit, as $K \to \infty$, of the sequence of reference functions, $h_K^*(\theta_2|\theta_1)$, as before. The first step of the algorithm is completed, using $\pi(\theta_2|\theta_1)$ to integrate out the nuisance parameter, thus obtaining the integrated likelihood $p(\mathbf{y}_K^*|\theta_1)$:

$$p(\mathbf{y}_K^*|\theta_1) = \int p(\mathbf{y}_K^*|\theta_1, \theta_2)\, \pi(\theta_2|\theta_1)\, d\theta_2.$$

This step assumes that $\pi(\theta_2|\theta_1)$ is a proper density. Recall, however, that reference analysis can lead to improper reference priors. If this is the case, the algorithm will not work. We return to this point briefly at the end of this section.

In the second step, the algorithm is again applied using as reference function

$$h_K^{**}(\theta_1) \propto \exp\left\{\int [\log h^{**}(\theta_1|\mathbf{y}_K^*)]\, p(\mathbf{y}_K^*|\theta_1)\, d\mathbf{y}_K^*\right\}, \quad K = 1, 2, \ldots, \tag{7.88}$$

where $h^{**}(\theta_1|\mathbf{y}_K^*)$ is any asymptotic approximation to the posterior distribution of the parameter of interest, but constructed using the integrated likelihood $p(\mathbf{y}_K^*|\theta_1)$. The marginal reference prior $\pi(\theta_1)$ is obtained as the limit, as $K \to \infty$, of the sequence of reference functions $h_K^{**}(\theta_1)$. Finally, the reference posterior of θ_1 is obtained as

$$p(\theta_1|\mathbf{y}) \propto p(\mathbf{y}|\theta_1)\, \pi(\theta_1),$$

which completes the algorithm.

We shall now consider the regular case where joint posterior asymptotic normality can be established. It is also assumed that the conditional reference prior is proper. The asymptotic approximation to the joint posterior distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ is written as

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} |\widehat{\boldsymbol{\theta}}_{kn} \sim N\left(\begin{bmatrix} \widehat{\theta}_{1kn} \\ \widehat{\theta}_{2kn} \end{bmatrix}, \begin{bmatrix} I_{\theta_1\theta_1} & I_{\theta_1\theta_2} \\ I_{\theta_1\theta_2} & I_{\theta_2\theta_2} \end{bmatrix}^{-1}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right).$$

Recall that $I_{\theta_2\theta_2}$ is the information about θ_2 in a model, either without θ_1 or when assuming that this parameter is known. Also, let

$$\begin{bmatrix} I_{\theta_1\theta_1} & I_{\theta_1\theta_2} \\ I_{\theta_1\theta_2} & I_{\theta_2\theta_2} \end{bmatrix}^{-1} = \begin{bmatrix} I^{\theta_1\theta_1} & I^{\theta_1\theta_2} \\ I^{\theta_1\theta_2} & I^{\theta_2\theta_2} \end{bmatrix},$$

and note than an asymptotic approximation to the marginal posterior distribution of θ_1 is given by

$$\theta_1|\widehat{\boldsymbol{\theta}}_{kn} \sim N\left(\widehat{\theta}_{1kn},\, I^{\theta_1\theta_1}\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right). \tag{7.89}$$

With independent samples, the information matrix is nk times the information from a single observation, so $I^{\theta_1\theta_1}$ is equal to $I_1^{\theta_1\theta_1}/nk$, where $I_1^{\theta_1\theta_1}$ is the appropriate part of the inverse of Fisher's information matrix for a single observation. Further, $I_1^{\theta_1\theta_1}$ is typically a function of both θ_1 and θ_2; hence, it is instructive to write $I_1^{\theta_1\theta_1}$ as $I_1^{\theta_1\theta_1}(\theta_1,\theta_2)$.

Employing the argument in (7.87), the reference function for arriving at the conditional reference prior of the nuisance parameter is

$$h_K^*(\theta_2|\theta_1) \propto \sqrt{I_{1(\theta_2\theta_2)}}. \tag{7.90}$$

Hence, the density of the conditional reference prior for the nuisance parameter is

$$\pi(\theta_2|\theta_1) \propto \sqrt{I_{1(\theta_2\theta_2)}}.$$

Since we assume that this is proper, the integrated likelihood is

$$\begin{aligned} p(\mathbf{y}_K^*|\theta_1) &= \int p(\mathbf{y}_K^*|\theta_1,\theta_2)\,\pi(\theta_2|\theta_1)\,d\theta_2 \\ &\propto \int p(\mathbf{y}_K^*|\theta_1,\theta_2)\,\sqrt{I_{1(\theta_2\theta_2)}}\,d\theta_2. \end{aligned}$$

Consequently, the reference function needed to arrive at the reference prior for the parameter of interest θ_1 in (7.88) has the form

$$\begin{aligned} &h_K^{**}(\theta_1) \\ &\propto \exp\left\{\int \left[\log h^{**}(\theta_1|\mathbf{y}_K^*)\right]\left[\int p(\mathbf{y}_K^*|\theta_1,\theta_2)\,\sqrt{I_{1(\theta_2\theta_2)}}d\theta_2\right]d\mathbf{y}_K^*\right\}. \end{aligned} \tag{7.91}$$

Rearranging the integral expression

$$\begin{aligned} &h_K^{**}(\theta_1) \\ &\propto \exp\left\{\int \sqrt{I_{1(\theta_2\theta_2)}}\left\{\int \left[\log h^{**}(\theta_1|\mathbf{y}_K^*)\right]p(\mathbf{y}_K^*|\theta_1,\theta_2)\,d\mathbf{y}_K^*\right\}d\theta_2\right\} \\ &\propto \exp\left\{\int \sqrt{I_{1(\theta_2\theta_2)}}E_{\mathbf{y}_K^*|\theta_1,\theta_2}\left[\log h^{**}(\theta_1|\mathbf{y}_K^*)\right]d\theta_2\right\}, \end{aligned} \tag{7.92}$$

which is a function of θ_1 only since θ_2 is integrated out. Now, using the asymptotic approximation (7.89) to the marginal posterior of θ_1, and taking expectations over the distribution of the observations arising in the K-fold replicated experiment, for large K (recall that $\widehat{\theta}_{1kn}$ is asymptotically

unbiased)

$$
\begin{aligned}
&E_{\mathbf{y}_K^*|\theta_1,\theta_2}\left[\log h^{**}\left(\theta_1|\mathbf{y}_K^*\right)\right] \\
&= E_{\mathbf{y}_K^*|\theta_1,\theta_2}\left[\log \frac{1}{\sqrt{2\pi\left[\frac{I_1^{\theta_1\theta_1}(\theta_1,\theta_2)}{nk}\right]}} - \frac{nk\left(\theta_1-\widehat{\theta}_{1kn}\right)^2}{2\left[I_1^{\theta_1\theta_1}\left(\theta_1,\theta_2\right)\right]}\right] \\
&= -\log\sqrt{2\pi\left[\frac{I_1^{\theta_1\theta_1}\left(\theta_1,\theta_2\right)}{nk}\right]} - \frac{1}{2} \propto \log\left[I_1^{\theta_1\theta_1}\left(\theta_1,\theta_2\right)\right]^{-\frac{1}{2}}.
\end{aligned}
$$

Using this in (7.92):

$$
h_K^{**}\left(\theta_1\right) \propto \exp\left\{\int \sqrt{I_{1(\theta_2\theta_2)}}\log\left[I_1^{\theta_1\theta_1}\left(\theta_1,\theta_2\right)\right]^{-\frac{1}{2}} d\theta_2\right\}, \tag{7.93}
$$

and recall that $\pi\left(\theta_1\right) \propto h_K^{**}\left(\theta_1\right)$ for $K \to \infty$. Since (7.93) involves information measures (or their inverses) for a single observation, this gives the reference prior directly. Hence, if the conditional reference prior $\pi\left(\theta_2|\theta_1\right)$ is proper, the reference prior of the primary parameter is obtained as follows:
(1) Compute Fisher's information measure about the nuisance parameter from a single observation (acting as if the primary parameter were known or absent from the model).
(2) Compute Fisher's information matrix for a single observation for the full, two-parameter model, and invert it (this may depend on both θ_1 and θ_2).
Then proceed to evaluate the integral and the exponential function in (7.93).

Suppose now that $\sqrt{I_{1(\theta_2\theta_2)}}$ factorizes as

$$
\sqrt{I_{1(\theta_2\theta_2)}} \propto f_{\theta_2}\left(\theta_1\right) g_{\theta_2}\left(\theta_2\right).
$$

This implies that the conditional reference prior of the nuisance parameter is

$$
\pi\left(\theta_2|\theta_1\right) = a g_{\theta_2}\left(\theta_2\right), \tag{7.94}
$$

where $a^{-1} = \int g_{\theta_2}\left(\theta_2\right) d\theta_2$. Also suppose that the inverse of Fisher's information measure (the θ_1 part) factorizes as

$$
\left[I_1^{\theta_1\theta_1}\left(\theta_1,\theta_2\right)\right]^{-\frac{1}{2}} \propto f_{\theta_1}\left(\theta_1\right) g_{\theta_1}\left(\theta_2\right).
$$

Then, if the space of the nuisance parameter does not depend on θ_1, application of these conditions in (7.93) gives

$$h_K^{**}(\theta_1) \propto \exp\left\{\int a g_{\theta_2}(\theta_2) \log\left[f_{\theta_1}(\theta_1) g_{\theta_1}(\theta_2)\right] d\theta_2\right\}$$
$$\propto \exp\left\{\log\left[f_{\theta_1}(\theta_1)\right] \int a g_{\theta_2}(\theta_2)\, d\theta_2 + \int \log\left[g_{\theta_1}(\theta_2)\right] a g_{\theta_2}(\theta_2)\, d\theta_2\right\}$$
$$\propto \exp\left\{\log\left[f_{\theta_1}(\theta_1)\right]\right\} \propto f_{\theta_1}(\theta_1), \tag{7.95}$$

since $\int a g_{\theta_2}(\theta_2)\, d\theta_2 = 1$. Combining (7.94) and (7.95) yields as θ_1-reference prior,

$$\pi_{\theta_1}(\theta_1, \theta_2) \propto f_{\theta_1}(\theta_1) g_{\theta_2}(\theta_2). \tag{7.96}$$

This result holds irrespective of whether the conditional reference prior is proper or not.

If the conditional reference prior of the nuisance parameter is not proper, the technical arguments are more involved. Bernardo and Smith (1994) indicate that the entire parameter space of θ_2, say Θ_2, must be broken into increasing sequences Θ_{2i}, possibly dependent on θ_1. For each sequence, one obtains a normalized conditional reference prior such that

$$\pi_i(\theta_2|\theta_1) = \frac{\pi(\theta_2|\theta_1)}{\int\limits_{\Theta_{2i}} \pi(\theta_2|\theta_1)\, d\theta_2}, \quad i = 1, 2, \ldots.$$

An integrated likelihood is obtained for each i, and a marginal reference prior $\pi_i(\theta_1)$ is identified using the procedures described above, to arrive at the joint prior $\{\pi_i(\theta_1)\pi_i(\theta_2|\theta_1)\}$. The limit of this sequence yields the desired reference prior. The strategy requires identifying suitable "cuts" of the parameter space.

Example 7.25 *Reference prior for the mean of a normal distribution: unknown standard deviation*

Let n samples be drawn from a normal distribution with mean and standard deviation μ and σ, respectively, with both parameters unknown. First we shall derive the reference posterior distribution for the ordered parameterization (μ, σ) in which σ acts as a nuisance parameter. Second, the reference analysis is carried out for the ordered parameterization (σ, μ).

Fisher's information measure for a single observation is formed from the expected negative second derivatives

$$E_{y|\mu,\sigma}\left\{-\frac{\partial^2}{(\partial\mu)^2} \log\left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]\right\}\right\} = \frac{1}{\sigma^2},$$

$$E_{y|\mu,\sigma}\left\{-\frac{\partial^2}{\partial\mu\partial\sigma} \log\left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]\right\}\right\} = 0,$$

$$E_{y|\mu,\sigma}\left\{-\frac{\partial^2}{(\partial\sigma)^2}\log\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]\right\}\right\}=\frac{2}{\sigma^2}.$$

Thus, Fisher's information matrix for $n=1$, and arranging it consistently with the ordered parameterization (μ,σ), is

$$\begin{bmatrix} I_{1(\mu\mu)} & I_{1(\mu\sigma)} \\ I_{1(\mu\sigma)} & I_{1(\sigma\sigma)} \end{bmatrix}=\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}.$$

The inverse is

$$\begin{bmatrix} I_1^{\mu\mu} & I_1^{\mu\sigma} \\ I_1^{\mu\sigma} & I_1^{\sigma\sigma} \end{bmatrix}=\begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}.$$

The conditional reference prior of σ, given μ, using (7.90), is

$$\pi(\sigma|\mu)\propto\sqrt{I_{1(\sigma\sigma)}}\propto\sqrt{\frac{2}{\sigma^2}}\propto\frac{1}{\sigma}.$$

Since the conditional reference prior is not proper, one encounters the technical difficulty mentioned earlier. However, we proceed to check whether the conditions leading to (7.96) hold. Following Bernardo (2001), note that

$$\sqrt{I_{1(\sigma\sigma)}}=\sqrt{2}\sigma^{-1}$$

factorizes as

$$\sqrt{I_{1(\sigma\sigma)}}=f_\sigma(\mu)\,g_\sigma(\sigma),$$

where $f_\sigma(\mu)=\sqrt{2}$ and $g_\sigma(\sigma)=\sigma^{-1}$. Also, $[I_1^{\mu\mu}]^{-\frac{1}{2}}=\sigma^{-1}$ factorizes as

$$[I_1^{\mu\mu}]^{-\frac{1}{2}}\propto f_\mu(\mu)\,g_\mu(\sigma),$$

where $f_\mu(\mu)=1$ and $g_\mu(\sigma)=\sigma^{-1}$. Thus (7.96) leads to the $\mu-$reference prior

$$\pi_\mu(\mu,\sigma)=\pi(\sigma|\mu)\,\pi(\mu)\propto f_\mu(\mu)\,g_\sigma(\sigma)=1\times\sigma^{-1}=\sigma^{-1}.$$

Hence, the reference prior for the mean is the improper uniform prior and the joint μ-reference prior is the reciprocal of the standard deviation. The reference posterior distribution of μ has density

$$\begin{aligned}\pi_\mu(\mu|\mathbf{y}) &\propto \int p(\mathbf{y}|\mu,\sigma)\,\pi_\mu(\mu,\sigma)\,d\sigma \\ &\propto \int(\sigma)^{-n-1}\exp\left\{-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i-\mu)^2\right\}d\sigma.\end{aligned}$$

Carrying out the integration, as seen earlier in the book, leads to a univariate-t process with mean $\overline{y}$, scale parameter

$$\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1},$$

and $n-1$ degrees of freedom as reference posterior distribution. ■

Example 7.26 *Reference prior for the standard deviation of a normal distribution: unknown mean*

The setting is as in Example 7.25 but we now consider the situation where μ acts as nuisance parameter. Here, $\sqrt{I_{1(\mu\mu)}} = 1/\sigma$ factorizes as $f_\mu(\mu) = 1$, $g_\mu(\sigma) = \sigma^{-1}$, and $[I_1^{\sigma\sigma}]^{-\frac{1}{2}} = \sqrt{2}/\sigma$ factorizes as $f_\sigma(\mu) = \sqrt{2}$ times $g_\sigma(\sigma) = \sigma^{-1}$. This leads to the σ-reference prior

$$\pi_\sigma(\mu,\sigma) \propto f_\mu(\mu)\, g_\sigma(\sigma) = 1 \times \sigma^{-1} = \sigma^{-1},$$

which is identical to the joint μ-reference prior in this case. The reference posterior density is then

$$\pi_\sigma(\sigma|\mathbf{y}) \propto \int (\sigma)^{-n-1} \exp\left\{-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i-\mu)^2\right\} d\mu.$$

This integration leads to a scaled inverted chi-square density with parameters $(n-1)\sum_{i=1}^{n}(y_i-\overline{y})^2$ and $n-1$. Equivalently, the reference posterior distribution for inferring σ is the process

$$Ga\left(\frac{n-1}{2}, \frac{(n-1)\sum_{i=1}^{n}(y_i-\overline{y})^2}{2}\right).$$

■

Example 7.27 *Reference prior for a standardized mean*

Following Bernardo and Smith (1994) and Bernardo (2001), consider inferring $\phi = \mu/\sigma$, a standardized mean or reciprocal of the coefficient of variation of a normal distribution, with the standard deviation σ acting as a nuisance parameter. The sampling model is as in the two preceding examples, but parameterized in terms of ϕ. In order to form Fisher's information measure for a single observation, the required derivatives are

$$E_{y|\phi,\sigma}\left\{-\frac{\partial^2}{(\partial\phi)^2}\log\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\phi\sigma)^2\right]\right\}\right\} = 1,$$

$$E_{y|\phi,\sigma}\left\{-\frac{\partial^2}{\partial\phi\partial\sigma}\log\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\phi\sigma)^2\right]\right\}\right\} = \frac{\phi}{\sigma},$$

$$E_{y|\phi,\sigma}\left\{-\frac{\partial^2}{(\partial\sigma)^2}\log\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\phi\sigma)^2\right]\right\}\right\}=\frac{2+\phi^2}{\sigma^2}.$$

The information matrix is, thus,

$$\begin{bmatrix} I_{1(\phi\phi)} & I_{1(\phi\sigma)} \\ I_{1(\phi\sigma)} & I_{1(\sigma\sigma)} \end{bmatrix}=\begin{bmatrix} 1 & \frac{\phi}{\sigma} \\ \frac{\phi}{\sigma} & \frac{2+\phi^2}{\sigma^2} \end{bmatrix},$$

with inverse

$$\begin{bmatrix} I_1^{\phi\phi} & I_1^{\phi\sigma} \\ I_1^{\phi\sigma} & I_1^{\sigma\sigma} \end{bmatrix}=\begin{bmatrix} 1+\frac{\phi^2}{2} & \frac{-\phi\sigma}{2} \\ \frac{-\phi\sigma}{2} & \frac{\sigma^2}{2} \end{bmatrix}.$$

If σ is the nuisance parameter, then

$$\sqrt{I_{1(\sigma\sigma)}}\propto\sqrt{2+\phi^2}\sigma^{-1},$$

factorizes into $f_\sigma(\phi)=\sqrt{2+\phi^2}$ and $g_\sigma(\sigma)=\sigma^{-1}$. Further

$$\left(I_1^{\phi\phi}\right)^{-\frac{1}{2}}\propto\left(1+\frac{\phi^2}{2}\right)^{-\frac{1}{2}}$$

factorizes into $f_\phi(\phi)=\left(1+\frac{\phi^2}{2}\right)^{-\frac{1}{2}}$ and $g_\phi(\sigma)=1$. The $\phi-$reference prior is then given by

$$\pi_\phi(\phi,\sigma)\propto f_\phi(\phi)\,g_\sigma(\sigma)\propto\left(1+\frac{\phi^2}{2}\right)^{-\frac{1}{2}}\sigma^{-1}.$$

■

Multiparameter Models

In a multiparameter situation, the reference posterior distribution is relative to an ordered parameterization of inferential interest. For example, if the ordered parameter vector is:

$$\boldsymbol{\theta}=\begin{pmatrix}\theta_1 & \theta_2 & . & . & . & \theta_m\end{pmatrix}',$$

so that θ_m is of primary inferential interest, the corresponding reference prior can be represented as

$$\pi_{\theta_m}(\boldsymbol{\theta})=\pi_{\theta_m}(\theta_m|\theta_1,\theta_2,...,\theta_{m-1})\,\pi_{\theta_m}(\theta_{m-1}|\theta_1,\theta_2,...,\theta_{m-2})\times\cdots$$
$$\cdots\times\pi_{\theta_m}(\theta_2|\theta_1)\,\pi_{\theta_m}(\theta_1).$$

An algorithm for deriving the reference prior under asymptotic normality (an extension of the procedure described for the two parameter situation) is presented in Berger and Bernardo (1992) and in Bernardo and Smith (1994). The method is quite involved algebraically and is not presented here.

As a final comment, the reader should be aware that there is no consensus among Bayesians on this topic. For example, Lindley, in the discussion of Bernardo (1979), remarks that "... the distributions derived by this procedure violate the likelihood principle, and therefore violate requirements of coherence ...." Lindley is referring to the foundational inconsistency of the method which is based on integrating over the sample space of the data. Further, McCulloch concludes his discussion of Berger and Bernardo (1992) with the following words: "We all use 'noninformative' priors and this work is probably the most important current work in the area. I found the papers very, um ... er ... ah ..., interesting. If the authors obtain impossible solutions it is because they are working on an impossible problem."

This page intentionally left blank

8

Bayesian Assessment of Hypotheses and Models

8.1 Introduction

The three preceding chapters gave an overview of how Bayesian probability models are constructed. Once a prior distribution is elicited and the form of the likelihood function is agreed upon, Bayesian analysis is conceptually straightforward: nuisance parameters are eliminated via integration and parameters of interest are inferred from their marginal posterior distributions. Further, yet-to-be-observed random variables can be predicted from the corresponding predictive distributions. In all cases, probability is the sole measure of uncertainty at each and everyone of the stages of Bayesian learning.

Inferences are expected to be satisfactory if the entire probability model (the prior, the likelihood, and all assumptions made) is a "good one", in some sense. In practice, however, agreement about the model to be used is more the exception than the rule, unless there is some well-established theory or mechanism underlying the problem. For example, a researcher may be uncertain about which hypothesis or theory holds. Further, almost always, there are alternative choices about the distributional form to be adopted or about the explanatory variables that should enter into a regression equation, say. Hence, it is important to take into account uncertainties about the model-building process. This is perfectly feasible in Bayesian analysis and new concepts do not need to be introduced in this respect. If there is a set of competing models in a certain class, each of the models in the set can be viewed as a different state of a random variable.

The prior distribution of this variable (the model) is updated using the information contained in the data, to arrive at the posterior distribution of the possible states of the model. Then inferences are drawn, either from the most probable model, a posteriori, or from the entire posterior distribution of the models, in a technique called Bayesian model averaging.

In this chapter, several concepts and techniques for the Bayesian evaluation of hypotheses and models are presented. Some of the approaches described are well founded theoretically; others are of a more exploratory nature. The next section defines the posterior probability of a model and an intimately related concept: the Bayes factor. Subsequently, the issue of "testing hypotheses" is presented from a Bayesian perspective. Approximations to the Bayes factor and extensions to the concept are suggested. A third section presents some methods for calculating the Bayes factor, including Monte Carlo procedures, since it is seldom the case that one can arrive at the desired quantities by analytical methods. The fourth and fifth sections present techniques for evaluating goodness of fit and the predictive ability of a model. The final section provides an introduction to Bayesian model averaging, with emphasis on highlighting its theoretical appeal from the point of view of predicting future observations.

8.2 Bayes Factors

8.2.1 Definition

Suppose there are several competing theories, hypotheses, or models about some aspect of a biological system. For example, consider different theories explaining how a population evolves. These theories are mutually exclusive and exhaustive (at least temporarily). The investigator assigns prior probability $p(H_i)$, $(i = 1, 2, ..., K)$ to hypothesis, or theory i, with $\sum_i p(H_i) = 1$. There is no limit to K and nesting requirements are not involved. After observing data $\mathbf{y}$, the posterior probability of hypothesis i is

$$p(H_i|\mathbf{y}) = \frac{p(H_i)\, p(\mathbf{y}|H_i)}{\sum_{i=1}^{K} p(H_i)\, p(\mathbf{y}|H_i)}, \quad i = 1, 2, \ldots, K, \tag{8.1}$$

where $p(\mathbf{y}|H_i)$ is the probability of the data under hypothesis i. If all hypotheses are equally likely a priori, which is the maximum entropy or reference prior in the discrete case (Bernardo, 1979), then

$$p(H_i|\mathbf{y}) = \frac{p(\mathbf{y}|H_i)}{\sum_{i=1}^{K} p(\mathbf{y}|H_i)}.$$

The posterior odds ratio of hypothesis i relative to hypothesis j takes the form

$$\frac{p(H_i|\mathbf{y})}{p(H_j|\mathbf{y})} = \frac{p(H_i)}{p(H_j)} \frac{p(\mathbf{y}|H_i)}{p(\mathbf{y}|H_j)}. \tag{8.2}$$

It follows that the posterior odds ratio is the product of the prior odds ratio and of the ratio between the marginal probabilities of observing the data under each of the hypotheses. The Bayes factor is defined to be

$$B_{ij} = \frac{p(\mathbf{y}|H_i)}{p(\mathbf{y}|H_j)} = \frac{\frac{p(H_i|\mathbf{y})}{p(H_j|\mathbf{y})}}{\frac{p(H_i)}{p(H_j)}} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}. \tag{8.3}$$

According to Kass and Raftery (1995) this terminology is apparently due to Good (1958). A $B_{ij} > 1$ means that H_i is more plausible than H_j in the light of $\mathbf{y}$. While the priors are not visible in the ratio $p(\mathbf{y}|H_i)/p(\mathbf{y}|H_j)$, because algebraically they cancel out, this does not mean that B_{ij} in general is not affected by prior specifications. This point is discussed below.

It is instructive to contrast this approach with the one employed in standard statistical analysis. In classical hypothesis testing, a null hypothesis $H_0 : \theta \in \theta_0$ and an alternative hypothesis $H_1 : \theta \in \theta_1$ are specified. The choice between these hypotheses is driven by the distribution under H_0 of a test statistic that is a function of the data (it could be the likelihood ratio), $T(\mathbf{y})$, and by the so-called p-value. This is defined as

$$\Pr[T(\mathbf{y}) \text{ at least as extreme as the value observed}|\theta, H_0]. \tag{8.4}$$

Then H_0 is accepted (or rejected, in which case H_1 is accepted) if the p-value is large (small) enough, or one may just quote the p-value and leave things there. Notice that (8.4) represents the probability of obtaining results larger than the one actually obtained; that is, (8.4) is concerned with events that might have occurred, but have not. Thus, the famous quotation from Jeffreys (1961):

> "What the use of p implies, therefore, is that a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred. ... On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it."

Often (and incorrectly), the p-value is interpreted as the probability that H_0 holds true. The interpretation in terms of probability of hypotheses, $p[H_0|T(\mathbf{y}) = t(\mathbf{y})]$, which is the Bayesian formulation of the problem, is conceptually more straightforward than the one associated with (8.4). Despite its conceptual clarity, the Bayesian approach is not free from problems. Perhaps not surprisingly, these arise especially in cases when prior information is supposed to convey vague knowledge.

8.2.2 Interpretation

The appeal of the Bayes factor as formulated in (8.3), is that it provides a measure of whether the data have increased or decreased the odds of H_i relative to H_j. This, however, does not mean that in general, the Bayes factor is driven by the data only. It is only when both H_i and H_j are simple hypotheses, that the prior influence vanishes and the Bayes factor takes the form of a likelihood ratio. In general, however, the Bayes factor depends on prior input, a point to which we will return.

Kass and Raftery (1995) give guidelines for interpreting the evidence against some "null hypothesis", H_0. For example, they suggest that a Bayes factor larger than 100 should be construed as "decisive evidence" against the null. Note that a Bayes factor under 1 means that there is evidence in support of H_0. When working in a logarithmic scale, $2 \log B_{ij}$, for example, the values are often easier to interpret by those who are familiar with likelihood ratio tests. It should be made clear from the onset that the Bayes factor cannot be viewed as a statistic having an asymptotic chi-square distribution under the null hypothesis. Again, B_{ij} is the quantity by which prior odds ratios are increased (or decreased) to become posterior odd ratios.

There are many differences between the Bayes factor and the usual likelihood ratio statistic. First, the intervening $p(\mathbf{y}|H_i)$ is not the classical likelihood, in general. Recall that the Bayesian marginal probability (or density) of the data is arrived at by integrating the joint density of the parameters and of the observations over all values that the parameters can take in their allowable space. For example, if hypothesis or model H_i has parameters $\boldsymbol{\theta}_i$, then for continuous data and continuous valued parameter vector

$$\begin{aligned} p(\mathbf{y}|H_i) &= \int p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\, p(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i \\ &= E_{\boldsymbol{\theta}_i|H_i}\left[p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\right]. \end{aligned} \tag{8.5}$$

The marginal density is, therefore, the expected value of all possible likelihoods, where the expectation is taken with respect to the prior distribution of the parameters. In likelihood inference, no such integration takes place unless the "parameters" are random variables having a frequentist interpretation. Since, in turn, these random variables have distributions indexed by parameters, the classical likelihood always depends on some fixed, unknown parameters. In the Bayesian approach, on the other hand, any dependence of the marginal distribution of the data is with respect to any hyperparameters the prior distribution may have, and with respect to the form of the model. In fact, $p(\mathbf{y}|H_i)$ is the prior predictive distribution and gives the density or probability of the data calculated before observation, unconditionally with respect to parameter values.

A second important difference is that the Bayes factor is not explicitly related to any critical value defining a rejection region of a certain size. For example, the usual p-values in classical hypothesis testing cannot be interpreted as the probabilities that either the null or the alternative hypotheses are "true". The p-value arises from the distribution of the test statistic (under the null hypothesis) in conceptual replications of the experiment. In contrast, the Bayes factor and the prior odds contribute directly to forming the posterior probabilities of the hypotheses. In order to illustrate, suppose that two models are equally probable, a priori. Then a Bayes factor $B_{01} = 19$, would indicate that the null hypothesis or model is 19 times more probable than its alternative, and that the posterior probability that the null model is true is 0.95. On the other hand, in a likelihood ratio test, a value of the test statistic generating a p-value of 0.95 as defined by (8.4) cannot be construed as evidence that the null hypothesis has a 95% chance of being true.

8.2.3 The Bayes Factor and Hypothesis Testing

Decision-Theoretic View

In Bayesian analysis, "hypothesis testing" is viewed primarily as a decision problem (e.g., Zellner, 1971). Suppose there are two hypotheses or models: H_0 (null) and H_1 (alternative). If one chooses H_0 when H_1 is "true", then a loss L_{10} is incurred. Similarly, when H_1 is adopted when the null holds, the loss is L_{01}. Otherwise, there are no losses.

The posterior expectation of the decision "accept the null hypothesis" is

$$\begin{aligned} E\left(\text{loss}|\text{accept } H_0, \mathbf{y}\right) &= 0 \times p\left(H_0|\mathbf{y}\right) + L_{10}\, p\left(H_1|\mathbf{y}\right) \\ &= L_{10}\, p\left(H_1|\mathbf{y}\right). \end{aligned}$$

Likewise, the expected posterior loss of the decision "accept the alternative" is

$$\begin{aligned} E\left(\text{loss}|\text{reject } H_0, \mathbf{y}\right) &= 0 \times p\left(H_1|\mathbf{y}\right) + L_{01}\, p\left(H_0|\mathbf{y}\right) \\ &= L_{01}\, p\left(H_0|\mathbf{y}\right). \end{aligned}$$

Naturally, if the expected posterior loss of accepting H_0 is larger than that of rejecting it, one would decide to reject the null hypothesis. Then the decision rule is

$$\text{if } E\left(\text{loss}|\text{reject } H_0, \mathbf{y}\right) < E\left(\text{loss}|\text{accept } H_0, \mathbf{y}\right) \rightarrow \text{reject } H_0.$$

The preceding is equivalent to

$$L_{01}\, p\left(H_0|\mathbf{y}\right) < L_{10}\, p\left(H_1|\mathbf{y}\right),$$

or, in terms of (8.3),

$$B_{10} = \frac{p\,(\mathbf{y}|H_1)}{p\,(\mathbf{y}|H_0)} > \frac{L_{01}\,p\,(H_0)}{L_{10}\,p\,(H_1)}. \tag{8.6}$$

This indicates that the null hypothesis is to be rejected if the Bayes factor (ratio of marginal likelihoods under the two hypotheses or models) for the alternative, relative to the null, exceeds the ratio of the prior expected losses. Note that $L_{01}\,p\,(H_0)$ is the expected prior loss of rejecting the null when this is true; $L_{10}\,p\,(H_1)$ is the expected prior loss that results when H_1 is true and one accepts H_0. Then the ratio of prior to posterior expected losses

$$\frac{L_{01}\,p\,(H_0)}{L_{10}\,p\,(H_1)}$$

plays the role of the "critical" value in classical hypothesis testing. If one views the Bayes factor as the "test statistic", the critical value is higher when one expects to loose more from rejecting the null than from accepting it. In other words, the larger the prior expected loss from rejecting the null (when this hypothesis is true) relative to the prior expected loss of accepting it (when H_1 holds), the larger the weight of the evidence should be in favor of the alternative, as measured by the Bayes factor.

If the losses are such that $L_{01} = L_{10}$, it follows from (8.6) that the decision rule is simply

$$B_{10} = \frac{p\,(\mathbf{y}|H_1)}{p\,(\mathbf{y}|H_0)} > \frac{p\,(H_0)}{p\,(H_1)}.$$

This implies that if the two models or hypotheses are equiprobable a priori, then the alternative should be chosen over the null whenever the Bayes factor exceeds 1. Similarly, a "critical value" of 10 should be adopted if it is believed a priori that the null hypothesis is 10 times more likely than the alternative. In all cases, it must be noted that the "accept" or "reject" framework depends nontrivially on the form of the loss function, and that adopting $L_{01} = L_{10}$ may not be realistic in many cases.

The definition in the form of (8.6) highlights the importance, in Bayesian testing, of defining non-zero a priori probabilities. This is so even though the Bayes factor can be calculated without specifying $p\,(H_0)$ and $p\,(H_1)$. If H_0 or H_1 are a priori impossible, the observations will not modify this information.

Bayesian Comparisons

Contrasting Two Simple Hypotheses

The definition of the Bayes factor in (8.3) as a ratio of marginal densities does not make explicit the influence of the prior distributions. With one

exception, Bayes factors are affected by prior specifications. The exception occurs when the comparison involves two simple hypotheses. In this case, under H_0, a particular value $\boldsymbol{\theta}_0$ is assigned to the parameter vector, whereas under H_1, another value $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ is posited. There is no uncertainty about the value of the parameter under any of the two competing hypotheses. Then one can express the discrete prior probability of hypothesis i as $p(H_i) = \Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_i)$, and the conditional p.d.f. for $\mathbf{y}$ given H_i as $p(\mathbf{y}|H_i) = p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_i)$. The Bayes factor for the alternative against the null is then

$$B_{10} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_1)}{p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_0)}. \tag{8.7}$$

In this particular situation, the Bayes factor is the odds for H_1 relative to H_0 given by the data only. Expression (8.7) is a ratio of standard likelihoods, where the values of the parameters are completely specified. In general, however, B_{10} depends on prior input. When a hypothesis is not simple, in order to arrive at the form equivalent to (8.7), one must compute the expectation of the likelihood of $\boldsymbol{\theta}_i$ with respect to the prior distribution. For continuously distributed values of the vector $\boldsymbol{\theta}_i$ and prior density $p(\boldsymbol{\theta}_i|H_i)$, one writes

$$p(\mathbf{y}|H_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\, p(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i.$$

In contrast to the classical likelihood ratio frequentist test, the Bayes factor does not impose nesting restrictions concerning the form of the likelihood functions, as illustrated in the following example adapted from Bernardo and Smith (1994).

Example 8.1 *Two fully specified models: Poisson versus negative binomial process*

Two completely specified models are proposed for counts. A sample of size n with values $y_1, y_2, \ldots, y_n$ is drawn independently from some population. Model P states that the distribution of the observations is Poisson with parameter θ_P. Then the likelihood under this model is

$$p(\mathbf{y}|\theta = \theta_P) = \prod_{i=1}^{n} \left[\frac{\theta_P^{y_i} \exp(-\theta_P)}{y_i!} \right] = \frac{\theta_P^{n\overline{y}}}{\exp(n\theta_P) \prod_{i=1}^{n} y_i!}. \tag{8.8}$$

Model N proposes a negative binomial distribution with parameter θ_P. The corresponding likelihood is

$$p(\mathbf{y}|\theta = \theta_N) = \prod_{i=1}^{n} \left[\theta_N (1 - \theta_N)^{y_i} \right] = \theta_N^n (1 - \theta_N)^{n\overline{y}}. \tag{8.9}$$

The Bayes factor for Model N relative to Model P is then

$$B_{NP} = \left(\frac{1-\theta_N}{\theta_P}\right)^{n\overline{y}} \frac{\theta_N^n}{\left[\exp\left(n\theta_P\right)\prod_{i=1}^{n} y_i!\right]^{-1}},$$

and its logarithm can be expressed as

$$\log B_{NP} = n\left[\overline{y}\log\left(\frac{1-\theta_N}{\theta_P}\right) + \log\left(\theta_N\right)\right] + n\theta_P + \sum_{i=1}^{n}\log\left(y_i!\right).$$

■

Simple Versus Composite Hypotheses

A second type of comparison is one where one of the models (Model 0 = M_0) postulates a given value of the parameter, whereas the other model (Model 1 = M_1) allows the unknown parameter to take freely any of its values in the allowable parameter space. This is called a simple versus composite test (Bernardo and Smith, 1994), and the Bayes factor in this case takes the form

$$B_{10} = \frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_0\right)} = \frac{\int p\left(\mathbf{y}|\boldsymbol{\theta},M_1\right)p\left(\boldsymbol{\theta}|M_1\right)d\boldsymbol{\theta}}{p\left(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0\right)},$$

where $p\left(\boldsymbol{\theta}|M_1\right)$ is the density of the prior distribution of the parameter vector under the assumptions of Model 1.

There is an interesting relationship between the posterior probability that $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ and the Bayes factor (Berger, 1985). Denote the prior probability of models 0 and 1 as $p\left(M_0\right)$ and $p\left(M_1\right)$, respectively, with $p\left(M_0\right)+p\left(M_1\right)=1$. The term $p\left(M_0\right)$ can also be interpreted as the prior probability that $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. Then, the posterior probability that $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ is

$$\Pr\left(\boldsymbol{\theta}=\boldsymbol{\theta}_0|\mathbf{y}\right) = \frac{p\left(M_0\right)p\left(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0\right)}{p\left(\mathbf{y}\right)}.$$

The constant term in the denominator is given by

$$\begin{aligned} p\left(\mathbf{y}\right) &= p\left(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0\right)\Pr\left(\boldsymbol{\theta}=\boldsymbol{\theta}_0|M_0\right)p\left(M_0\right) \\ &\quad + \int p\left(\mathbf{y}|\boldsymbol{\theta},M_1\right)p\left(\boldsymbol{\theta}|M_1\right)p\left(M_1\right)d\boldsymbol{\theta} \\ &= p\left(M_0\right)p\left(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0\right) + p\left(M_1\right)\int p\left(\mathbf{y}|\boldsymbol{\theta},M_1\right)p\left(\boldsymbol{\theta}|M_1\right)d\boldsymbol{\theta}, \end{aligned}$$

	Genotypes			
	AB/ab	Ab/ab	aB/ab	ab/ab
Phenotype	AB	Ab	aB	ab
Frequency	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$
Observed	a	b	c	d

TABLE 8.1. Genotypic distribution in offspring from a backcross design.

with the equality arising in the last line because $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_0) = 1$. Substituting above yields

$$\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|\mathbf{y}) = \left[1 + \frac{p(M_1)}{p(M_0)} B_{10}\right]^{-1}.$$

It is important to mention that in evaluating a point null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, say, $\boldsymbol{\theta}_0$ must be assigned a positive probability a priori. The point null hypothesis cannot be tested invoking a continuous prior distribution, since any such prior will give $\boldsymbol{\theta}_0$ prior (and therefore posterior) probability of zero.

In contrast to the traditional likelihood ratio, the test of a parameter value on the boundary of the parameter space using the Bayes factor does not in principle create difficulties. This being so because asymptotic distributions and series expansions do not come into play. Such a test is illustrated in the example below.

Example 8.2 *A point null hypothesis: assessing linkage between two loci* The problem consists of inferring the probability of recombination r between two autosomal loci A and B, each with two alleles. The parameter r is defined in the closed interval $\left[0, \frac{1}{2}\right]$, with the upper value corresponding to the situation where there is no linkage. We wish to derive the posterior distribution of the recombination fraction between the two loci, and to contrast the two models that follow. The null model (M_0) postulates that segregation is independent (that is, $r = \frac{1}{2}$), whereas the alternative model (M_1) claims that the loci are linked (that is, $r < \frac{1}{2}$).

Let the alleles at the corresponding loci be A, a, B, and b, where A and B are dominant alleles. Suppose that a line consisting of coupling heterozygote individuals AB/ab is crossed to homozygotes ab/ab. Hence four offspring classes can be observed: AB/ab, Ab/ab, aB/ab, and ab/ab. Let $n = a + b + c + d$ be the total number of offspring observed. The four possible genotypes resulting from this cross, their phenotypes, the expected frequencies and the observed numbers are shown in Table 8.1. The expected relative frequencies follow from the fact that if the probability of observing a recombinant is r, the individual can be either Ab/ab or aB/ab, with the two classes being equally likely. A similar reasoning applies to the observation of a non-recombinant type.

Suppose that the species under consideration has 22 pairs of autosomal chromosomes. In the absence of prior information about loci A and B, it may be reasonable to assume that the probability that these are located on the same chromosome (and therefore, linked, so that $r < \frac{1}{2}$) is $\frac{1}{22}$. This is so because the probability that 2 randomly picked alleles are in a given chromosome is $\left(\frac{1}{22}\right)^2$, and there are 22 chromosomes in which this can occur. Hence, a priori, $p(M_1) = \frac{1}{22}$ and $p(M_0) = \frac{21}{22}$; the two models are viewed as mutually exclusive and exhaustive.

Next, we must arrive at some reasonable prior distribution for r under M_1. Here, a development by Smith (1959) is followed. First, note that the recombination fraction takes the value $r = \frac{1}{2}$ with prior probability $\frac{21}{22}$. Further, assume a uniform distribution for r otherwise (provided one can view the values $0 < r < \frac{1}{2}$ as "equally likely"). Then the density of this uniform distribution, $p(r|M_1)$ must be such that

$$\Pr\left(r < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} p(u|M_1)\, du = \frac{1}{22}.$$

Solving for the desired uniform density gives

$$p(r|M_1) = \frac{1}{11}.$$

Therefore the prior is the uniform process $p(r|M_1) = \frac{1}{11}$ for $0 < r < \frac{1}{2}$, and the point mass $\frac{21}{22}$ at $r = \frac{1}{2}$. That is, $\Pr\left(r = \frac{1}{2}\right) = p(M_0) = \frac{21}{22}$. Note that given M_0, $\Pr\left(r = \frac{1}{2}|M_0\right) = 1$.

Given the data $\mathbf{y} = (a, b, c, d)'$ in Table 8.1, the conditional distribution of the observations under linkage has the multinomial form

$$\begin{aligned} p(\mathbf{y}|r) &= \frac{n!}{a!b!c!d!}\left[\frac{1}{2}(1-r)\right]^a\left(\frac{1}{2}r\right)^b\left(\frac{1}{2}r\right)^c\left[\frac{1}{2}(1-r)\right]^d \\ &\propto \left(\frac{1}{2}\right)^n (1-r)^{a+d}\, r^{b+c}. \end{aligned}$$

Under no linkage

$$p\left(\mathbf{y}|r = \frac{1}{2}\right) = \frac{n!}{a!b!c!d!}\left(\frac{1}{4}\right)^{a+b+c+d} \propto \left(\frac{1}{4}\right)^n.$$

Therefore the posterior odds ratio is given by

$$\begin{aligned} \frac{p(M_1|\mathbf{y})}{p(M_0|\mathbf{y})} &= \frac{p(M_1)}{p(M_0)} \frac{\int_0^{\frac{1}{2}} p(r|M_1)\, p(\mathbf{y}|r, M_1)\, dr}{p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)} \\ &= \frac{1}{21} \frac{\frac{1}{11}\left(\frac{1}{2}\right)^n \int_0^{\frac{1}{2}} r^{b+c}(1-r)^{a+d}\, dr}{\left(\frac{1}{4}\right)^n}, \end{aligned}$$

where

$$B_{10} = \frac{\int_0^{\frac{1}{2}} p\left(r|M_1\right) p\left(\mathbf{y}|r, M_1\right) dr}{p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)}.$$

The posterior probability of linkage is

$$\Pr\left(r < \frac{1}{2}|\mathbf{y}, M_1\right) = \int_0^{\frac{1}{2}} p\left(r|\mathbf{y}, M_1\right) dr,$$

where

$$p\left(r|\mathbf{y}, M_1\right) = \frac{p\left(r|M_1\right) p\left(\mathbf{y}|r, M_1\right)}{p\left(\mathbf{y}\right)}, \tag{8.10}$$

whereas the posterior probability of no linkage is

$$\Pr\left(r = \frac{1}{2}|\mathbf{y}, M_0\right) = 1 - \Pr\left(r < \frac{1}{2}|\mathbf{y}, M_1\right).$$

The denominator in (8.10) is equal to

$$p\left(\mathbf{y}\right) = p\left(M_0\right) p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)$$
$$+ p\left(M_1\right) \int_0^{\frac{1}{2}} p\left(r|M_1\right) p\left(\mathbf{y}|r, M_1\right) dr.$$

The integrals in these expressions can easily be evaluated numerically. ■

Example 8.3 *Lindley's paradox*

This problem was brought to light initially by Lindley (1957). The data sampling involves n independent draws from $N\left(\mu, \sigma^2\right)$, with σ^2 known and μ to be inferred. Model 0 corresponds to the simple or sharp hypothesis that $\mu = \mu_0$. Model 1 takes σ^2 as known and μ as unknown, with its prior distribution being $N\left(\mu_1, \sigma_1^2\right)$; the hyperparameters are assumed to be known. This corresponds to a classical setting in which Model 0 is the null hypothesis $\mu = \mu_0$, and Model 1 is the alternative that the parameter can take any value other than $\mu = \mu_0$.

The marginal density of the data under Model 0 is

$$p_0\left(\mathbf{y}|\mu_0, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \mu_0\right)^2\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \overline{y}\right)^2\right] \exp\left[-\frac{n}{2\sigma^2}\left(\overline{y} - \mu_0\right)^2\right]. \tag{8.11}$$

The marginal density under Model 1 can be written as

$$p_1\left(\mathbf{y}|\mu_1,\sigma_1^2,\sigma^2\right)=\int p\left(\mathbf{y}|\mu,\sigma^2\right)p\left(\mu|\mu_1,\sigma_1^2\right)d\mu$$
$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_1^2}}$$
$$\int\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu\right)^2\right]\exp\left[-\frac{\left(\mu-\mu_1\right)^2}{2\sigma_1^2}\right]d\mu. \tag{8.12}$$

The Bayes factor for Model 0 relative to Model 1 is given by the ratio between (8.11) and (8.12)

$$B_{01}=\frac{\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right]}{\frac{1}{\sqrt{2\pi\sigma_1^2}}\int\exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}\left(\mu-\overline{y}\right)^2+\frac{\left(\mu-\mu_1\right)^2}{\sigma_1^2}\right]\right\}d\mu}. \tag{8.13}$$

Now the two quadratic forms on μ in the integrand can be combined in the usual manner, leading to

$$\frac{n}{\sigma^2}\left(\mu-\overline{y}\right)^2+\frac{\left(\mu-\mu_1\right)^2}{\sigma_1^2}=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)\left(\mu-\widehat{\mu}\right)^2$$
$$+\frac{n}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\frac{1}{\sigma_1^2}\left(\overline{y}-\mu_1\right)^2.$$

Above

$$\widehat{\mu}=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\mu_1}{\sigma_1^2}\right).$$

Carrying out the integration, the Bayes factor becomes, after some algebra,

$$B_{01}=\sqrt{\frac{\sigma_1^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}}}\frac{\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right]}{\exp\left[-\frac{n}{2\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\frac{1}{\sigma_1^2}\left(\overline{y}-\mu_1\right)^2\right]}. \tag{8.14}$$

Now examine what happens when the prior information becomes more and more diffuse, that is, eventually σ_1^2 is so large that $1/\sigma_1^2$ is near 0. The Bayes factor is, approximately,

$$B_{01}\approx\sqrt{\frac{\sigma_1^2\sigma^2}{n}}\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right].$$

For any fixed value of $\overline{y}$, the Bayes factor goes to ∞ when $\sigma_1^2\to\infty$, which implies that $p\left(Model\ 0|\mathbf{y}\right)\to 1$. This means that no matter what the value of $\overline{y}$ is, the null hypothesis would tend to be favored, even for values of

$\left|(\overline{y}-\mu_0)/\sqrt{\sigma^2/n}\right|$ that are large enough to cause rejection of the null at any, arbitrary, "significance" level in classical testing. This result, known as "Lindley's paradox", illustrates that a comparison of models in which one of the hypothesis is "sharp" (simple), strongly depends on the form of the prior distribution. In particular, when the distribution is improper, the Bayes factor leads to acceptance of the null. O'Hagan (1994) concludes that improper priors cannot be used when comparing models. However, the problem is not avoided entirely by adopting vague uniform priors over some large but finite range. This will be discussed later. ■

Comparing Two Composite Hypotheses

Third, the comparison may be a "composite versus composite", that is, one where the two models allow their respective parameters to take any values in the corresponding spaces. Here the Bayes factor is

$$B_{10}=\frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_0\right)}=\frac{\int p\left(\mathbf{y}|\boldsymbol{\theta}_1,M_1\right)p_1\left(\boldsymbol{\theta}_1|M_1\right)d\boldsymbol{\theta}}{\int p\left(\mathbf{y}|\boldsymbol{\theta}_0,M_0\right)p_0\left(\boldsymbol{\theta}_0|M_0\right)d\boldsymbol{\theta}}. \tag{8.15}$$

In general, as is the case with likelihood ratios, all constants appearing in $p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)$ must be included when computing B_{10}.

Example 8.4 *Marginal distributions and the Bayes factor in Poisson and negative binomial models*
The setting is as in Example 8.1, but the parameter values are allowed to take any values in their spaces. Following Bernardo and Smith (1994), take as prior distribution for the Poisson parameter

$$\theta_P\sim Ga\left(a_P,b_P\right),$$

and for the parameter of the negative binomial model adopt as prior the Beta distribution

$$\theta_N\sim Be\left(a_N,b_N\right).$$

The a's and b's are known hyperparameters. The marginal distribution of the data under the Poisson model, using (8.8) as likelihood function (suppressing the dependence on hyperparameters in the notation), is obtained as

$$\begin{aligned}p\left(\mathbf{y}|P\right)&=\int\frac{\theta_P^{n\overline{y}}\exp\left(-n\theta_P\right)}{\prod_{i=1}^{n}y_i!}\frac{b_P^{a_P}}{\Gamma\left(a_P\right)}\theta_P^{a_P-1}\exp\left(-b_P\theta_P\right)d\theta_P\\&=\frac{b_P^{a_P}}{\Gamma\left(a_P\right)\prod_{i=1}^{n}y_i!}\int\theta_P^{n\overline{y}+a_P-1}\exp\left[-\left(n+b_P\right)\theta_P\right]d\theta_P.\end{aligned} \tag{8.16}$$

The integrand is the kernel of the density of the

$$\theta_P\sim Ga\left(n\overline{y}+a_P,n+b_P\right)$$

distribution. Hence, the marginal of interest is

$$p(\mathbf{y}|P) = \frac{\Gamma(a_P + n\overline{y})\, b_P^{a_P}}{\Gamma(a_P)\,(n+b_P)^{a_P+n\overline{y}} \prod_{i=1}^{n} y_i!}. \tag{8.17}$$

Similarly, using (8.9), the marginal distribution of the data under the negative binomial model takes the form

$$\begin{aligned} p(\mathbf{y}|N) &= \int \theta_N^n (1-\theta_N)^{n\overline{y}} \frac{\Gamma(a_N+b_N)}{\Gamma(a_N)\Gamma(b_N)} \theta_N^{a_N-1} (1-\theta_N)^{b_N-1}\, d\theta_N \\ &= \frac{\Gamma(a_N+b_N)}{\Gamma(a_N)\Gamma(b_N)} \int \theta_N^{a_N+n-1} (1-\theta_N)^{b_N+n\overline{y}-1}\, d\theta_N. \end{aligned}$$

The integrand is the kernel of a beta density, so the integral can be evaluated analytically, yielding

$$p(\mathbf{y}|N) = \frac{\Gamma(a_N+b_N)}{\Gamma(a_N)\Gamma(b_N)} \frac{\Gamma(a_N+n)\Gamma(b_N+n\overline{y})}{\Gamma(a_N+n+b_N+n\overline{y})}. \tag{8.18}$$

The Bayes factor in favor of the N model relative to the P model is given by the ratio between (8.18) and (8.17). Note that the two marginal densities and the Bayes factor depend only on the data and on the hyperparameters, contrary to the ratio of likelihoods. This is because all unknown parameters are integrated out in the process of finding the marginals. It cannot be overemphasized that all integration constants must be kept when calculating the Bayes factors. In classical likelihood ratio tests, on the other hand, only those parts of the density functions that depend on the parameters are kept. ■

8.2.4 Influence of the Prior Distribution

From its definition, and from Example 8.4, it should be apparent that the Bayes factor depends on the prior distributions adopted for the competing models. The exception is when two simple hypotheses are at play. For the Poisson versus Negative Binomial setting discussed above, Bernardo and Smith (1994) give numerical examples illustrating that minor changes in the values of the hyperparameters produce changes in the direction of the Bayes factors. This dependence is illustrated with a few examples in what follows. Before we do so, note that one can write

$$\begin{aligned} \int p(\mathbf{y}|M_i)\, d\mathbf{y} &= \int\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i d\mathbf{y} \\ &= \int p(\boldsymbol{\theta}_i|M_i) \left[\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\mathbf{y} \right] d\boldsymbol{\theta}_i \\ &= \int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i, \end{aligned}$$

where the last equality follows because $\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\mathbf{y} = 1$. The message here is that when $p(\boldsymbol{\theta}_i|M_i)$ is improper, so is $p(\mathbf{y}|M_i)$. In this case, the Bayes factor is not well defined. This is discussed further in Subsection 8.2.5 below.

Example 8.5 *Influence of the bounds of a uniform prior*
Let the sampling model be $y_i|\mu \sim N(\mu, 1)$ and let the prior distribution adopted for μ under Model 1 be uniform over $[-L, L]$. Model 2 postulates the same sampling model but the bounds are $[-\alpha L, \alpha L]$, where α is a known, positive, real number. Suppose n independent samples are drawn, so that the marginal density of the data under Model 2 is

$$p(\mathbf{y}|\alpha, L) = \int_{-\alpha L}^{\alpha L} \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \frac{1}{2\alpha L} d\mu$$

$$= \frac{1}{2\alpha L}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2\right] \int_{-\alpha L}^{\alpha L} \exp\left[-\frac{n}{2}(\mu - \overline{y})^2\right] d\mu.$$

The integrand is in a normal form and can be evaluated readily, yielding

$$p(\mathbf{y}|\alpha, L) = \frac{1}{2\alpha L}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2\right]$$
$$\times \left[\Phi\left(\frac{\alpha L - \overline{y}}{\sqrt{\frac{1}{n}}}\right) - \Phi\left(\frac{-\alpha L - \overline{y}}{\sqrt{\frac{1}{n}}}\right)\right] \sqrt{2\pi \frac{1}{n}}.$$

The Bayes factor for Model 2 relative to Model 1 is

$$B_{21} = \frac{p(\mathbf{y}|\alpha, L)}{p(\mathbf{y}|1, L)} = \frac{\Phi\left(\frac{\alpha L - \overline{y}}{\sqrt{\frac{1}{n}}}\right) - \Phi\left(\frac{-\alpha L - \overline{y}}{\sqrt{\frac{1}{n}}}\right)}{\alpha\left[\Phi\left(\frac{L - \overline{y}}{\sqrt{\frac{1}{n}}}\right) - \Phi\left(\frac{-L - \overline{y}}{\sqrt{\frac{1}{n}}}\right)\right]}.$$

This clearly shows that the Bayes factor is sensitive with respect to the value of α. This is relevant in conjunction with the problem outlined in Example 8.3: the difficulties caused by improper priors in model selection via the Bayes factors are not solved satisfactorily by adopting, for example, bounded uniform priors. The Bayes factor depends very strongly on the width of the interval used. ■

Example 8.6 *The Bayes factor for a simple linear model*
Consider the linear model

$$\mathbf{y} = \boldsymbol{\beta} + \mathbf{e},$$

where the variance of the residual distribution, σ^2, is known, so it can be set equal to 1 without loss of generality. Model 1 posits as prior distribution $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_1^2\right)$, whereas for Model 2 the prior distribution is $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_2^2\right)$. Since the sampling model and the priors are both normal, it follows that the marginal distributions of the data are normal as well. The means and variances of these distributions are arrived at directly by taking expectations of the sampling model with respect to the appropriate prior. One gets $\mathbf{y}|\sigma^2, \sigma_1^2 \sim N\left(\mathbf{0}, \mathbf{I}\left(\sigma_1^2 + \sigma^2\right)\right)$ and $\mathbf{y}|\sigma^2, \sigma_2^2 \sim N\left(\mathbf{0}, \mathbf{I}\left(\sigma_2^2 + \sigma^2\right)\right)$ for Models 1 and 2, respectively. The Bayes factor for Model 1 relative to Model 2 is

$$\begin{aligned} B_{12} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\left(\sigma_1^2+1\right)}} \exp\left[-\frac{y_i^2}{2\left(\sigma_1^2+1\right)}\right]}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\left(\sigma_2^2+1\right)}} \exp\left[-\frac{y_i^2}{2\left(\sigma_2^2+1\right)}\right]} \\ &= \left(\frac{\sigma_2^2+1}{\sigma_1^2+1}\right)^{\frac{n}{2}} \frac{\exp\left[\frac{\mathbf{y}'\mathbf{y}}{2\left(\sigma_2^2+1\right)}\right]}{\exp\left[\frac{\mathbf{y}'\mathbf{y}}{2\left(\sigma_1^2+1\right)}\right]}. \end{aligned}$$

Taking logarithms and multiplying by 2, to arrive at the same scale as the likelihood ratio statistic, yields

$$2\log\left(B_{12}\right) = n\log\left(\frac{\sigma_2^2+1}{\sigma_1^2+1}\right) + \mathbf{y}'\mathbf{y}\left[\frac{\sigma_1^2-\sigma_2^2}{\left(\sigma_1^2+1\right)\left(\sigma_2^2+1\right)}\right].$$

The first term will contribute toward favoring Model 1 whenever σ_2^2 is larger than σ_1^2, whereas the opposite occurs in the second term. ■

8.2.5 Nested Models

As seen in Chapter 3, a nested model is one that can be viewed as a special case of a more general, larger model, and is typically obtained by fixing or "zeroing in" some parameters in the latter. Following O'Hagan (1994), let the bigger model have parameters (θ, ϕ) and denote it Model 1 whereas in the nested model fix $\phi = \phi_0$, with this value being usually 0. This is Model 0.

Let the prior probability of the larger model (often called the "alternative" one) be π_1, and let the prior density of its parameters be $p_1(\theta, \phi)$. The prior probability of the nested model is $\pi_0 = 1 - \pi_1$, which can be interpreted as the prior probability that $\phi = \phi_0$. This is somewhat perplexing at first sight, since the probability that a continuous parameter takes a given value is 0. However, the fact that consideration is given to the nested model as a plausible model implies that one is assigning some probability to the special situation that $\phi = \phi_0$ holds. In the nested model,

the prior density of the "free parameters" is $p_0(\theta) = p(\theta|\phi = \phi_0)$, that is, the density of the conditional distribution of the theta parameter, given that $\phi = \phi_0$. Now, for the larger model, write

$$p_1(\theta, \phi) = p_1(\theta|\phi)\, p_1(\phi),$$

where $p_1(\theta|\phi)$ is the density of the conditional distribution of θ, given ϕ. In practice, it is reasonable to assume that the conditional density of θ, given ϕ, is continuous at $\phi = \phi_0$ (O'Hagan, 1994).

In order to obtain the marginal density of the data under Model 0 one must integrate the joint density of the observations, and of the free parameters (given $\phi = \phi_0$) with respect to the latter, to obtain

$$\begin{aligned} p(\mathbf{y}|\text{Model 0}) &= \int p(\mathbf{y}|\theta, \phi = \phi_0)\, p(\theta|\phi = \phi_0)\, d\theta \\ &= p(\mathbf{y}|\phi = \phi_0). \end{aligned} \tag{8.19}$$

For the larger model

$$\begin{aligned} p(\mathbf{y}|\text{Model 1}) &= \int \left[\int p(\mathbf{y}|\theta, \phi)\, p(\theta|\phi)\, d\theta \right] p_1(\phi)\, d\phi \\ &= \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi = E_\phi \left[p(\mathbf{y}|\phi) \right]. \end{aligned} \tag{8.20}$$

The expectation above is an average of the sampling model marginal densities (after integrating out θ) taken over all values of ϕ (other than ϕ_0) and with plausibility as conveyed by the prior density $p_1(\phi)$ under the larger model. The posterior probability of the null model is then

$$\begin{aligned} p(\text{Model 0}|\mathbf{y}) &= \frac{p(\mathbf{y}|\text{Model 0})\, \pi_0}{p(\mathbf{y}|\text{Model 0})\, \pi_0 + p(\mathbf{y}|\text{Model 1})\, (1 - \pi_0)} \\ &= \frac{p(\mathbf{y}|\phi = \phi_0)\, \pi_0}{p(\mathbf{y}|\phi = \phi_0)\, \pi_0 + \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi\, (1 - \pi_0)}, \end{aligned} \tag{8.21}$$

and $p(\text{Model 1}|\mathbf{y}) = 1 - p(\text{Model 0}|\mathbf{y})$.

Consider now the case where there is a single parameter, so that Model 0 poses $\phi = \phi_0$ and Model 1 corresponds to the "alternative" hypothesis $\phi \neq \phi_0$ (the problem then consists of one of evaluating the "sharp" null hypothesis $\phi = \phi_0$). Then (8.21) holds as well, with the only difference being that the marginal distributions of the data are calculated directly as

$$p(\mathbf{y}|\text{Model 0}) = p(\mathbf{y}|\phi = \phi_0),$$

and

$$p(\mathbf{y}|\text{Model 1}) = \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi. \tag{8.22}$$

It is instructive to study the consequences of using a vague prior distribution on the Bayes factor. Following O'Hagan (1994), suppose that ϕ is a scalar parameter on $(-\infty, \infty)$. Vague prior knowledge is expressed as the limit of a uniform distribution

$$p_1(\phi) = (2c)^{-1}, \text{ for } -c \leq \phi \leq c,$$

by letting $c \to \infty$, in which case, $p_1(\phi) \to 0$. Then (8.22) is

$$\int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi = (2c)^{-1} \int_{-c}^{c} p(\mathbf{y}|\phi)\, d\phi.$$

Often, $p(\mathbf{y}|\phi)$ will tend to zero as ϕ tends to infinity, such that the limit of the integral above is finite. Then as $c \to \infty$, (8.22) tends to zero and the Bayes factor B_{01} tends to infinity. Thus, using a prior with very large spread on ϕ in an attempt to describe vague prior knowledge, forces the Bayes factor to favor Model 0.

Example 8.7 *Normal model: known versus unknown variance*
The setting will be the usual $N(\mu, \sigma^2)$ for each of n independent observations. In the larger model, both the mean and variance are taken as unknown. In the nested model, the variance is assumed to be known, such that $\sigma^2 = \sigma_0^2$. As in O'Hagan (1994), it will be assumed that the conditional prior distribution of the mean is the normal process $\mu|\mu_1, w\sigma^2 \sim N(\mu_1, w\sigma^2)$, where w is a known scalar. This implies that the variance of the prior distribution is proportional to that of the sampling model. Further, it will be assumed that the prior distribution of σ^2 is a scaled inverted chi-square distribution with parameters ν and S^2.

Under the null or nested model (known variance), the prior distribution is then $\mu|\mu_1, w\sigma_0^2 \sim N(\mu_1, w\sigma_0^2)$, and the marginal distribution of the data, following (8.19) and making use of (8.12), is

$$\begin{aligned} p(\mathbf{y}|\text{Model 0}) &= \int p(\mathbf{y}|\mu, \sigma_0^2) p(\mu|\mu_1, w\sigma_0^2)\, d\mu \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \overline{y})^2\right] \\ \times \frac{1}{\sqrt{2\pi w\sigma_0^2}} &\int \exp\left[-\frac{n}{2\sigma_0^2}(\overline{y} - \mu)^2\right] \exp\left[-\frac{(\mu - \mu_1)^2}{2w\sigma_0^2}\right] d\mu. \end{aligned}$$

Combining the two quadratics in μ gives

$$p(\mathbf{y}|\text{Model 0}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i-\overline{y})^2\right]\frac{1}{\sqrt{2\pi w\sigma_0^2}}$$
$$\exp\left[-\frac{n}{2\sigma_0^2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)^{-1}\frac{1}{w\sigma_0^2}(\overline{y}-\mu_1)^2\right]$$
$$\int \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)(\mu-\widehat{\mu})^2\right]d\mu,$$

where $\widehat{\mu}$ has the same form as in Example 8.3. After the integration is carried out, one gets

$$p(\mathbf{y}|\text{Model 0}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i-\overline{y})^2\right]\frac{1}{\sqrt{2\pi w\sigma_0^2}}$$
$$\exp\left[-\frac{1}{2}\frac{n}{\sigma_0^2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)^{-1}\frac{1}{w\sigma_0^2}(\overline{y}-\mu_1)^2\right]\sqrt{2\pi\sigma_0^2\left(n+\frac{1}{w}\right)^{-1}}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \frac{1}{\sqrt{nw+1}}\exp\left[-\frac{Q_y}{2\sigma_0^2}\right] = p\left(\mathbf{y}|\sigma^2{=}\sigma_0^2\right), \tag{8.23}$$

where

$$Q_y = \sum_{i=1}^n (y_i-\overline{y})^2 + n\left(n+\frac{1}{w}\right)^{-1}\frac{1}{w}(\overline{y}-\mu_1)^2.$$

In order to obtain the marginal density of the data under Model 1, use is made of (8.20) and of (8.23), although noting that σ^2 is now a free parameter. Then, recalling that the prior distribution of σ^2 is scaled inverted chi-square

$$p(\mathbf{y}|\text{Model 1}) = \int p\left(\mathbf{y}|\sigma^2\right)p\left(\sigma^2|\nu,S^2\right)d\sigma^2$$
$$= \int \frac{\left(2\pi\sigma^2\right)^{-\frac{n}{2}}}{(nw+1)^{\frac{1}{2}}}\exp\left[-\frac{Q_y}{2\sigma^2}\right]\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\left(\sigma^2\right)^{-\left(\frac{\nu+2}{2}\right)}\exp\left(-\frac{\nu S^2}{2\sigma^2}\right)d\sigma^2$$
$$= \frac{(2\pi)^{-\frac{n}{2}}}{(nw+1)^{\frac{1}{2}}}\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int\left(\sigma^2\right)^{-\left(\frac{n+\nu+2}{2}\right)}\exp\left(-\frac{\nu S^2+Q_y}{2\sigma^2}\right)d\sigma^2$$
$$= \frac{(2\pi)^{-\frac{n}{2}}}{(nw+1)^{\frac{1}{2}}}\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu S^2+Q_y}{2}\right)^{-\left(\frac{n+\nu}{2}\right)}. \tag{8.24}$$

In order to arrive at the last result, use is made of the gamma integrals (see Chapter 1). The Bayes factor in favor of Model 1 relative to Model 0

is given by the ratio between (8.24) and (8.23) yielding

$$B_{10} = \frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}} \Gamma\left(\frac{n+\nu}{2}\right)}{\left(\sigma_0^2\right)^{-\frac{n}{2}} \exp\left[-\frac{Q_y}{2\sigma_0^2}\right] \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu S^2 + Q_y}{2}\right)^{-\left(\frac{n+\nu}{2}\right)}.$$

■

8.2.6 Approximations to the Bayes Factor

There is extensive literature describing various approximate criteria for Bayesian model selection. Some have been motivated by the desire for suppressing the dependence of the final results on the prior. Ease of computation has also been an important consideration, especially in the pre-MCMC era. Some of these methods are still a useful part of the toolkit for comparing models. This section introduces widely used approximations to the Bayes factor based on asymptotic arguments. The latter are based on regularity conditions which fail when the parameter lies on a boundary of its parameter space (Pauler et al., 1999), a restriction not encountered with the Bayes factor.

The marginal density of the data under Model i, say, is

$$p\left(\mathbf{y}|M_i\right) = \int p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right) d\boldsymbol{\theta}_i, \tag{8.25}$$

where $\boldsymbol{\theta}_i$ is the $p_i \times 1$ vector of parameters under this model. In what follows, it will be assumed that the dimension of the parameter vector does not increase with the number of observations or that, if this occurs, it does so in a manner that, for n being the number of observations, p_i/n goes to 0 as $n \to \infty$. This is important for asymptotic theory to hold. In the context of quantitative genetic applications, there are models in which the number of parameters, e.g., the additive genetic effects, increases as the number of observations increases. For such models the approximations hold provided that these effects are first integrated out, in which case $p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)$ would be an integrated likelihood. For example, suppose a Gaussian linear model has f location parameters, n additive genetic effects (one for each individual), and two variance components. Then analytical integration of the additive effects (over their prior distribution) would need to be effected before proceeding. On the other hand, if the model is one of repeated measures taken on subjects or clusters (such as a family of half-sibs), it is reasonable to defend the assumption that p_i/n goes to 0 asymptotically.

Using the Posterior Mode

As in Chapter 7, expand the logarithm of the integrand in (8.25) around the posterior mode, $\widetilde{\boldsymbol{\theta}}_i$, using a second-order Taylor series expansion, to

obtain (recall that the gradient vanishes at the maximum value)

$$\begin{aligned}&\log\left[p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right)\right]\\ &\approx \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right] - \frac{1}{2}\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)' \left(\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}\right)\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right), \quad (8.26)\end{aligned}$$

where $\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}$ is the corresponding negative Hessian matrix. Then, using this in (8.25),

$$\begin{aligned}p\left(\mathbf{y}|M_i\right) &= \int \exp\left\{\log\left[p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right)\right]\right\} d\boldsymbol{\theta}_i\\ &\approx \exp\left\{\log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right]\right\}\\ &\times \int \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)' \left(\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}\right)\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)\right] d\boldsymbol{\theta}_i.\end{aligned}$$

The integral is in a Gaussian form (this approach to integration is called Laplace's method for integrals), so it can be evaluated readily. Hence

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} \left|\mathbf{H}^{-1}_{\widetilde{\boldsymbol{\theta}}_i}\right|^{\frac{1}{2}}, \quad (8.27)$$

where $\mathbf{H}^{-1}_{\widetilde{\boldsymbol{\theta}}_i}$ is the variance–covariance matrix of the Gaussian approximation to the posterior distribution. Further

$$\begin{aligned}\log\left[p\left(\mathbf{y}|M_i\right)\right] &\approx \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right]\\ &+ \frac{p_i}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\left|\mathbf{H}^{-1}_{\widetilde{\boldsymbol{\theta}}_i}\right|\right). \quad (8.28)\end{aligned}$$

Twice the logarithm of the Bayes factor for Model i relative to Model j, to express the "evidence brought up by the data" in support of Model i relative to j in the same scale as likelihood ratio tests, is then

$$\begin{aligned}2\log\left(B_{ij}\right) &\approx 2\log\left[\frac{p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_j, M_j\right)}\right] + 2\log\frac{p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widetilde{\boldsymbol{\theta}}_j|M_j\right)}\\ &+ \left(p_i - p_j\right)\log\left(2\pi\right) + \log\left(\frac{\left|\mathbf{H}^{-1}_{\widetilde{\boldsymbol{\theta}}_i}\right|}{\left|\mathbf{H}^{-1}_{\widetilde{\boldsymbol{\theta}}_j}\right|}\right). \quad (8.29)\end{aligned}$$

Note that the criterion depends on the log-likelihood ratios (evaluated at the posterior modes), on the log-prior ratios (also evaluated at the modes), on the difference between the dimensions of the two competing models, and on a Hessian adjustment.

Using the Maximum Likelihood Estimator

A variant to approximation (8.26) is when the expansion of the logarithm of the product of the prior density and of the conditional distribution of the observations (given the parameters) is about the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_i$, instead of the mode of the posterior distribution (Tierney and Kadane, 1989; O'Hagan, 1994; Kass and Raftery, 1995). Here one obtains in (8.27),

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right) p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} \left|\mathbf{H}^{-1}_{\widehat{\boldsymbol{\theta}}_i}\right|^{\frac{1}{2}}, \tag{8.30}$$

where $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}$ is the observed information matrix evaluated at the maximum likelihood estimator. In particular, if the observations are i.i.d. one has $\mathbf{H}_{\widehat{\boldsymbol{\theta}}} = n\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}}$, where $\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}}$ is the observed information matrix calculated from a single observation. Then

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right) p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} (n)^{-\frac{p_i}{2}} \left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_i}\right|^{\frac{1}{2}}. \tag{8.31}$$

The approximation to twice the logarithm of the Bayes factor becomes

$$\begin{aligned} 2\log\left(B_{ij}\right) \approx 2\log\left[\frac{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j, M_j\right)}\right] + 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right)} \\ - \left(p_i - p_j\right)\log\frac{n}{2\pi} + \log\frac{\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_i}\right|}{\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_j}\right|}. \end{aligned} \tag{8.32}$$

It is important to note that even though the asymptotic approximation to the posterior distribution (using the maximum likelihood estimator) does not depend on the prior, the resulting approximation to the Bayes factor does depend on the ratio of priors evaluated at the corresponding maximum likelihood estimators. If the term on the logarithm of the prior densities is excluded, the resulting expression is called the Bayesian information criterion (or BIC) (Schwarz, 1978; Kass and Raftery, 1995; Leonard and Hsu, 1999).

Suppose that the prior conveys some sort of "minimal" information represented by the distribution $\boldsymbol{\theta}_i|M_i \sim N\left(\widehat{\boldsymbol{\theta}}_i, \mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right)$. This is a unit information prior centered at the maximum likelihood estimator and having a precision (inverse of the covariance matrix) equivalent to that brought up

by a sample of size $n = 1$. Using this in (8.31):

$$p(\mathbf{y}|M_i) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)(2\pi)^{-\frac{p_i}{2}}\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right|^{-\frac{1}{2}}$$
$$\times \exp\left[-\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)'\left(\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}1,\widehat{\boldsymbol{\theta}}}\right)\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)\right](2\pi)^{\frac{p_i}{2}}(n)^{-\frac{p_i}{2}}\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right|^{\frac{1}{2}}$$
$$= p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)(n)^{-\frac{p_i}{2}}. \tag{8.33}$$

Hence

$$2\log(B_{ij}) \approx 2\log\left[\frac{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j, M_j\right)}\right] - (p_i - p_j)\log n. \tag{8.34}$$

This is Schwarz (1978) BIC in its most commonly presented form (Kass and Raftery, 1995; O'Hagan, 1994). Some authors (Leonard and Hsu, 1999; Congdon, 2001) use the term BIC to refer just to the approximated marginal densities, e.g., the logarithm of (8.33). At any rate, note that (8.34) is twice the maximized log-likelihood ratio, plus an adjustment that penalizes the model with more parameters. If $n = 1$, there is no penalty. However, the term $(p_i - p_j)$ becomes more important as a sample size increases. When $p_i > p_j$, (8.34) is smaller than twice the log-likelihood ratio, so the adjustment favors parsimony. In contrast, classical testing based on the traditional likelihood ratio tends to favor the more complex models. Contrary to the traditional likelihood ratio, BIC is well defined for nonnested models.

Denoting S the right hand side of (8.34) divided by 2, as sample size $n \to \infty$, this quantity satisfies:

$$\frac{S - \log B_{ij}}{\log B_{ij}} \to 0,$$

so it is consistent in this sense (Kass and Raftery, 1995). Recent extensions of BIC can be found in Kass (1995).

A related criterion is AIC (or the Akaike's information criterion) (Akaike, 1973), where the penalty is $2(p_i - p_j)$. The argument underlying the AIC is that if two models favor the data equally well, then the more parsimonious one should be favored. The BIC produces an even more drastic penalty, which increases with sample size, as noted.

The differences between the likelihood ratio criterion, the BIC, and the AIC are discussed by O'Hagan (1994) in the context of a nested model. The larger model has parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and dimension p_2, whereas the "smaller or null" model has a parameter vector $\boldsymbol{\theta}$ with p_1 elements and $p_2 - p_1$ fixed components $\boldsymbol{\phi} = \boldsymbol{\phi}_0$. For a large sample size, the log-likelihood ratio may favor the larger model, yet the penalty, $(p_2 - p_1)\log n$, may be severe enough so that the Bayes factor may end up favoring the null model.

It is instructive to examine the behavior of the approximation to the Bayes factor under repeated sampling from the appropriate model. Consider the BIC as given in (8.32), and take its expected value under the null model, with only the likelihood ratio viewed as a random variable. Recalling that the expected value of twice the log-likelihood ratio statistic under the null hypothesis is equal to the difference in dimension between the competing models, or $(p_2 - p_1)$, one gets

$$E\left[2\log\left(B_{21}\right)\right] \approx 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_2|M_2\right)}{p\left(\widehat{\boldsymbol{\theta}}_1|M_1\right)} - \left(p_2 - p_1\right)\left[\log\frac{n}{2\pi} - 1\right] + \text{constant}.$$

Hence, as $n \to \infty$, the expected value of the log of the Bayes factor in favor of the larger model goes to $-\infty$. This implies that the posterior probability of the larger model goes to 0 when the null model is true, regardless of the prior odds ratios as conveyed by $p\left(\widehat{\boldsymbol{\theta}}_2|M_2\right)/p\left(\widehat{\boldsymbol{\theta}}_1|M_1\right)$. Conversely, when the larger model is true, the expected value of twice the log-likelihood ratio statistic is approximately equal to $nQ\left(\boldsymbol{\phi}, \boldsymbol{\phi}_0\right)$,where $Q\left(\cdot\right)$ is a quadratic form (O'Hagan, 1994). This is a consequence of the asymptotically normal distribution of the maximum likelihood estimator (see Chapters 3 and 4). Then, under the larger model,

$$E\left[2\log\left(B_{ij}\right)\right] \approx nQ\left(\boldsymbol{\phi}, \boldsymbol{\phi}_0\right) + 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right)} - \left(p_2 - p_1\right)\log\frac{n}{2\pi} + \text{constant}.$$

As $n \to \infty$, the logarithm of the Bayes factor in favor of the larger model goes to ∞, since n grows faster than $\log n$. Consequently, the posterior probability of the larger model goes to 1, no matter what the prior odds are. Strictly from a classical point of view, and no matter how large n is, the null model will be rejected with probability equal to the significance level even when the model is true. Hence, more stringent significance levels should be adopted in classical hypothesis testing when sample sizes are large. Classical theory does not give a procedure for modifying the type-1 error as a function of sample size, and the probability of this error is prescribed arbitrarily. As noted by O'Hagan (1994), the Bayesian approach gives an automatic procedure in which in a single formula, such as (8.32), the evidence from the data, the prior odds, the model dimensionality, and the sample size are combined automatically.

8.2.7 Partial and Intrinsic Bayes Factors

The Bayes factor is only defined up to arbitrary constants when prior distributions are improper (i.e., Berger and Pericchi, 1996), as was illustrated at the end of Subsection 8.2.5. Further, when the priors are proper, the

Bayes factor depends on the form of the chosen prior distribution, as seen in connection with (8.32). This dependence does not decrease as sample size increases, contrary to the case of estimation of parameters from posterior distributions. In estimation problems and under regularity conditions, one can obtain an asymptotic approximation centered at the maximum likelihood estimator that does not involve the prior.

Berger and Pericchi (1996) suggested what are called intrinsic Bayes factors, in an attempt to circumvent the dependence on the prior, and to allow for the use of improper prior distributions, such as those based on Jeffreys' rule. Here, a brief overview of one of the several proposed types of Bayes factors (the arithmetic intrinsic Bayes factor) is presented.

Let the data vector of order n be partitioned as

$$\mathbf{y} = \left[\mathbf{y}'_{(1)}, \mathbf{y}'_{(2)}, \ldots, \mathbf{y}'_{(L)}\right]',$$

where $\mathbf{y}_{(l)}$, $(l = 1, 2, ..., L)$ denotes what is called the minimal training sample. This is the minimal number of observations needed for the posterior distribution to be proper. For example, if the minimal size of the training sample is m, there would be C_m^n different possible training samples. The posterior distribution based on the minimal training sample has density $p\left(\theta_i|\mathbf{y}_{(l)}, M_i\right)$. Further, put

$$\mathbf{y} = \left[\mathbf{y}'_{(l)}, \mathbf{y}'_{(-l)}\right]',$$

where $\mathbf{y}_{(-l)}$ is the data vector with $\mathbf{y}_{(l)}$ removed. Then the predictive density of $\mathbf{y}_{(-l)}$ under model i, conditionally on the data of the training sample $\mathbf{y}_{(l)}$, is

$$p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right) = \int p\left(\mathbf{y}_{(-l)}|\theta_i, \mathbf{y}_{(l)}, M_i\right) p\left(\theta_i|\mathbf{y}_{(l)}, M_i\right) d\theta_i.$$

The Bayes factor for model j relative to model i, conditionally on $\mathbf{y}_{(l)}$, or partial Bayes factor (O'Hagan, 1994) is

$$B_{ji}\left(\mathbf{y}_{(l)}\right) = \frac{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_j\right)}{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right)}. \tag{8.35}$$

Clearly, the partial Bayes factor depends on the choice of the training sample $\mathbf{y}_{(l)}$. To eliminate this dependence, Berger and Pericchi (1996) propose averaging $B_{ji}\left(\mathbf{y}_{(l)}\right)$ over all $C_m^n = K$ training samples. This yields the arithmetic intrinsic Bayes factor, defined formally as

$$B_{ji}^{AI} = \frac{1}{K}\sum_{l=1}^{K} \frac{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_j\right)}{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right)}. \tag{8.36}$$

This expression can be computed for any pair of models, irrespective of whether these are nested or not. Although the procedure is appealing, some difficulties arise. First, for most realistic hierarchical models it is not possible to determine in advance what the minimum sample size should be in order for the posterior to be proper. Second, and especially in animal breeding, the data sets are very large so, at best, just a few minimal training samples could be processed in practice.

There have been several other attempts to circumvent the need to using proper priors and to restrict the dependence on the prior. These are reviewed in O'Hagan (1994).

8.3 Estimating the Marginal Likelihood from Monte Carlo Samples

Except in highly stylized models, the integration indicated in (8.25) is not feasible by analytical means. An alternative is to use Monte Carlo methods. Here we shall consider the method of importance sampling, which will be encountered again in Chapters 12 and 15, where more details on the technique are given. Suppose samples of $\boldsymbol{\theta}_i$, the parameter vector under Model i, can be obtained from some known distribution that is relatively easy to sample from. This distribution, having the same support as the prior or posterior, is called the importance sampling distribution, and its density will be denoted as $g(\boldsymbol{\theta}_i)$. Then since $\int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i = 1$, the marginal density of the data under Model i is expressible as

$$\begin{aligned} p(\mathbf{y}|M_i) &= \frac{\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i}{\int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i} \\ &= \frac{\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}. \end{aligned} \tag{8.37}$$

Various Monte Carlo sampling schemes can be derived from (8.37), depending on the importance sampling function adopted. Suppose m samples can be obtained from the distribution with density $g(\boldsymbol{\theta}_i)$; let the samples be $\boldsymbol{\theta}_i^{[j]}$, $(j = 1, 2, \ldots, m)$. Then note that the denominator of (8.37) can be written as

$$\int \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i = \lim_{m\to\infty} \left[\frac{1}{m} \sum_{j=1}^{m} \frac{p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)}{g\left(\boldsymbol{\theta}_i^{[j]}\right)} \right],$$

where $p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)$ is the prior density under Model i evaluated at sampled value j. Likewise, the numerator can be written as

$$\int p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{g\left(\boldsymbol{\theta}_i\right)} g\left(\boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_i$$
$$= \lim_{m\to\infty} \left[\frac{1}{m}\sum_{j=1}^{m} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right) \frac{p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)}{g\left(\boldsymbol{\theta}_i^{[j]}\right)}\right],$$

where $p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right)$ is the density of the sampling model evaluated at the jth sample obtained from the importance distribution. Hence for large m, and putting $w_i^{[j]} = p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)/g\left(\boldsymbol{\theta}_i^{[j]}\right)$, a consistent estimator of (8.37) is given by the ratio

$$\widehat{p}\left(\mathbf{y}|M_i\right) = \frac{\sum_{j=1}^{m} w_i^{[j]} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right)}{\sum_{j=1}^{m} w_i^{[j]}}, \tag{8.38}$$

which is a weighted average of the density of the sampling distribution evaluated at the corresponding sampled values of the parameter vector under the appropriate model.

Sampling from the Prior

If the importance distribution is the prior, each of the weights $w_i^{[j]}$ are equal to 1, and the Monte Carlo estimator (8.38) of the marginal density at the observed value of $\mathbf{y}$ becomes

$$\widehat{p}\left(\mathbf{y}|M_i\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right), \tag{8.39}$$

where the $\boldsymbol{\theta}_i^{[j]}$ are draws from the prior distribution. The procedure is very simple because the joint prior distribution of the parameters is often simple to sample from. However, the estimator is imprecise because, typically, the $\boldsymbol{\theta}_i^{[j]}$ drawn from the prior are conferred little likelihood by the data. There will be just a few draws that will have appreciable likelihood and these will "dominate" the average (Kass and Raftery, 1995). Numerical studies can be found in McCulloch and Rossi (1991).

Sampling from the Posterior

If the importance distribution is the posterior, then

$$w_i = \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right)} = \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{\frac{p(\mathbf{y}|\boldsymbol{\theta}_i,M_i)p(\boldsymbol{\theta}_i|M_i)}{p(\mathbf{y}|M_i)}} = \frac{p\left(\mathbf{y}|M_i\right)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)}.$$

Using this in (8.38):

$$\begin{aligned}\widehat{p}\left(\mathbf{y}|M_i\right) &= \frac{\sum_{j=1}^{m} \frac{p(\mathbf{y}|M_i)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}{\sum_{j=1}^{m} \frac{p(\mathbf{y}|M_i)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}} \\ &= \frac{m}{\sum_{j=1}^{m} \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}} \\ &= \left[\frac{1}{m}\sum_{j=1}^{m} \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}\right]^{-1}. \end{aligned} \tag{8.40}$$

This estimator, the harmonic mean of the likelihood values, was derived by Newton and Raftery (1994), but arguing directly from Bayes theorem. Observe that a rearrangement of the theorem leads to

$$\frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{p\left(\mathbf{y}|M_i\right)} = \frac{p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)}.$$

Then, integrating both sides with respect to $\boldsymbol{\theta}_i$, yields

$$\frac{1}{p\left(\mathbf{y}|M_i\right)} \int p\left(\boldsymbol{\theta}_i|M_i\right) d\boldsymbol{\theta}_i = \int \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)} p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right) d\boldsymbol{\theta}_i.$$

Since the prior must be proper for the marginal density of the data to be defined, the integral on the left is equal to 1 leading directly to

$$p\left(\mathbf{y}|M_i\right) = \frac{1}{E_{\boldsymbol{\theta}_i|\mathbf{y},M_i}\left[p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)\right]}. \tag{8.41}$$

The Monte Carlo estimator of the reciprocal of the posterior expectation of the reciprocal of the likelihood values is precisely (8.40). An advantage of the harmonic mean estimator is that one does not need to know the form of the posterior distribution. The Markov chain Monte Carlo methods presented in the next part of the book enable one to draw samples from complex, unknown, distributions. The disadvantage, however, is its

numerical instability. The form of (8.40) reveals that values of $\boldsymbol{\theta}_i$ with very small likelihood can have a strong impact on the estimator. An alternative is to form some robust estimator of the harmonic mean (Congdon, 2001) such as a trimmed average. Kass and Raftery (1995) state that, in spite of the lack of stability, the estimator is accurate enough for interpretation on a logarithmic scale.

Caution must be exercised in the actual computation of (8.40), to avoid numerical over- or under-flows. A possible strategy could be as follows. Let

$$\begin{aligned} v &= \frac{1}{m}\sum_{j=1}^{m} p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}^{[j]}, M_i\right) \\ &= \frac{1}{m}\sum_{j=1}^{n} S_i^{[j]}, \end{aligned}$$

where $S_i^{[j]} = p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}^{[j]}, M_i\right)$, and store $\log S_i^{[j]}$ in a file for each sampled value. Then, since

$$\exp(x) = \exp(x - c + c) = \exp(x - c)\exp c,$$

one can write v in the form

$$v = \frac{1}{m}\sum_{j=1}^{m} \exp\left(\log S_i^{[j]} - c\right)\exp c,$$

where c is the largest value of $\log S_i^{[j]}$. Taking logarithms yields

$$\log v = \log\left[\frac{1}{m}\sum_{j=1}^{m} \exp\left(\log S_i^{[j]} - c\right)\right] + c.$$

Hence

$$\log\left[\widehat{p}\,(\mathbf{y}|M_i)\right] = -\log v.$$

Chib's Method

Most often, the marginal posterior distributions cannot be identified. However, there are many models where the conditional posterior distributions can be arrived at from inspection of the joint posterior densities. Advantage of this is taken in a Markov chain-based method called the Gibbs sampler, which will be introduced in Chapter 11. Chib (1995) outlined a procedure for estimating the marginal density of the data under a given model when the fully conditional posterior distributions can be identified. These distributions are defined in Section 11.5.1 of Chapter 11. We will

suppress the dependency on the model in the notation, for simplicity. Suppose the parameter vector is partitioned as $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$. The logarithm of the marginal density of the data can be expressed as

$$\begin{aligned}\log p(\mathbf{y}) &= \log\left[p(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] + \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] - \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\right] \\ &= \log\left[p(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] + \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] - \log\left[p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})\right] - \log\left[p(\boldsymbol{\theta}_1|\mathbf{y})\right].\end{aligned}$$

Suppose now that samples of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ have been drawn from the posterior distribution using the Gibbs sampler. Inspection of a large number of samples permits us to calculate, e.g., the posterior mean, mode, or median for each of the elements of the parameter vector, such that one can form, say, the vector of posterior medians $\widetilde{\boldsymbol{\theta}} = \left[\widetilde{\boldsymbol{\theta}}_1', \widetilde{\boldsymbol{\theta}}_2'\right]'$. An estimate of the marginal density of the data can be obtained as

$$\begin{aligned}\log \widehat{p}(\mathbf{y}) = \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1, \mathbf{y}\right)\right] \\ - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right)\right]. \qquad (8.42)\end{aligned}$$

If the conditional density $\log\left[p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})\right]$ is known, the third term can be evaluated readily. The difficulty resides in the fact that the marginal posterior density may not be known. However, recall that

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\right].$$

Hence

$$p\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right) = E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2, \mathbf{y}\right)\right],$$

and an estimate of the marginal posterior density can be obtained as

$$\widehat{p}\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{[j]}, \mathbf{y}\right),$$

where $\boldsymbol{\theta}_2^{[j]}$, $(j = 1, 2, \ldots, m)$ are samples from the marginal posterior distribution of $\boldsymbol{\theta}_2$ obtained with the Gibbs sampler. Then, using this in (8.42), the estimated marginal density of the data is arrived at as

$$\begin{aligned}\log \widehat{p}(\mathbf{y}) = \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1, \mathbf{y}\right)\right] \\ - \log\left[\frac{1}{m}\sum_{j=1}^{m} p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{[j]}, \mathbf{y}\right)\right]. \qquad (8.43)\end{aligned}$$

The procedure is then repeated for each of the models in order to calculate the Bayes factor. However, the fully conditional posterior distribution of one parameter given the other must be identifiable in each of the models. The method can be extended from two to several parameter blocks (Chib, 1995; Han and Carlin, 2001). Additional refinements of the procedure are in Chib and Jeliazkov (2001).

8.4 Goodness of Fit and Model Complexity

In general, as a model becomes increasingly more complex, i.e., by increasing the number of parameters, its fit gets better. For example, it is well known that if one fits n regression coefficients to a data set consisting of n points, the fit is perfect. As seen earlier, the AIC and BIC introduce penalties against more highly parameterized models. A slightly different approach was suggested by Spiegelhalter et al. (2002), and it is based on calculating the expected posterior deviance, i.e., a measure of fit.

Consider a model with parameter vector $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$. For example, in a mixed linear model $\boldsymbol{\theta}_1$, $(p_1 \times 1)$ may be a vector of "fixed" effects such as breed or sex of animal and variance parameters, while $\boldsymbol{\theta}_2$ may be a vector of random effects or missing data. Hence, in some sense, the dimension of $\boldsymbol{\theta}_1$ can be viewed as fixed, as the dimension of the data vector increases, whereas the order of $\boldsymbol{\theta}_2$ may perhaps increase with n. Clearly, neither the AIC nor the BIC can be used in models where the parameters outnumber the observations (Gelfand and Dey, 1994) unless some parameters are integrated out. In what follows it will be assumed that $\boldsymbol{\theta}_2$ can be integrated somehow, possibly by analytical means, and that one arrives at the integrated likelihood

$$p\left(\mathbf{y}|\boldsymbol{\theta}_1\right) = \int p\left(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) p\left(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1\right) d\boldsymbol{\theta}_2.$$

Here $p\left(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1\right)$ is the density of the conditional distribution of the nuisance parameters, given the primary model parameters. The dependency on the model will be suppressed in the notation. Now rearrange the components of Bayes theorem as

$$\frac{p\left(\boldsymbol{\theta}_1|\mathbf{y}\right)}{p\left(\boldsymbol{\theta}_1\right)} = \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{p\left(\mathbf{y}\right)}.$$

Taking expectations of the logarithm of both sides with respect to the posterior distribution of $\boldsymbol{\theta}_1$, one gets

$$\int \log\left[\frac{p\left(\boldsymbol{\theta}_1|\mathbf{y}\right)}{p\left(\boldsymbol{\theta}_1\right)}\right] p\left(\boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta}_1 = \int \log\left[\frac{p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{p\left(\mathbf{y}\right)}\right] p\left(\boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta}_1.$$

The left-hand side is the Kullback–Leibler distance between the posterior and prior distributions, so it is at least null. Following Dempster (1974, 1997), Spiegelhalter et al. (2002) suggest to view (given the model) $p\left(\mathbf{y}\right)$ as a standardizing term and to set $p\left(\mathbf{y}\right) = 1$. This can be construed as a "perfect predictor", giving probability 1 to each of the observations in the data set (prior to observation). Hence, the larger the logarithm of $p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)$ (the log-likelihood), the closer the model is to "perfect prediction". Setting $p\left(\mathbf{y}\right) = 1$ and multiplying the right-hand side of the above expression by

-2, define

$$\begin{aligned}\overline{D} &= -2\int \left[\log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)\right] p\left(\boldsymbol{\theta}_1|\mathbf{y}\right) d\boldsymbol{\theta}_1 \\ &= E_{\boldsymbol{\theta}_1|\mathbf{y}}\left[-2\log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)\right] \\ &= E_{\boldsymbol{\theta}_1|\mathbf{y}}\left[D\left(\boldsymbol{\theta}_1\right)\right], \end{aligned} \tag{8.44}$$

where $D\left(\boldsymbol{\theta}_1\right) = -2\log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)$ is called the deviance (a function of the unknown parameter), and $\overline{D}$ is its expected value taken over the posterior distribution. Note that when the deviance is evaluated at the maximum likelihood estimator, one obtains the numerator (or denominator) of the usual likelihood ratio statistic. Thus, one averages out the deviance criterion over values whose plausibilities are dictated by the posterior distribution. The expected deviance is interpreted as a posterior summary of the fit of the model. In general, $\overline{D}$ will need to be computed using Monte Carlo procedures for sampling from the posterior distribution: samples from the posterior are obtained, and then one averages the log-likelihoods evaluated at each of the draws.

Concerning model complexity (degree of parameterization), Spiegelhalter et al. (2002) suggest using the "effective number of parameters"

$$p_D = \overline{D} - D\left(\overline{\boldsymbol{\theta}}_1\right), \tag{8.45}$$

where $D\left(\overline{\boldsymbol{\theta}}_1\right)$ is the deviance evaluated at the posterior mean of the primary parameter vector. In order to motivate this concept, expand the deviance around the posterior mean $\overline{\boldsymbol{\theta}}_1$, to obtain

$$\begin{aligned} D\left(\boldsymbol{\theta}_1\right) \approx & -2\log p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_1\right) - 2\left[\frac{\partial \log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1}\right]'_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\left(\boldsymbol{\theta}_1 - \overline{\boldsymbol{\theta}}_1\right) \\ & - \left(\boldsymbol{\theta}_1 - \overline{\boldsymbol{\theta}}_1\right)'\left[\frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'}\right]\left(\boldsymbol{\theta}_1 - \overline{\boldsymbol{\theta}}_1\right). \end{aligned} \tag{8.46}$$

Taking the expectation of (8.46), with respect to the posterior distribution of the parameter vector, gives the expected deviance

$$\begin{aligned} \overline{D} &\approx -2\log p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_1\right) + tr\left[-\frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'}\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1} Var\left(\boldsymbol{\theta}_1|\mathbf{y}\right) \\ &= D\left(\overline{\boldsymbol{\theta}}_1\right) + tr\left\{\left[-\frac{\partial^2 \log p\left(\mathbf{y}|\boldsymbol{\theta}_1\right)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'}\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1} Var\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right\} \\ &= D\left(\overline{\boldsymbol{\theta}}_1\right) + tr\left\{\left[\mathbf{I}\left(\boldsymbol{\theta}_1\right)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1} Var\left(\boldsymbol{\theta}_1|\mathbf{y}\right)\right\}, \end{aligned} \tag{8.47}$$

where $\mathbf{I}\left(\boldsymbol{\theta}_1\right)$ is the observed information matrix and $Var\left(\boldsymbol{\theta}_1|\mathbf{y}\right)$ is the variance–covariance matrix of the posterior distribution. The trace adjustment in (8.47) is called the "effective number of parameters" and is denoted

p_D, following (8.45). Recall from Chapter 7, that an asymptotic approximation to the posterior distribution is given by a normal process having a covariance matrix that is equal to the inverse of the sum of the observed information matrix (evaluated at some mode), plus the negative Hessian of the log-prior density (evaluated at some mode); the latter will be denoted as $\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}$ when evaluated at the posterior mean. Hence, approximately,

$$\begin{aligned} p_D &\approx tr\left\{\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1} Var(\boldsymbol{\theta}_1|\mathbf{y})\right\} \\ &\approx tr\left\{\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\left(\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}+\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\right)^{-1}\right\}. \end{aligned} \tag{8.48}$$

Thus, the effective number of parameters can be interpreted as the information about $\boldsymbol{\theta}_1$ contained in the likelihood relative to the total information in both the likelihood and the prior. Some additional algebra yields

$$\begin{aligned} p_D &\approx tr\left\{\left(\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}+\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}-\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\right)\right. \\ &\qquad \left.\times\left(\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}+\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\right)^{-1}\right\} \\ &= p_1 - tr\left[\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\left(\left[\mathbf{I}(\boldsymbol{\theta}_1)\right]_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}+\mathbf{P}(\boldsymbol{\theta}_1)_{\boldsymbol{\theta}_1=\overline{\boldsymbol{\theta}}_1}\right)^{-1}\right]. \end{aligned} \tag{8.49}$$

This representation leads to the interpretation that the effective number of parameters is equal to the number of parameters in $\boldsymbol{\theta}_1$, minus an adjustment measuring the amount of information in the prior relative to the total information contained in the asymptotic approximation to the posterior.

Further, Spiegelhalter et al. (2002) suggested combining the measure of fit given by $\overline{D}$ (the posterior expectation of the deviance) in (8.44) with the effective number of parameters in (8.48) or (8.49), into a deviance information criterion (DIC). This is defined as

$$\begin{aligned} DIC &= \overline{D} + p_D \\ &= D\left(\overline{\boldsymbol{\theta}}_1\right) + 2p_D, \end{aligned} \tag{8.50}$$

with the last expression resulting from (8.45). Models having a smaller DIC should be favored, as this indicates a better fit and a lower degree of model complexity. The authors emphasize that they consider DIC to be a preliminary device for screening alternative models.

Example 8.8 *Deviance information criterion in the mixed linear model* Consider a hierarchical model with structure

$$\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \mathbf{e},$$

where $\mathbf{y}|\boldsymbol{\theta}, \mathbf{R} \sim N(\mathbf{W}\boldsymbol{\theta}, \mathbf{R})$. This model has been discussed several times, especially in Chapter 6. In animal breeding $\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{u}'\right]'$ is typically a vector of "fixed" and "random" effects, and the corresponding known incidence matrix is then $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$. Suppose that the dimension of $\boldsymbol{\beta}$ (p_β) does not

increase with the number of observations, and that the vector $\mathbf{u}$ (having order p_u) contains the effects of clusters, e.g., half-sib families. Hence, one can conceptually let the number of observations per cluster go to infinity (or, equivalently, think that the number of observations increases more rapidly than the number of clusters). Under these conditions, one can employ the asymptotic approximations discussed earlier. The second level of the hierarchy poses

$$\boldsymbol{\theta}|\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\mu}_{\mathbf{u}},\mathbf{V}_{\beta},\sigma^2_{\beta},\mathbf{G}_u \sim N\left(\left[\begin{array}{c}\boldsymbol{\mu}_{\beta}\\ \boldsymbol{\mu}_u\end{array}\right],\left[\begin{array}{cc}\mathbf{V}_{\beta}\sigma^2_{\beta} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_u\end{array}\right]\right).$$

The dispersion parameters $\mathbf{R},\mathbf{V}_{\beta},\sigma^2_{\beta},\mathbf{G}_u$, and the location vectors $\boldsymbol{\mu}_{\beta}$ and $\boldsymbol{\mu}_u$ are assumed known. As mentioned in Chapter 1, Example 1.18, and shown in Chapter 6, the posterior distribution of $\boldsymbol{\theta}$ is normal, with mean vector

$$\overline{\boldsymbol{\theta}}=\left[\begin{array}{c}\overline{\boldsymbol{\beta}}\\ \overline{\mathbf{u}}\end{array}\right]=\left[\begin{array}{cc}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\frac{\mathbf{V}_{\beta}^{-1}}{\sigma^2_{\beta}} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}_u^{-1}\end{array}\right]^{-1}$$
$$\times\left[\begin{array}{c}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}+\frac{\mathbf{V}_{\beta}^{-1}}{\sigma^2_{\beta}}\boldsymbol{\mu}_{\beta}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}+\mathbf{G}_{\mathbf{u}}^{-1}\boldsymbol{\mu}_u\end{array}\right],$$

and variance–covariance matrix

$$\mathbf{C}^{-1}=\left[\begin{array}{cc}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\frac{1}{\sigma^2_{\beta}}\mathbf{V}_{\beta}^{-1} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}_u^{-1}\end{array}\right]^{-1}.$$

The deviance is

$$\begin{aligned}D(\boldsymbol{\theta}) &= -2\log p(\mathbf{y}|\boldsymbol{\theta},\mathbf{R})\\ &= N\log(2\pi)+\log|\mathbf{R}|+(\mathbf{y}-\mathbf{W}\boldsymbol{\theta})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\theta}).\end{aligned}$$

Then,

$$D(\overline{\boldsymbol{\theta}})=N\log(2\pi)+\log|\mathbf{R}|+\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right),$$

and the expected deviance becomes

$$\begin{aligned}\overline{D}=N\log(2\pi)+\log|\mathbf{R}|+\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)\\ +tr\left(\mathbf{R}^{-1}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}'\right).\end{aligned}$$

Employing (8.45), the effective number of parameters is

$$\begin{aligned}p_D &= \overline{D}-D(\overline{\boldsymbol{\theta}})\\ &= tr\left(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}\right).\end{aligned}$$

For example, let $\mathbf{R} = \mathbf{I}\sigma_e^2$ and $\mathbf{G}_u = \mathbf{I}\sigma_u^2$, which results in a variance component model. Further, let $\sigma_\beta^2 \to \infty$, to make prior information about $\boldsymbol{\beta}$ vague. Here

$$\begin{aligned}\mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \sigma_e^2 \\ &= \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \sigma_e^2,\end{aligned}$$

and

$$\begin{aligned}\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \\ &= \mathbf{I}_{p_\beta + p_u} - \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix}.\end{aligned}$$

Hence

$$\begin{aligned}p_D &= tr\left(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}\right) \\ &= tr\left(\mathbf{I}_{p_\beta + p_u}\right) - tr\left\{ \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix} \right\} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} tr \begin{bmatrix} \mathbf{0} & \mathbf{C}^{\beta u} \\ \mathbf{0} & \mathbf{C}^{uu} \end{bmatrix} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} tr\left[\mathbf{C}^{uu}\right].\end{aligned}$$

Note that the prior information about the $\mathbf{u}$ vector results in that the effective number of parameters is smaller than the dimension of $\boldsymbol{\theta}$.

■

8.5 Goodness of Fit and Predictive Ability of a Model

The posterior probability of a model and the Bayes factors can be viewed as global measures of model relative plausibility. However, one often needs to go further than that. For example, a model can be the most plausible within a set of competing models and, yet, either be unable to predict the data at hand well or to give reasonable predictions of future observations. Here we will provide just a sketch of some of the procedures that can be used for gauging the quality of fit and predictive performance of a model.

8.5.1 Analysis of Residuals

A comprehensive account of techniques for examination of residuals is given by Barnett and Lewis (1995). In order to illustrate some of the basic ideas, consider, for example, a linear regression analysis. One of the most widely used techniques for assessing fit is to carry out a residual analysis (e.g., Draper and Smith, 1981). In the context of classical regression, one calculates the predicted value of an observation, $\widehat{y}$, and forms the Studentized fitted residual

$$\frac{y-\widehat{y}}{\sqrt{\widehat{\sigma}_e^2}},$$

where $\widehat{\sigma}_e^2$ is typically the unbiased estimator of the residual variance. If the absolute value of the Studentized residual exceeds a certain critical value of the t or normal distributions, then the observation is viewed as suspicious and regarded as a potential outlier. This may be construed as an indication that the model does not fit well.

The Bayesian counterpart of this classical regression analysis consists of examining the posterior distribution of the unobserved standardized quantity

$$r_i = \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sqrt{\sigma_e^2}},$$

where the row vector $\mathbf{x}_i'$ contains known explanatory variables linking the unknown regression vector $\boldsymbol{\beta}$ to y_i. Using the standard normality assumptions with independent and identically distributed errors, the distribution of r_i under the sampling model is $r_i \sim N(0,1)$, provided σ_e^2 is known. If $\boldsymbol{\beta}$ has the prior distribution $\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{V}_\beta \sim N(\boldsymbol{\alpha}, \mathbf{V}_\beta)$, where the hyperparameters are also known, one obtains as prior (or predictive) distribution of the residual above, given σ_e^2,

$$r_i|\boldsymbol{\alpha}, \sigma_e^2, \mathbf{V}_\beta \sim N\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\alpha}}{\sqrt{\sigma_e^2}}, \frac{\mathbf{x}_i'\mathbf{V}_\beta\mathbf{x}_i}{\sigma_e^2}\right).$$

The unconditional (with respect to σ_e^2) prior distribution of the standardized residual will depend on the prior adopted for σ_e^2. Then one could carry out an analysis of the residuals prior to proceeding with Bayesian learning about the parameters. More commonly, however, the residual analysis will be undertaken based on the joint posterior distribution of $\boldsymbol{\beta}$ and σ_e^2. As seen in Chapter 6, given σ_e^2, the posterior distribution of $\boldsymbol{\beta}$ is the normal process

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y} \sim N\left[\widetilde{\boldsymbol{\beta}}, \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\right],$$

where

$$\widetilde{\boldsymbol{\beta}} = \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{y}}{\sigma_e^2} + \mathbf{V}_{\boldsymbol{\beta}}^{-\mathbf{1}}\boldsymbol{\alpha}\right).$$

Further, given σ_e^2, the posterior distribution of the Studentized residual will have the form

$$r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y} \sim N\left[\frac{y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}}{\sqrt{\sigma_e^2}}, \frac{\mathbf{x}_i'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\mathbf{x}_i}{\sigma_e^2}\right].$$

The unconditional (with respect to σ_e^2) posterior distribution will depend on the form of the marginal posterior distribution of the residual variance, and its density is obtained as

$$p\left(r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y}\right) = \int p\left(r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y}\right) p\left(\sigma_e^2|\mathbf{y}\right) d\sigma_e^2,$$

where $p\left(\sigma_e^2|\mathbf{y}\right)$ is the marginal posterior density of the residual variance. Unless standard conjugate priors are adopted, the marginal posterior distribution of the Studentized residual cannot be arrived at in closed form. In such a situation, one can adopt the sampling techniques described in the third part of the book and obtain draws from the posterior distribution of the standardized residual. This is done simply by drawing from the posterior distribution of the model parameters. Then, for observation i, one forms samples

$$r_i^{[j]} = \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}^{[j]}}{\sqrt{\sigma_e^{2[j]}}}, \quad j = 1, 2, \ldots, m,$$

where $\boldsymbol{\beta}^{[j]}$ and $\sigma_e^{2[j]}$ are samples from the joint posterior distribution of the regression vector and of the residual variance. Thus, one obtains an entire distribution for each Studentized residual, which can be used to decide whether or not the observation is in reasonable agreement with what the model predicts. If the value 0 appears at high density in the posterior distribution, this can be construed as an indication that the observation is in conformity with the model.

This simple idea extends naturally to other models in which residuals are well defined. For example, for binary $(0, 1)$ responses analyzed with a probit model, Albert and Chib (1993, 1995) define the Bayesian residual $r_i = y_i - \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, which is real valued on the interval $[y_i - 1, y_i]$. If samples are taken from the posterior distribution of $\boldsymbol{\beta}$, one can form corresponding draws from the posterior distribution of each residual. Since $\Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$ takes values between 0 and 1, an observation $y_i = 0$ will be outlying if the posterior distribution of r_i is concentrated towards the endpoint -1, and an observation $y_i = 1$ is suspect if the posterior of r_i is concentrated towards the value 1. A value of 0 appearing at high density in the posterior distribution of the residuals can be interpreted as an indication of reasonable fit. Albert and Chib (1995) propose an alternative residual defined at the level of a latent variable called the liability (see Chapter 14 for a definition of this concept). The reader is referred to their paper for details.

8.5.2 Predictive Ability and Predictive Cross-Validation

Predictive ability and goodness of fit are distinct features of a model. A certain model may explain and predict adequately the observations used for model building. However, it may yield poor predictions of future observations or of data points that are outside the range represented in the data employed for model building. A number of techniques is available for gauging the predictive ability of a Bayesian model. Even though some attention is paid to foundational issues, the approaches here are often eclectic and explorative. They constitute an important set of tools for understanding the predictive ability of a model.

Cross-validation methods involve constructing the posterior distribution of the parameters but leaving some observations out. Then the predictive distributions of the observations that have been removed are derived to examine whether or not the actual data points fall in regions of reasonably high density. Partition the data as $\mathbf{y}' = \left[y_{\text{out}}, \mathbf{y}'_{-\text{out}}\right]$, where y_{out} is the observation to be removed, and $\mathbf{y}_{-\text{out}}$ is the vector of the remaining observations. The density of the posterior predictive distribution can be written as

$$p\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}, M\right) = \int p\left(y_{\text{out}}|\boldsymbol{\theta}, \mathbf{y}_{-\text{out}}, M\right) p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}, M\right) d\boldsymbol{\theta}, \tag{8.51}$$

where $p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}, M\right)$ is the density of the posterior distribution built from $\mathbf{y}_{-\text{out}}$ and model M. In hierarchical modeling, one typically writes the sampling distribution of the data such that conditional independence can be exploited. Thus, given the parameters, y_{out} is independent of $\mathbf{y}_{-\text{out}}$, and one can write (suppressing the notation denoting model M)

$$p\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \int p\left(y_{\text{out}}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}\right) d\boldsymbol{\theta}. \tag{8.52}$$

Since, in general, the form of the posterior density is unknown or analytically intractable, the predictive density will be calculated via Monte Carlo methods (Gelfand et al., 1992; Gelfand, 1996). For example, if m draws from the posterior distribution can be made via MCMC procedures, the form of (8.52) suggests the estimator

$$\widehat{p}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(y_{\text{out}}|\boldsymbol{\theta}^{[j]}\right),$$

where $\boldsymbol{\theta}^{[j]}$ is a draw from $\left[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}\right]$. The mean and variance of the predictive distribution can also be computed by Monte Carlo procedures. Since the expected value of the sampling model can almost always be deduced readily, e.g., in regression $E\left(y_{\text{out}}|\boldsymbol{\theta}\right) = \mathbf{x}'_{\text{out}}\boldsymbol{\beta}$, the mean of the predictive distribution can be estimated as

$$\widehat{E}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \frac{1}{m}\sum_{j=1}^{m} E\left(y_{\text{out}}|\boldsymbol{\theta}^{[j]}\right). \tag{8.53}$$

Similarly, a Monte Carlo estimate of the variance of the predictive distribution can be obtained as

$$\widehat{Var}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \widehat{E}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[Var\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right] + \widehat{Var}\left[E\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right]. \quad (8.54)$$

This can be illustrated with a regression model, although in this situation there is an analytical solution under the standard assumptions. For example, if the regression model postulates $y_{\text{out}}|\boldsymbol{\beta}, \sigma_e^2 \sim N\left(x'_{out}\boldsymbol{\beta}, \sigma_e^2\right)$, then

$$\widehat{E}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[Var\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right] = \frac{1}{m}\sum_{j=1}^{m}\sigma_e^{2[j]},$$

and

$$\begin{aligned}&\widehat{Var}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[E\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right] = \widehat{Var}\left[x'_{out}\boldsymbol{\beta}\right]\\ &= \frac{1}{m}\sum_{j=1}^{m}\left(x'_{out}\boldsymbol{\beta}^{[j]}\right)^2 - \left(\frac{1}{m}\sum_{j=1}^{m}x'_{out}\boldsymbol{\beta}^{[j]}\right)^2.\end{aligned}$$

Subsequently, the following composite statistic can be used to evaluate the overall predictive ability of the model (Congdon, 2001):

$$D^2 = \sum_{out=1}^{n}\left[\frac{y_{\text{out}} - \widehat{E}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right)}{\sqrt{\widehat{Var}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right)}}\right]^2. \quad (8.55)$$

Models having a smaller value of D^2 would be viewed as having a better predictive ability. Clearly, if n is very large, the computations may be taxing, since n posterior and predictive distributions need to be computed. Other statistics are described in Gelfand et al. (1992) and in Gelfand (1996).

A related idea has been advocated by Gelman et al. (1996). Rather than working with the leave-one-out method in (8.52), they propose generating data $\widetilde{\mathbf{y}}$ from the posterior predictive distribution with density

$$p\left(\widetilde{\mathbf{y}}|\mathbf{y},M\right) = \int p\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta},M\right)p\left(\boldsymbol{\theta}|\mathbf{y},M\right)d\boldsymbol{\theta}. \quad (8.56)$$

One then wishes to study whether the simulated value $\widetilde{\mathbf{y}}$ agrees with the observed data $\mathbf{y}$. Systematic differences between the simulations and the observed data indicate potential failure of model M. Various criteria or test quantities can be used to carry out the comparisons. Examples of these are given in Gelfand (1996). The choice of test quantities should be driven by the aspect of the model whose fit is in question and/or by the purpose with which the model will be used. The method of composition (introduced in Chapter 1), can be used to obtain draws from (8.56), and can be described as follows:

1. Draw $\boldsymbol{\theta}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, M)$. Ways of achieving this are discussed later in this book.

2. Draw $\widetilde{\mathbf{y}}$ from the sampling distribution $p(\widetilde{\mathbf{y}}|\boldsymbol{\theta},M)$. One has now a single realization from the joint distribution $p(\widetilde{\mathbf{y}}, \boldsymbol{\theta}|M)$.

3. Repeat steps 1 and 2 many times.

The set of $\widetilde{\mathbf{y}}'s$ drawn using this algorithm constitutes samples from (8.56). Letting $h(\mathbf{y})$ be a particular test quantity, for example, the average of the top 10 observations, one can then study whether $h(\mathbf{y})$ falls in a region of high posterior probability in the distribution $[h(\widetilde{\mathbf{y}})|\mathbf{y},M]$. This can be repeated for all the models under investigation. Gelman et al. (1996) propose the calculation of Bayesian p-values, p_B, for given test quantities $h(\mathbf{y}, \boldsymbol{\theta})$. The notation emphasizes that, in contrast with classical p-values, the test quantity can depend on both data and parameters. Then,

$$\begin{aligned} p_B &= p[h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) > h(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}] \\ &= \int\int I[h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) > h(\mathbf{y}, \boldsymbol{\theta})]\, p(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} d\widetilde{\mathbf{y}}, \end{aligned} \tag{8.57}$$

gives the probability that the simulated data $\widetilde{\mathbf{y}}$ is more extreme than the observed data $\mathbf{y}$, averaged over the distribution $[\boldsymbol{\theta}|\mathbf{y}]$. A possible test quantity could be

$$h(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{[y_i - E(Y_i|\boldsymbol{\theta})]^2}{Var(Y_i|\boldsymbol{\theta})}$$

and

$$h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\left[\widetilde{y}_i - E\left(\widetilde{Y}_i|\boldsymbol{\theta}\right)\right]^2}{Var\left(\widetilde{Y}_i|\boldsymbol{\theta}\right)}.$$

These are then used for computing (8.57). A cross-validation approach can also be implemented using this idea. An application of these techniques in animal breeding is in Sorensen et al. (2000).

Another way of assessing global predictive ability of a set of models was proposed by Geisser and Eddy (1979) and by Geisser (1993) via the conditional predictive ordinate (CPO). The logarithm of the CPO for Model i is

$$\log[CPO_{\text{Model } i}] = \sum_{out=1}^{n} \log[p(y_{\text{out}}|\mathbf{y}_{-\text{out}}, \text{Model } i)].$$

Gelfand and Dey (1994) describe techniques for calculating the CPO that avoid carrying out the n implementations of the sampling procedure described above. Chapter 12, especially Section 12.4, discusses Monte Carlo implementation of these quantities in more detail.

8.6 Bayesian Model Averaging

8.6.1 General

Consider a survival analysis of sheep or of dairy cows. The information available may consist of covariates such as herd or flock, sire, year-season of birth, molecular markers, and last known survival status, since censoring is pervasive. The objective of the analysis may be to assess the effects of explanatory variables, or to predict the survival time of the future progeny of some of the sires. Hence, one searches for some reasonable survival model (e.g., Gross and Clark, 1975; Collet, 1994) and finds that a proportional hazards model M_1 fits well and that it gives sensible parameter estimates. Then one proceeds to make predictions. However, another proportional hazards model M_2 also fits well, but it gives different estimates and predictions. Which model should be used at the end?

Now imagine a standard regression analysis in which 15 predictor variables are available, and suppose that some "best" model must be sought. Even if second-order and cross-product terms are ignored, there would be 2^{15} different models. For example, suppose that the variables are Y, X_1, X_2. Then, using the standard notation, there are the following four possible models

$$\begin{aligned}
\text{model 1} &: Y = \beta_0 + e, \\
\text{model 2} &: Y = \beta_0 + \beta_1 X_1 + e, \\
\text{model 3} &: Y = \beta_0 + \beta_2 X_2 + e, \\
\text{model 4} &: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e.
\end{aligned}$$

These models may differ little in relative plausibility. Again, which model ought to be used for predictions?

A third example is that of choosing between genetic models to infer parameters, and to predict the genetic merit of future progeny. One specification may be the classical infinitesimal model. A second specification may be a model with a finite number of loci. If so, how many? A third model may pose polygenic variation, plus the effects of some marked QTL.

The preceding three examples illustrate that the problem of model choice is pervasive. Typically, models are chosen in some ad-hoc manner, and inferences are based on the model eventually chosen, as if there were no uncertainty about it. In Bayesian analysis, however, it is possible to view the model as an item subject to uncertainty. Then the "model random variable" is treated as a nuisance, and the posterior distribution of the "model random variable" is used to obtain inferences that automatically take into account the relative plausibility of the models under consideration. This is called Bayesian model averaging, or BMA for short. We will outline the basic ideas, and refer the reader to Madigan and Raftery (1994), Raftery et al. (1997), and Hoeting et al. (1999) for additional details. These authors

argue as follows: since part of the evidence must be used in the process of model selection, ignoring the uncertainty about the model leads to an overstatement of precision in the analysis. In turn, this can lead to declaring "false positives", and the analysis lacks robustness unless, by chance, one stumbles into the "right" model. It will be shown at the end of this section that BMA can be used to enhance the predictive ability of an analysis.

8.6.2 Definitions

Let

$$\begin{aligned} \Delta &= \text{parameter or future data point,} \\ \mathbf{y} &= \text{data,} \\ M &= \{M_1, M_2, \ldots, M_K\} \text{ set of models,} \\ p(M_i) &= \text{prior probability of model } i, \\ p(M_i|\mathbf{y}) &= \text{posterior probability of model } i. \end{aligned}$$

The "usual" Bayesian approach gives, as posterior distribution (or density) of Δ,

$$p(\Delta|\mathbf{y}, M_i) = \frac{p(\mathbf{y}|\Delta, M_i)\, p(\Delta|M_i)}{p(\mathbf{y}|M_i)},$$

and the notation indicates clearly that inferences are conditional on M_i, as if the model were known to be true for sure. In BMA, on the other hand, the idea is to average out over the posterior distribution of the models, leading to

$$\begin{aligned} p(\Delta|\mathbf{y}) &= p(\Delta \text{ and } M_1|\mathbf{y}) + \cdots + p(\Delta \text{ and } M_K|\mathbf{y}) \\ &= \sum_{i=1}^{K} p(\Delta \text{ and } M_i|\mathbf{y}) = \sum_{i=1}^{K} p(\Delta|\mathbf{y}, M_i)\, p(M_i|\mathbf{y}). \end{aligned} \tag{8.58}$$

The preceding expression reveals that, in BMA, the model is treated as a nuisance parameter. Hence, the nuisance is eliminated in the usual manner, by integration or by summing. Then the inferences about a parameter can be viewed as a weighted average of the inferences that would be drawn if each of the models were true, using the posterior probability of the model as a mixing distribution.

In BMA, the posterior expectation and variance are calculated in the usual manner. For example, let the posterior mean of Δ under model k be

$$E(\Delta|M_k, \mathbf{y}) = \int \Delta\; p(\Delta|M_k, \mathbf{y})\, d\Delta = \widehat{\Delta}_k$$

Then, unconditionally with respect to the model, one obtains

$$E(\Delta|\mathbf{y}) = E_{M|\mathbf{y}}\left[E(\Delta|M_k, \mathbf{y})\right] = \sum_{k=1}^{K} \widehat{\Delta}_k\, p(M_k|\mathbf{y}). \tag{8.59}$$

Similarly, one can use the variance decomposition

$$Var\left(\Delta|\mathbf{y}\right) = E_{M|\mathbf{y}}\left[Var\left(\Delta|M,\mathbf{y}\right)\right] + Var_{M|\mathbf{y}}\left[E\left(\Delta|M,\mathbf{y}\right)\right],$$

leading to

$$Var\left(\Delta|\mathbf{y}\right) = \sum_{k=1}^{K} Var\left(\Delta|M_k,\mathbf{y}\right) p\left(M_k|\mathbf{y}\right) + \sum_{k=1}^{K} \left(\widehat{\Delta}_k\right)^2 p\left(M_k|\mathbf{y}\right) - \left[\sum_{k=1}^{K} \widehat{\Delta}_k \, p\left(M_k|\mathbf{y}\right)\right]^2. \tag{8.60}$$

The idea is straightforward, and it makes eminent sense, at least from a Bayesian perspective. The difficulty resides in that there can be many models, as in a regression equation, where there may be at least 2^p (for p being the number of covariates) models, and even more when interactions are included. Hoeting et al. (1999) discusses some of the methods that have been used for reducing the number of terms to be included in the sums appearing in (8.59) and (8.60).

8.6.3 Predictive Ability of BMA

Suppose one partitions the data as

$$\mathbf{y} = \left[\mathbf{y}'_{\text{Build}}, \mathbf{y}'_{\text{Pred}}\right]',$$

where $\mathbf{y}_{\text{Build}}$ is the data used for model building, and $\mathbf{y}_{\text{Pred}}$ includes the data points to be predicted, as in predictive cross-validation. Good (1952) introduced the predictive logscore (PLS) which, for Model k, is

$$PLS_k = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log p\left(y|M_k,\mathbf{y}_{\text{Build}}\right)$$

$$= -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log \int p\left(y|\boldsymbol{\theta}_k,M_k,\mathbf{y}_{\text{Build}}\right) p\left(\boldsymbol{\theta}_k|M_k,\mathbf{y}_{\text{Build}}\right) d\boldsymbol{\theta}_k, \tag{8.61}$$

where $\boldsymbol{\theta}_k$ is the parameter vector under Model k. It it desirable to have a model with as small a PLS as possible. Under BMA

$$PLS_{BMA} = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log\left[\sum_{k=1}^{K} p\left(y|M_k,\mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)\right]. \tag{8.62}$$

Suppose that the model and the data to be predicted are unknown, which is the usual situation. Now consider the difference

$$PLS_{BMA} - PLS_k = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log \frac{\sum_{k=1}^{K} p\left(y|M_k,\mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)}{p\left(y|M_k,\mathbf{y}_{\text{Build}}\right)}.$$

Next, take expectations of this difference with respect to the predictive distribution under BMA (that is, averaging over all possible models). This distribution has density

$$p\left(\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}\right) = \sum_{k=1}^{K} p\left(\mathbf{y}_{\text{Pred}}|M_k, \mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)].$$

Thus,

$$\begin{aligned} &E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_{BMA} - PLS_k\right) \\ &= -\sum_{y \in \mathbf{y}_{\text{Pred}}} E\left[\log \frac{\sum_{k=1}^{K} p\left(y|M_k, \mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)}{p\left(y|M_k, \mathbf{y}_{\text{Build}}\right)}\right]. \end{aligned}$$

The expected value in the right hand side, taken over the distribution $[\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}]$, is the Kullback–Leibler discrepancy between the predictive distributions of datum y under BMA and under Model k. Since the discrepancy is at least 0, it follows that the right-hand side is at most null. Hence

$$E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_{BMA}\right) \leq E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_k\right),$$

as in Madigan and Raftery (1994). This implies that under model uncertainty, the predictive performance of BMA (at least in the PLS sense) is expected to be better than that obtained under a single model, even if the latter is the most probable one. Raftery et al. (1997) and Hoeting et al. (1999) present several study cases supporting this theoretical result.

Typically, BMA leads to posterior distributions that are more spread than those under a single model. This illustrates that inferences based on a single model may give an unrealistic statement of precision; this may lead to false positive results.

The reader is now equipped with the foundations on which Bayesian inference rests. As stated before and especially for complex models, it is seldom the case that exact methods of inference can be used. Fortunately, methods for sampling from posterior distributions are available, and these are discussed in Part III of this book.

9

Approximate Inference Via the EM Algorithm

9.1 Introduction

The classical paradigm of maximum likelihood estimation is based on finding the supremum of the likelihood function (if it exists), and on attaching a measure of uncertainty via Fisher's information measure, which has an asymptotic justification. An overview of the classical first-order asymptotic ML theory was presented in Chapters 3 and 4. The Bayesian counterpart of this large sample theory consists of using an asymptotic approximation to the posterior distribution. The most commonly used approximation relies on computing the posterior mode and the observed information matrix. Computation of maximum likelihood estimates with the Newton–Raphson or scoring algorithms was dealt with in Chapter 4, but little has been said so far of how the calculations should proceed in the approximate Bayesian analysis.

In this chapter, an introductory account is given of one of the most versatile iterative algorithms for computing maximum likelihood and posterior modes: the expectation–maximization, or EM algorithm. This algorithm is conceptually simple, at least in its basic form, and brings considerable insight into the statistical structure of a maximum likelihood or posterior mode problem, contrary to Newton–Raphson or scoring, which are based primarily on numerical considerations. The chapter begins with a definition of the concepts of complete and incomplete data. The subsequent section presents a derivation of the algorithm in its basic form. Additional sections of the chapter discuss properties of the algorithm, the special form it takes

when applied to exponential families, and extensions that have been suggested for recovering measures of uncertainty. The chapter concludes with a set of examples. Many developments and extensions have become available since its introduction by Dempster et al. (1977). Several of these can be found in the comprehensive book of McLachlan and Krishnan (1997).

9.2 Complete and Incomplete Data

The EM algorithm was given its name in a celebrated paper by Dempster et al. (1977). As mentioned above, this is an iterative method for finding ML estimates or posterior modes in what are called incomplete-data problems. The influence of the EM algorithm has been far reaching, not only as a computational tool but as a way of solving difficult statistical problems. A main reason for this impact is because it is easy to implement. The basic idea behind the method is to transform an incomplete- into a complete-data problem for which the required maximization is computationally more tractable. Also, the algorithm is numerically stable: each iteration increases the likelihood or posterior density and convergence is nearly always to a local maximum.

The concept of missing data is fairly broad. It includes, for example, missing data in an unbalanced layout, but it extends to observations from truncated distributions, censored data, and latent variables. In these cases, one can view the complete data $\mathbf{x}$ as consisting of the vectors $(\mathbf{y}, \mathbf{z})$, where $\mathbf{y}$ is the observed data or incomplete data, and $\mathbf{z}$ is the missing data. More generally, many statistical problems which at first glance do not appear to involve missing data can be reformulated into missing-data problems, by judicious augmentation of the data set, with unobserved values. As such, one can view the observations at hand and the parameters of the posed model as data: part of these data is observed (the records) and another part is missing (the parameters). Mixed effects and hierarchical models, and models with latent variables, such as the threshold model, are typically amenable to an EM formulation. An example is an additive genetic model where inference may focus on $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \sigma_a^2, \sigma_e^2\right)'$, where $\boldsymbol{\beta}$ is a vector of location parameters and $\left(\sigma_a^2, \sigma_e^2\right)$ are variance components. Here, one may augment the observed data $\mathbf{y}$, with the missing data $\mathbf{a}$, the unobserved vector of additive genetic values. As shown later, this simplifies the computations involved in finding the ML estimates of $\boldsymbol{\theta}$, or the maximum of $p\left(\boldsymbol{\beta}, \sigma_a^2, \sigma_e^2|\mathbf{y}\right)$, or mode of the posterior distribution $\left[\boldsymbol{\beta}, \sigma_a^2, \sigma_e^2|\mathbf{y}\right]$. On the other hand, if one wishes to find the mode of the distribution with density $p\left(\sigma_a^2, \sigma_e^2|\mathbf{y}\right)$, an EM strategy is to consider $(\boldsymbol{\beta}, \mathbf{a})$ as the missing data. Here, if improper priors are adopted for the variance components and for the location vector $\boldsymbol{\beta}$, the mode is identical to the REML estimates of $\left(\sigma_a^2, \sigma_e^2\right)$. Another example is a regression model with t-distributed errors;

this can be formulated as a standard weighted least-squares problem where the missing data are related to the "weights".

9.3 The EM Algorithm

9.3.1 Form of the Algorithm

Suppose that the objective is to draw inferences about the $d \times 1$ vector $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ using the mode of $[\boldsymbol{\theta}|\mathbf{y}]$ as point estimator. We will use $p(\boldsymbol{\theta}|\mathbf{y})$ to denote a posterior density (or a likelihood function if flat priors are adopted for the parameters) where, as usual, $\mathbf{y}$ is the vector of observed data. Let $\mathbf{z}$ represent a vector of missing data, such as missing records or unobserved "parameters" of the model, and let its conditional, p.d.f. be $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$. The marginal posterior density of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\, d\mathbf{z},$$

where $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ is the joint posterior density of $\boldsymbol{\theta}$ and $\mathbf{z}$. The integration above typically leads to an expression which makes $p(\boldsymbol{\theta}|\mathbf{y})$ difficult to maximize, even though maximization of $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \propto p(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$ with respect to $\boldsymbol{\theta}$ may be trivial if $\mathbf{z}$ were observed. The EM algorithm formalizes an old idea for dealing with missing-data problems. Starting with a guessed value for the parameter $\boldsymbol{\theta}$, carry out the following iteration:

- Replace the missing data $\mathbf{z}$ by their expectation given the guessed value of the parameters and the observed data. Let this conditional expectation be $\widetilde{\mathbf{z}}$.
- Maximize $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$ replacing the missing data $\mathbf{z}$ by their expected values. This is equivalent to maximizing $p(\boldsymbol{\theta}|\widetilde{\mathbf{z}}, \mathbf{y})$.
- Reestimate the missing values $\mathbf{z}$ using their conditional expectation based on the updated $\boldsymbol{\theta}$.
- Reestimate $\boldsymbol{\theta}$ and continue until convergence is reached.

9.3.2 Derivation

Consider the identity

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})}.$$

Taking logarithms on both sides leads to

$$\ln p(\boldsymbol{\theta}|\mathbf{y}) = \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) - \ln p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}), \tag{9.1}$$

where the first term on the right-hand side is known as the complete-data log-likelihood (more generally, as complete-data log-posterior). Now take expectations of both sides with respect to $\left[\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right]$, where $\boldsymbol{\theta}^{[t]}$ is the current guess of $\boldsymbol{\theta}$. The left-hand side of (9.1) does not depend on $\mathbf{z}$, so averaging over $\mathbf{z}$, providing the integrals exist, gives

$$\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \int \ln p\left(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z} - \int \ln p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z}. \tag{9.2}$$

The first term on the right-hand side of (9.2) is a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$ and fixed $\boldsymbol{\theta}^{[t]}$, and it is denoted as $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$ in the EM literature. The second term is denoted $H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$. Thus,

$$\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right) = Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) - H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right). \tag{9.3}$$

The EM algorithm involves working with the first term only, $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$, disregarding $H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$. The two steps are:

1. E-step: calculation of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$, that is, the expectation of the complete data log-likelihood (log-posterior) with respect to the conditional distribution of the missing data, given the observed data and the current guess for $\boldsymbol{\theta}$.

2. M-step: maximization of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$ with respect to $\boldsymbol{\theta}$, solving for $\boldsymbol{\theta}$, and setting the result equal to $\boldsymbol{\theta}^{[t+1]}$, the new value of the parameter. Thus, if $\boldsymbol{\theta}^{[t+1]}$ maximizes $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$, the M-step is such that

$$Q\left(\boldsymbol{\theta}^{[t+1]}|\boldsymbol{\theta}^{[t]}\right) \geq Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right), \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Omega}, \tag{9.4}$$

which implies that $\boldsymbol{\theta}^{[t+1]}$ is a solution to the equation

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}} = 0. \tag{9.5}$$

The two steps are repeated iteratively until convergence is reached. It is shown below that this iterative sequence leads to a monotonic increase of $\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)$. That is,

$$\ln p\left(\boldsymbol{\theta}^{[t+1]}|\mathbf{y}\right) \geqslant \ln p\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right). \tag{9.6}$$

Since the marginal posterior density increases in each step, the EM algorithm, with few exceptions, converges to a local mode.

In many important applications, the E-step involves replacing the missing data by their conditional expectations. In some cases, computation of the E-step as formally dictated by step 1 above, can more easily disclose pathologies rendering the EM algorithm inapplicable. For instance, see Flury and Zoppe (2000) for a case where the EM is not applicable because the log-likelihood function takes the value zero in a subset of the parameter space, so the relevant conditional expectation is not defined.

In some models the calculation of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$ in the E-step may be difficult. Wei and Tanner (1990) propose a Monte Carlo approach for overcoming this difficulty. This consists of simulating $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m$ from $p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right)$ and then forming the simulation consistent estimator

$$\widehat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) \approx \frac{1}{m}\sum_{i=1}^{m} \ln p\left(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{y}\right).$$

9.4 Monotonic Increase of $\ln p(\boldsymbol{\theta}|\mathbf{y})$ with Each EM Iteration

Consider a sequence of iterates $\boldsymbol{\theta}^{[0]}, \boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[t+1]}$. The difference in value of $\ln p(\boldsymbol{\theta}|\mathbf{y})$ in successive iterates is obtained from (9.3) as

$$\begin{aligned}\ln p\left(\boldsymbol{\theta}^{[t+1]}|\mathbf{y}\right) - \ln p\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right) &= Q\left(\boldsymbol{\theta}^{[t+1]}|\boldsymbol{\theta}^{[t]}\right) - Q\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right) \\ &\quad - \left[H\left(\boldsymbol{\theta}^{[t+1]}|\boldsymbol{\theta}^{[t]}\right) - H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)\right]. \end{aligned} \tag{9.7}$$

The difference between $Q(\cdot)$ functions on the right-hand side is nonnegative due to (9.4). Therefore (9.6) holds if the difference in $H(\cdot)$ above is nonpositive; that is, if

$$H\left(\boldsymbol{\theta}^{[t+1]}|\boldsymbol{\theta}^{[t]}\right) - H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right) \leq 0. \tag{9.8}$$

Now for any $\boldsymbol{\theta}$,

$$\begin{aligned} H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) - H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right) &= \int \ln p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z} \\ &\quad - \int \ln p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{[t]}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z} \\ &= \int \ln\left[\frac{p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}\right)}{p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{[t]}\right)}\right] p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z} \\ &= -\int \ln\left[\frac{p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{[t]}\right)}{p\left(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}\right)}\right] p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{z}. \end{aligned} \tag{9.9}$$

The integral is the Kullback–Leibler distance between the distributions with densities $p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)$and $p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)$, which is at least 0. Hence, this integral preceded by a negative sign is at most 0. Thus, $H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) \leq H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)$. This establishes (9.8) and, hence, inequality (9.6). In well behaved problems the sequence of EM iterates converges to a stationary point which is a global maximum, in which case EM yields the unique posterior mode or ML estimate of $\boldsymbol{\theta}$, the maximizer of $\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)$.

Since

$$H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) \leq H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)$$

for all $\boldsymbol{\theta}$, this implies that $H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$ has a maximum at $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$. Hence,

$$\left.\frac{\partial H\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}} = \frac{\partial H\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{9.10}$$

Therefore from (9.3),

$$\left.\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}} = \left.\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}}. \tag{9.11}$$

9.5 The Missing Information Principle

9.5.1 Complete, Observed and Missing Information

Recall the identity

$$\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right) - \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right).$$

Differentiating twice with respect to $\boldsymbol{\theta}$ and multiplying by -1 gives

$$-\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} = -\frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} + \frac{\partial^2 \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}. \tag{9.12}$$

Note that the left-hand side is not a function of z. Now taking expectations of both sides over the distribution $[\mathbf{z}|\mathbf{y},\boldsymbol{\theta}]$ gives

$$\begin{aligned} -\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} = &-\int \frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right) d\mathbf{z} \\ &+ \int \frac{\partial^2 \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right) d\mathbf{z}. \end{aligned} \tag{9.13}$$

Recalling the definition of the Q and H functions given in (9.2), and provided that the integral and differential operations are interchangeable, one arrives at

$$-\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = -\frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} - \left[-\frac{\partial^2 H\left(\boldsymbol{\theta}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]. \tag{9.14}$$

If one calls the first and second terms on the right-hand side as "complete" and "missing" information, respectively, (9.14) has the following interpretation (Louis, 1982):

Observed information = complete information − missing information.

This can be represented as

$$\mathbf{I}\left(\boldsymbol{\theta}|\mathbf{y}\right) = \mathbf{I}_c\left(\boldsymbol{\theta}|\mathbf{y}\right) - \mathbf{I}_m\left(\boldsymbol{\theta}|\mathbf{y}\right). \tag{9.15}$$

The rate of convergence of the EM algorithm is related to these matrix-valued quantities: the larger the proportion of missing information (relative to the complete information), the slower the rate of convergence. This can be represented in the following manner (a formal justification for the expression is given later on):

$$\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^* = \left[\mathbf{I}_c\left(\boldsymbol{\theta}^*|\mathbf{y}\right)\right]^{-1} \mathbf{I}_m\left(\boldsymbol{\theta}^*|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t]} - \boldsymbol{\theta}^*\right). \tag{9.16}$$

The preceding indicates that as the EM iteration proceeds, the distance between the iterates and $\boldsymbol{\theta}^*$ (a stationary point of the likelihood function or posterior density) is a function of the rate of convergence matrix $\left[\mathbf{I}_c\left(\boldsymbol{\theta}^*|\mathbf{y}\right)\right]^{-1} \mathbf{I}_m\left(\boldsymbol{\theta}^*|\mathbf{y}\right)$. In short, the larger the proportion of missing information, the slower the algorithm proceeds towards $\boldsymbol{\theta}^*$, and the form of (9.16) suggests a linear approach towards $\boldsymbol{\theta}^*$. Note from (9.15) that the rate of convergence matrix can be written also as

$$\left[\mathbf{I}_c\left(\boldsymbol{\theta}^*|\mathbf{y}\right)\right]^{-1} \mathbf{I}_m\left(\boldsymbol{\theta}^*|\mathbf{y}\right) = \mathbf{I}_d - \left[\mathbf{I}_c\left(\boldsymbol{\theta}^*|\mathbf{y}\right)\right]^{-1} \mathbf{I}\left(\boldsymbol{\theta}^*|\mathbf{y}\right), \tag{9.17}$$

where $\mathbf{I}_d$ is the identity matrix of dimension d, the number of elements in $\boldsymbol{\theta}$.

9.5.2 Rate of Convergence of the EM Algorithm

A formal derivation of (9.16) is presented here. Following McLachlan and Krishnan (1997), consider a Taylor series expansion of the score vector

$$\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}$$

about the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$. This yields

$$\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} \approx \frac{\partial \ln p\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} - \mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{[t]}\right).$$

Setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the term on the left-hand side vanishes and one obtains after rearrangement

$$\boldsymbol{\theta}^* \approx \boldsymbol{\theta}^{[t]} + \left[\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1} \frac{\partial \ln p\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}. \tag{9.18}$$

Now expand $\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)/\partial \boldsymbol{\theta}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t+1]}}$ in a linear Taylor series about $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$:

$$\frac{\partial Q\left(\boldsymbol{\theta}^{[t+1]}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}} \approx \frac{\partial Q\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 Q\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}\right).$$

From (9.5), the term on the left-hand side is equal to zero. Making use of this in the preceding expression and employing the notation in (9.15), this can be written as

$$\frac{\partial Q\left(\boldsymbol{\theta}^{[t]}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}} \approx \mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}\right)$$

which, from (9.11), is

$$\frac{\partial \ln p\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} \approx \mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}\right). \tag{9.19}$$

Substituting approximation (9.19) into (9.18) yields

$$\begin{aligned}\boldsymbol{\theta}^* - \boldsymbol{\theta}^{[t]} &\approx \left[\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}\right)\\ &= \left[\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \boldsymbol{\theta}^{[t]}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\{\mathbf{I}_d - \left[\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right\}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{[t]}\right)\\ &\quad\approx \left[\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^*\right).\end{aligned}$$

Premultiplying both sides by $\left[\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)$ yields

$$\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^* \approx \left\{\mathbf{I}_d - \left[\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right\}\left(\boldsymbol{\theta}^{[t]} - \boldsymbol{\theta}^*\right).$$

In view of (9.17), this is expressible as

$$\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^* \approx \left[\mathbf{I}_c\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_m\left(\boldsymbol{\theta}^{[t]}|\mathbf{y}\right)\left(\boldsymbol{\theta}^{[t]} - \boldsymbol{\theta}^*\right),$$

and for $\boldsymbol{\theta}^{[t]}$ close to $\boldsymbol{\theta}^*$ this leads to the desired result (9.16) directly.

9.6 EM Theory for Exponential Families

The EM algorithm has a simpler form when the complete data $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$ have a distribution from the regular exponential family. The density can be written in its canonical form as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{b(\mathbf{x}) \exp\left[\boldsymbol{\theta}' \mathbf{t}(\mathbf{x})\right]}{a(\boldsymbol{\theta})}. \tag{9.20}$$

In this expression, the vector $\boldsymbol{\theta}$ is the natural or canonical parameter indexing the distribution, $b(\mathbf{x})$ is a function of the complete data alone, $a(\boldsymbol{\theta})$ is a function of the vector $\boldsymbol{\theta}$ alone, and $\mathbf{t}(\mathbf{x})$ is the $d \times 1$ vector of complete-data sufficient statistics. Many common distributions can be put in the form (9.20), which is characterized by a number of nice statistical properties. In an exponential family, the statistic $\mathbf{t}(\mathbf{x})$ carries all the information about $\boldsymbol{\theta}$ contained in the data $\mathbf{x}$; therefore, inferences about $\boldsymbol{\theta}$ can be based solely on $\mathbf{t}(\mathbf{x})$. In a Bayesian context this means that the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ is identical to the posterior distribution $[\boldsymbol{\theta}|\mathbf{y}]$. The ML or posterior mode equations (the score) take a particularly simple form. To see this, take logarithms on both sides of (9.20):

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \ln b(\mathbf{x}) + \boldsymbol{\theta}' \mathbf{t}(\mathbf{x}) - \ln a(\boldsymbol{\theta}). \tag{9.21}$$

In the context of ML estimation, the score is

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{x}) - \frac{\partial \ln a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{9.22}$$

Taking expectations of both sides over the sampling model for the complete data with density $p(\mathbf{x}|\boldsymbol{\theta})$, and recalling from ML theory that

$$E\left[\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

leads to

$$\begin{aligned} \frac{1}{a(\boldsymbol{\theta})} \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \int \mathbf{t}(\mathbf{x})\, p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \\ &= E\left[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}\right]. \end{aligned} \tag{9.23}$$

This is the expected value of the vector of sufficient statistics, given $\boldsymbol{\theta}$. Now from (9.22), and in the context of likelihood-based inference, the ML estimator of $\boldsymbol{\theta}$ is the solution to the equations

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{x}) - E\left[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}\right] = 0,$$

or, equivalently,

$$\mathbf{t}(\mathbf{x}) = E\left[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}\right]. \tag{9.24}$$

If equations (9.24) can be solved for $\boldsymbol{\theta}$, then the solution is unique due to the convexity property of the log-likelihood for regular exponential families. When the equations are not solvable, the maximizer of $\boldsymbol{\theta}$ lies on the boundary of the parameter space (McLachlan and Krishnan, 1997).

The regular exponential family of distributions also leads to a simple representation of the complete-data information matrix. Note from (9.21), that the second derivatives of the complete-data log-likelihood do not depend on $\mathbf{x}$, and that

$$-\frac{\partial^2 \ln p\left(\mathbf{x}|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} = \frac{\partial^2 \ln a\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} = \mathbf{I}_c\left(\boldsymbol{\theta}\right), \tag{9.25}$$

where $\mathbf{I}_c\left(\boldsymbol{\theta}\right)$ is Fisher's expected information matrix.

Return now to the computation of EM to obtain the ML estimator of $\boldsymbol{\theta}$ when the complete data can be written in the form of (9.20). The E-step is given by

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) &= \int \ln p\left(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right) d\mathbf{z} \\ = \int \ln b\left(\mathbf{x}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right) d\mathbf{z} &+ \boldsymbol{\theta}' \int \mathbf{t}\left(\mathbf{x}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right) d\mathbf{z} - \ln a\left(\boldsymbol{\theta}\right). \end{aligned} \tag{9.26}$$

The M-step consists of differentiating this expression with respect to $\boldsymbol{\theta}$. Since the first term does not depend on $\boldsymbol{\theta}$,

$$\begin{aligned} \frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\boldsymbol{\theta}} &= \int \mathbf{t}\left(\mathbf{x}\right) p\left(\mathbf{z}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right) d\mathbf{z} - \frac{1}{a\left(\boldsymbol{\theta}\right)} \frac{\partial a\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} \\ &= E\left[\mathbf{t}\left(\mathbf{x}\right)|\boldsymbol{\theta}^{[t]},\mathbf{y}\right] - E\left[\mathbf{t}\left(\mathbf{x}|\boldsymbol{\theta}\right)\right], \end{aligned} \tag{9.27}$$

where the first term on the right-hand side is the conditional expected value of the vector of sufficient statistics, given the observed data and the current value of $\boldsymbol{\theta}$. It follows from (9.27) that in the M-step, $\boldsymbol{\theta}^{[t+1]}$ is chosen by solving the equations

$$E\left[\mathbf{t}\left(\mathbf{x}\right)|\boldsymbol{\theta}^{[t]},\mathbf{y}\right] = E\left[\mathbf{t}\left(\mathbf{x}|\boldsymbol{\theta}\right)\right], \tag{9.28}$$

so the form of the algorithm is quite simple.

9.7 Standard Errors and Posterior Standard Deviations

One of the early criticisms of the EM approach was that, unlike the Newton–Raphson and related methods, it does not automatically produce an estimate of the asymptotic covariance matrix of the ML estimators of $\boldsymbol{\theta}$ or some

indication of the uncertainty in the posterior distribution of a parameter of interest. A number of ways of computing estimates of the asymptotic covariance matrix of the ML estimates have been suggested over the last few years. Important contributions include those of Louis (1982), Meilijson (1989), Meng and Rubin (1991), Lange (1995), and Oakes (1999). All methods make use of asymptotic theory. Three of these approaches will be described here.

9.7.1 The Method of Louis

Louis (1982) showed that

$$\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[\frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right] + Var_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[\frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}\right], \tag{9.29}$$

which, on multiplication by -1, yields (9.15). Thus,

$$\mathbf{I}_c\left(\boldsymbol{\theta}|\mathbf{y}\right) = E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[-\frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \tag{9.30}$$

and

$$\mathbf{I}_m\left(\boldsymbol{\theta}|\mathbf{y}\right) = Var_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[\frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}\right]. \tag{9.31}$$

To prove (9.29), first note that

$$\begin{aligned}
\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} &= \frac{1}{p\left(\boldsymbol{\theta}|\mathbf{y}\right)} \frac{\partial}{\partial \boldsymbol{\theta}} \int p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right) d\mathbf{z} \\
&= \frac{1}{p\left(\boldsymbol{\theta}|\mathbf{y}\right)} \int \frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right) d\mathbf{z} \\
&= \int \frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right) d\mathbf{z} \\
&= E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[\frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}}\right].
\end{aligned} \tag{9.32}$$

Now, take derivatives with respect to $\boldsymbol{\theta}$ again

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}'}\right] &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\int \frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right) d\mathbf{z}\right] \\
= \int \frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right) d\mathbf{z} &+ \int \frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} \frac{\partial p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right)}{\partial \boldsymbol{\theta}'} d\mathbf{z} \\
&= E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}}\left[\frac{\partial^2 \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \\
&+ \int \frac{\partial \ln p\left(\boldsymbol{\theta},\mathbf{z}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}} \frac{\partial \ln p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right)}{\partial \boldsymbol{\theta}'} p\left(\mathbf{z}|\boldsymbol{\theta},\mathbf{y}\right) d\mathbf{z}.
\end{aligned} \tag{9.33}$$

The second term in (9.33) can be manipulated as follows

$$\begin{aligned}
&\int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})}{\partial \boldsymbol{\theta}'} p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z} \\
&= \int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} - \frac{\partial \ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z} \\
&= \int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z} \\
&- \int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}'} p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z} \\
&= E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] \\
&- \int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}'} p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z}.
\end{aligned}$$

In view of (9.32), the last two lines of the preceding expression can be written as

$$\begin{aligned}
&E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] \\
&- \int \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{z} \\
&= E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] \\
&- E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \right] E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}'} \right] \\
&= Var_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \right].
\end{aligned}$$

Hence, (9.33) becomes

$$\frac{\partial^2 \ln p(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = E_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial^2 \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] + Var_{\mathbf{z}|\boldsymbol{\theta},\mathbf{y}} \left[\frac{\partial \ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})}{\partial \boldsymbol{\theta}} \right],$$

thus establishing (9.29), and the matrix of second derivatives from which a measure of uncertainty can be derived. Tanner (1996) suggests a Monte Carlo approximation to (9.29) that can be used when the integrals over $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ are difficult to obtain analytically.

9.7.2 Supplemented EM Algorithm (SEM)

This method was proposed by Meng and Rubin (1991) who showed that the asymptotic covariance matrix of the ML of $\boldsymbol{\theta}$ (evaluated at $\widehat{\boldsymbol{\theta}}$, the ML

estimator) is equal to

$$\left[\mathbf{I}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} = \left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} + \left[\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}. \tag{9.34}$$

Here, the term $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right) = \left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_m\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$ is the rate of convergence matrix and, as before, $\mathbf{I}_d$ is an identity matrix of dimension $d \times d$. The first term on the right-hand side of (9.34) is the asymptotic covariance matrix based on the complete-data log-likelihood and averaged over the distribution $\left[\mathbf{z}|\widehat{\boldsymbol{\theta}}, \mathbf{y}\right]$

$$\begin{aligned}
\left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} &= \left[E\left(-\frac{\partial^2 \ln p\left(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\middle|\boldsymbol{\theta}, \mathbf{y}\right)\middle|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]^{-1} \\
&= \left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\int \ln p\left(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}\right) p\left(\mathbf{z}|\widehat{\boldsymbol{\theta}}, \mathbf{y}\right) d\mathbf{z}\middle|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]^{-1} \\
&= \left[-\frac{\partial^2 Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\middle|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]^{-1}.
\end{aligned}$$

Often, this can be computed analytically, and it is simple to calculate when the density of the complete-data distribution is in the exponential family form (9.20), as indicated in (9.25).

The derivation of (9.34) is as follows. From (9.15),

$$\begin{aligned}
\mathbf{I}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right) &= \mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\left[\mathbf{I}_d - \left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}\mathbf{I}_m\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right] \\
&= \mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\left[\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right].
\end{aligned}$$

Inverting this expression establishes (9.34)

$$\begin{aligned}
\left[\mathbf{I}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} &= \left[\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}\left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} \\
&= \left\{\mathbf{I}_d + \left[\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right\}\left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} \\
&= \left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1} + \left[\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\left[\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right]^{-1}.
\end{aligned}$$

The second line arises from the matrix algebra result

$$\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1},$$

after setting $\mathbf{A} = \mathbf{B} = \mathbf{D} = \mathbf{I}_d$ and $\mathbf{C} = \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$.

In order to compute the rate of convergence matrix, $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$, Meng and Rubin (1991) suggest the following approach:

- Run the EM algorithm until convergence, and find

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_d\right)',$$

the maximizer of $\ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)$.

- Choose a starting point for $\boldsymbol{\theta}$ different from $\widehat{\boldsymbol{\theta}}$ in all components. A possible starting point is the one used for the original EM calculation. Run the EM algorithm through t iterations.

1. INPUT: $\boldsymbol{\theta}^{[t]}$ and $\widehat{\boldsymbol{\theta}}$. Run the usual E- and M-steps to obtain $\boldsymbol{\theta}^{[t+1]}$. Repeat the following steps (a) and (b) for $i = 1, 2, \ldots, d$, to obtain a matrix $\mathbf{R}^{[t]}$ with element $r_{ij}^{[t]}$ defined below.

   (a) Construct

$$\boldsymbol{\theta}^{[t]}(i) = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_{i-1}, \theta_i^{[t]}, \widehat{\theta}_{i+1}, \ldots, \widehat{\theta}_d\right).$$

   (b) Use $\boldsymbol{\theta}^{[t]}(i)$ as the input value for one EM step to obtain $\boldsymbol{\theta}^{[t+1]}(i)$ with elements $\theta_j^{[t+1]}(i)$, $(j = 1, 2, \ldots, d)$. The ith row of $\mathbf{R}^{[t]}$ is obtained as

$$r_{ij}^{[t]} = \frac{\theta_j^{[t+1]}(i) - \widehat{\theta}_j}{\theta_i^{[t]} - \widehat{\theta}_i}, \quad j = 1, 2, \ldots, d. \tag{9.35}$$

2. OUTPUT: $\boldsymbol{\theta}^{[t+1]}$ and $\left\{r_{ij}^{[t]}, \quad (i, j = 1, 2, \ldots, d)\right\}$. Set $t = t + 1$ and GO TO 1.

When the value of an element r_{ij} no longer changes between successive iterates, this represents a numerical estimate of the corresponding element in $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$. It is possible that different values of t may be required for different r_{ij} components. Meng and Rubin (1991) discuss many relevant implementation issues. The reader is referred to their paper for details. The numerical estimate of $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$ and the expression derived analytically for $\mathbf{I}_c\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)$ are then used in expression (9.34) to obtain the desired asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$.

9.7.3 The Method of Oakes

Oakes (1999) provided an expression for the observed information matrix based on the second derivatives of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$, the conditional expectation of the complete-data log-likelihood given the observed data and a fixed value of $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$. Oakes' formula is

$$\begin{aligned}\mathbf{I}\left(\boldsymbol{\theta}|\mathbf{y}\right) &= -\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} \\ &= -\left[\left(\frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} + \frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'^{[t]}}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}}\right]. \quad (9.36)\end{aligned}$$

The second term on the right-hand side corresponds to the information from the missing data, and is equal to $\mathbf{I}_m\left(\boldsymbol{\theta}|\mathbf{y}\right)$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$.

In order to derive (9.36), results (3.20) and (3.26) from Chapter 3 will be used. These results are valid for any $\boldsymbol{\theta}$, and therefore

$$\begin{aligned}&\int \frac{\partial}{\partial\boldsymbol{\theta}^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right) p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right) d\mathbf{z} \\ &= E_{z|y,\theta^{[t]}}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right] = \mathbf{0}, \quad (9.37)\end{aligned}$$

and

$$\begin{aligned}&E_{z|y,\theta^{[t]}}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right)'\right] \\ &= -E_{z|y,\theta^{[t]}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^{[t]}\partial\boldsymbol{\theta}'^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right]. \quad (9.38)\end{aligned}$$

In the development that follows, both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{[t]}$ are regarded as arguments of Q. To prove (9.36), first differentiate (9.3) with respect to $\boldsymbol{\theta}$; this gives

$$\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial\boldsymbol{\theta}} = \frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\boldsymbol{\theta}} - \int \frac{\partial \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right) d\mathbf{z}. \quad (9.39)$$

Setting $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$, the second term vanishes because of (9.37), and the score for the observed data becomes

$$\frac{\partial \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial\boldsymbol{\theta}} = \frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}}. \quad (9.40)$$

Take now partial derivatives of (9.39) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{[t]}$. First note that the left hand side of (9.39) is not a function of $\boldsymbol{\theta}^{[t]}$, and that one

may write

$$\frac{\partial \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\left(\frac{\partial p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}^{[t]}}\right)' = \frac{\partial \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}^{[t]}}\right)' p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right).$$

Then differentiating (9.39 first with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned}\frac{\partial^2 \ln p\left(\boldsymbol{\theta}|\mathbf{y}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} &= \frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} - \int \frac{\partial^2 \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right) d\mathbf{z} \\ &= \frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} - E_{z|y,\theta^{[t]}}\left[\frac{\partial^2 \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\right],\end{aligned}$$

and, second, with respect to $\boldsymbol{\theta}^{[t]}$, yields

$$\begin{aligned}\mathbf{0} =& \frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'^{[t]}} - \int \frac{\partial \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\left(\frac{\partial p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}^{[t]}}\right)' d\mathbf{z} \\ &= \frac{\partial^2 Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'^{[t]}} \\ &- E_{z|y,\theta^{[t]}}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}^{[t]}} \ln p\left(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{[t]}\right)\right)'\right].\end{aligned}$$

Substituting $\boldsymbol{\theta} = \boldsymbol{\theta}^{[t]}$, adding, using identity (9.38), and multiplying both sides by -1, retrieves (9.36).

As Oakes (1999) points out, (9.36) together with (9.40), implies that the function $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)$ can be used to perform standard Newton–Raphson maximization of the observed data likelihood.

9.8 Examples

Illustrations of the EM-related computations are presented first with two examples involving discrete data and a multinomial sampling model. The third and fourth examples deal with inferences about variance components in a Gaussian hierarchical model.

Example 9.1 *A multinomial example*
Example 4.8 from Chapter 4 illustrated estimation of the recombination

Class	Genotypic class	Frequency
1	AA/BB	$\alpha^2/4$
2	Aa/Bb	$(\alpha^2 + (1-\alpha)^2)/2$
3	AA/Bb	$\alpha(1-\alpha)/2$
4	Aa/BB	$\alpha(1-\alpha)/2$
5	aa/bb	$\alpha^2/4$
6	Aa/bb	$\alpha(1-\alpha)/2$
7	aa/Bb	$\alpha(1-\alpha)/2$
8	AA/bb	$(1-\alpha)^2/4$
9	aa/BB	$(1-\alpha)^2/4$

TABLE 9.1. Genotypic classes and frequencies from a mating between repulsion heterozygotes.

fraction α between loci A and B using coupling heterozygotes. Here, α is estimated with data from repulsion heterozygotes. The gametes produced are AB, ab, Ab, and aB with respective frequencies $\alpha/2$, $\alpha/2$, $(1-\alpha)/2$, and $(1-\alpha)/2$. Random union of these gametes produces the genotypic classes shown in Table 9.1.

Assuming complete dominance at both loci, the phenotypic classes which are distinguishable (the observed data) are shown in Table 9.2, together with the number of observations for each phenotype and their expected frequencies. For example, the frequency of phenotype AB is obtained summing genotypic classes 1, 2, 3, and 4.

Maximum Likelihood Estimation

Denoting $\theta = \alpha^2$ and $\mathbf{n} = (n_1, n_2, n_3, n_4)'$, the observed data likelihood is in the form of a multinomial distribution

$$p(\theta|\mathbf{n}) \propto \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{n_1} \left(\frac{1}{4} - \frac{1}{4}\theta\right)^{n_3} \left(\frac{1}{4} - \frac{1}{4}\theta\right)^{n_4} \left(\frac{1}{4}\theta\right)^{n_2}$$

and the observed data log-likelihood, excluding an additive constant, is

$$\ln p(\theta|\mathbf{n}) = n_1 \ln(2+\theta) + (n_3 + n_4)\ln(1-\theta) + n_2 \ln\theta.$$

The score is equal to

$$\frac{\partial \ln p(\theta|\mathbf{n})}{\partial\theta} = \frac{n_1}{2+\theta} - \frac{n_3+n_4}{1-\theta} + \frac{n_2}{\theta}.$$

Phenotype	Frequency	Number observed
AB	$1/2 + \alpha^2/4$	n_1
ab	$\alpha^2/4$	n_2
aB	$1/4 - \alpha^2/4$	n_3
Ab	$1/4 - \alpha^2/4$	n_4

TABLE 9.2. Distribution of phenotypic classes with complete dominance.

The observed information is

$$-\frac{\partial^2 \ln p\left(\theta|\mathbf{n}\right)}{\partial\theta^2} = \frac{n_1}{\left(2+\theta\right)^2} + \frac{n_3+n_4}{\left(1-\theta\right)^2} + \frac{n_2}{\theta^2}. \tag{9.41}$$

Upon setting the score to 0, one obtains a quadratic equation in θ with one positive root, which is the ML estimator of θ. For example, for $\mathbf{n} = (20, 12, 120, 117)'$, one obtains $\widehat{\theta} = 0.050056$ and $\widehat{\alpha} = \sqrt{\widehat{\theta}} = 0.22$. The observed information evaluated at $\widehat{\theta}$ is

$$I\left(\widehat{\theta}|\mathbf{n}\right) = 5056.6997. \tag{9.42}$$

The estimate of the asymptotic variance of $\widehat{\theta}$ based on the observed information is:

$$\widehat{Var\left(\widehat{\theta}\right)} = \left[-\frac{\partial^2 \ln p\left(\theta|\mathbf{n}\right)}{\partial\theta^2}\bigg|_{\theta=\widehat{\theta}}\right]^{-1} = 0.000198, \tag{9.43}$$

and the estimate of the asymptotic variance of $\widehat{\alpha}$ is

$$\widehat{Var\left(\widehat{\alpha}\right)} = \widehat{Var\left(\widehat{\theta}\right)}\left[\left(\frac{d\alpha}{d\theta}\right)^2\bigg|_{\theta=\widehat{\theta}}\right] = 0.000986.$$

Note that the transformation $\theta = \alpha^2$ is not one-to-one; therefore the expression above for $\widehat{Var\left(\widehat{\alpha}\right)}$ can be misleading. However, it can be verified that in this case, parameterizing the likelihood in terms of α rather than in terms of θ, yields and estimate of the asymptotic $Var\left(\widehat{\alpha}\right)$ based on the inverse of the observed information equal to 0.000988.

Computation via the EM Algorithm

The ML estimator of θ is obtained now using the EM algorithm. The choice of missing data must be based on the form of the resulting complete-data likelihood. Suppose that we split the first of the four multinomial cells n_1, with associated frequency $\left(1/2 + \alpha^2/4\right)$ into two parts, n_{11} and n_{12}, with associated frequencies $1/2$ and $\alpha^2/4$, respectively. Thus, $n_1 = n_{11}+n_{12}$ and the missing data is $\mathbf{z} = (n_{11}, n_{12})'$. The complete-data likelihood is now

$$p\left(\theta, \mathbf{z}|\mathbf{n}\right) \propto \left(\frac{1}{2}\right)^{n_{11}}\left(\frac{1}{4}\theta\right)^{n_{12}}\left(\frac{1}{4}-\frac{1}{4}\theta\right)^{n_3}\left(\frac{1}{4}-\frac{1}{4}\theta\right)^{n_4}\left(\frac{1}{4}\theta\right)^{n_2}.$$

The complete-data log-likelihood, excluding an additive constant, is

$$\ln p\left(\theta, \mathbf{z}|\mathbf{n}\right) = \left(n_{12}+n_2\right)\ln\theta + \left(n_3+n_4\right)\ln\left(1-\theta\right),$$

which is the log-likelihood of the binomial probability model

$$Bi\left(n_{12}+n_2|\theta, n_{12}+n_2+n_3+n_4\right).$$

The score is

$$\frac{\partial \ln p\left(\theta, \mathbf{z}|\mathbf{n}\right)}{\partial \theta} = \frac{n_{12}+n_2}{\theta} - \frac{n_3+n_4}{1-\theta},$$

which is linear in θ. If n_{12} were observed, the ML estimator of θ is obtained by setting the score equal to 0. The explicit solution is

$$\widehat{\theta} = \frac{n_{12}+n_2}{n_{12}+n_2+n_3+n_4}. \tag{9.44}$$

Of course, $\widehat{\theta}$ cannot be obtained from (9.44) because n_{12} is not observed. Instead, n_{12} is replaced in (9.44) by its conditional expectation, given the observed data $\mathbf{n}$ and $\theta^{[t]}$. The E-step of the EM algorithm consists of evaluating

$$\begin{aligned} Q\left(\theta|\theta^{[t]}\right) &= E\left\{\left[(n_{12}+n_2)\ln\theta + (n_3+n_4)\ln(1-\theta)\right]|\,\mathbf{n}, \theta^{[t]}\right\} \\ &= \left[E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right) + n_2\right]\ln\theta + (n_3+n_4)\ln(1-\theta). \end{aligned} \tag{9.45}$$

The equality in the second line arises because, conditionally on $\mathbf{n}$, the only random variable in (9.45) is n_{12}. The M-step consists of choosing $\theta^{[t+1]}$ as the solution to the equation

$$\frac{\partial Q\left(\theta|\theta^{[t]}\right)}{\partial \theta} = 0.$$

This yields

$$\theta^{[t+1]} = \frac{E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right) + n_2}{E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right) + n_2 + n_3 + n_4}, \tag{9.46}$$

which has the same form as (9.44), with n_{12} replaced by its conditional expectation, $E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right)$. The EM scheme calculates $E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right)$ and (9.46) in an iterative fashion, until convergence is reached.

In order to obtain $E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right) = E\left(n_{12}|n_{11}+n_{12}, \theta^{[t]}\right)$, recall from Example 2.15 in Chapter 2 that

$$\begin{aligned} E\left(n_{12}|n_{11}+n_{12}, \theta^{[t]}\right) &= (n_{11}+n_{12})\,\frac{\theta^{[t]}/4}{\frac{1}{2}+\frac{\theta^{[t]}}{4}} \\ &= n_1\frac{\theta^{[t]}}{2+\theta^{[t]}}, \end{aligned}$$

because

$$n_{12}|\theta, n_1 \sim Bi\left(\frac{\theta/4}{\frac{1}{2}+\frac{\theta}{4}}, n_1\right). \tag{9.47}$$

Iteration	$\theta^{[t]}$	$r^{[t]}$
1	0.063241	0.03598
2	0.050530	0.03623
3	0.050073	0.03624
4	0.050056	0.03624
⋮	⋮	·
10	0.050056	·

TABLE 9.3.

Using $\theta^{[0]} = 0.5$ as the starting value, the results of the iterative EM sequence are shown in Table 9.3. After a few iterations the algorithm converges to $\widehat{\theta} = 0.050056$.

A Monte Carlo EM algorithm along the lines suggested by Wei and Tanner (1990) can be implemented by replacing $E\left(n_{12}|\mathbf{n}, \theta^{[t]}\right)$ in (9.46) with

$$\overline{n}_{12} = \frac{1}{m}\sum_{i=1}^{m} z_i$$

where $z_1, z_2, \ldots, z_m$ are draws from (9.47).

Rate of Convergence and Standard Errors

The third column of Table 9.3 shows the evolution of the rate of convergence given by (9.35), which for this one-dimensional example is

$$r^{[t]} = \frac{\theta^{[t+1]} - \widehat{\theta}}{\theta^{[t]} - \widehat{\theta}}.$$

The rate of convergence of the EM iteration is 0.03624.

Now we calculate $I_c\left(\widehat{\theta}|\mathbf{n}\right)$ and $I_m\left(\widehat{\theta}|\mathbf{n}\right)$.

$$I_c\left(\widehat{\theta}|\mathbf{n}\right) = -\frac{\partial^2 Q\left(\theta|\widehat{\theta}\right)}{(\partial\theta)^2}\bigg|_{\theta=\widehat{\theta}}.$$

Differentiating (9.45) twice with respect to θ gives

$$\begin{aligned} -\frac{\partial^2 Q\left(\theta|\widehat{\theta}\right)}{(\partial\theta)^2} &= \frac{E\left(n_{12}|\mathbf{n}, \widehat{\theta}\right) + n_2}{\theta^2} + \frac{n_3 + n_4}{(1-\theta)^2} \\ &= \frac{n_1 \frac{\widehat{\theta}}{2+\widehat{\theta}} + n_2}{\theta^2} + \frac{n_3 + n_4}{(1-\theta)^2}. \end{aligned} \tag{9.48}$$

Evaluated at $\theta = \widehat{\theta}$, (9.48) is equal to

$$I_c\left(\widehat{\theta}|\mathbf{n}\right) = 5246.84. \tag{9.49}$$

To calculate $I_m\left(\widehat{\theta}|n\right)$, use is made of (9.31). This produces

$$\begin{aligned} &Var\left[\frac{\partial}{\partial\theta}\left(n_{12}+n_2\right)\ln\theta + \left(n_3+n_4\right)\ln\left(1-\theta\right)\middle|\,\mathbf{n},\theta\right] \\ &= Var\left[\frac{\partial}{\partial\theta}\left(n_{12}\ln\theta\right)|\,\mathbf{n},\theta\right] = Var\left[\frac{n_{12}}{\theta}\middle|\,\mathbf{n},\theta\right] \\ &= \frac{2n_1}{\theta\left(2+\theta\right)^2}, \end{aligned}$$

where the last line follows from the variance of the binomial distribution (9.47). When evaluated at $\theta = \widehat{\theta}$, the expression above is equal to

$$I_m\left(\widehat{\theta}|\mathbf{n}\right) = 190.14. \tag{9.50}$$

From (9.42), (9.49), and (9.50), (9.15) can be verified:

$$5056.70 = 5246.84 - 190.14.$$

Second, the rate of convergence $\Delta\left(\widehat{\theta}|\mathbf{n}\right)$ is

$$\left[I_c\left(\widehat{\theta}|\mathbf{n}\right)\right]^{-1} I_m\left(\widehat{\theta}|\mathbf{n}\right) = \frac{190.14}{5246.84} = 0.036238,$$

as obtained in the third column of Table 9.3.

The asymptotic variance based on (9.34) is given by

$$\begin{aligned} \left[I\left(\widehat{\theta}|\mathbf{n}\right)\right]^{-1} &= \left(5246.84\right)^{-1} + \\ \left(1-0.03624\right)^{-1}&\left(0.03624\right)\left(5246.84\right)^{-1} \\ &= 0.0001978, \end{aligned}$$

in agreement with (9.43).

In order to obtain the asymptotic variance of $\widehat{\theta}$ based on (9.36), first compute

$$\frac{\partial^2 Q\left(\theta|\theta^{[t]}\right)}{\partial\theta\partial\theta^{[t]}} = \frac{2n_1}{\theta\left(2+\theta^{[t]}\right)^2}, \tag{9.51}$$

and, from (9.48),

$$\frac{\partial^2 Q\left(\theta|\theta^{[t]}\right)}{\left(\partial\theta\right)^2} = -\frac{n_1\frac{\theta}{2+\theta}+n_2}{\theta^2} - \frac{n_3+n_4}{\left(1-\theta\right)^2}. \tag{9.52}$$

When $\theta^{[t]}$ and θ are evaluated at $\widehat{\theta}$, the sum of these two expressions times -1 is equal to:

$$-\left[\frac{\partial^2 Q\left(\theta|\theta^{[t]}\right)}{(\partial\theta)^2}+\frac{\partial^2 Q\left(\theta|\theta^{[t]}\right)}{\partial\theta\partial\theta^{[t]}}\bigg|_{\theta^{[t]},\theta=\widehat{\theta}}\right] = -(-5246.84+190.14)$$
$$= 5056.70,$$

which agrees with (9.42) and (9.41). ■

Example 9.2 *Blood groups*
In Example 4.7 from Chapter 4, the ML estimates of the frequency of blood group alleles were computed via Newton–Raphson. Here the EM algorithm is used instead. Let $\mathbf{n} = (n_A, n_{AB}, n_B, n_O)'$ be the observed data, with $n_A = 725$, $n_{AB} = 72$, $n_B = 258$, $n_O = 1073$. It is sensible to treat the unobserved counts n_{AO}, n_{AA}, n_{BB} and n_{BO} as missing data. The resulting complete-data vector is

$$\mathbf{n}_c = (n_{AA}, n_{AO}, n_{AB}, n_{BB}, n_{BO}, n_O)'.$$

The complete-data log-likelihood excluding an additive constant is

$$\ln f(p_A, p_B|\mathbf{n}_c) = 2n_{AA}\ln(p_A) + n_{AO}\ln(2p_Ap_O) + n_{AB}\ln(2p_Ap_B) + 2n_{BB}\ln(p_B) + n_{BO}\ln(2p_Bp_O) + 2n_O\ln(p_O),$$

where $p_O = (1 - p_A - p_B)$. The E-step consists of computing the expected value of the complete-data log-likelihood, conditionally on the observed counts $\mathbf{n}$ and on the value of the parameters at iteration t, $\left(p_A^{[t]}, p_B^{[t]}\right)$. Explicitly, this is

$$\begin{aligned} Q\left(p_A,p_B|p_A^{[t]},p_B^{[t]}\right) &= E\left[\{2n_{AA}\ln(p_A) + n_{AO}\ln(2p_Ap_O) + n_{AB}\ln(2p_Ap_B) \right. \\ &\quad \left. + 2n_{BB}\ln(p_B) + n_{BO}\ln(2p_Bp_O) + 2n_O\ln(p_O)\}\,|p_A^{[t]}, p_B^{[t]}, \mathbf{n}\right] \\ &= 2\widetilde{n}_{AA}\ln(p_A) + \widetilde{n}_{AO}\ln(2p_Ap_O) + n_{AB}\ln(2p_Ap_B) + 2\widetilde{n}_{BB}\ln(p_B) \\ &\quad + \widetilde{n}_{BO}\ln(2p_Bp_O) + 2n_O\ln(p_O), \end{aligned} \tag{9.53}$$

where

$$\begin{aligned} \widetilde{n}_{AA} &= E\left(n_{AA}|p_A^{[t]}, p_B^{[t]}, \mathbf{n}\right), \\ \widetilde{n}_{AO} &= E\left(n_{AO}|p_A^{[t]}, p_B^{[t]}, \mathbf{n}\right), \\ \widetilde{n}_{BB} &= E\left(n_{BB}|p_A^{[t]}, p_B^{[t]}, \mathbf{n}\right), \\ \widetilde{n}_{BO} &= E\left(n_{BO}|p_A^{[t]}, p_B^{[t]}, \mathbf{n}\right). \end{aligned}$$

The M-step consists of maximizing (9.53) with respect to p_A and p_B. This yields the following closed-form solution for p_A and p_B at a round $(t+1)$:

$$p_A^{[t+1]} = \frac{2\widetilde{n}_{AA} + n_{AB} + \widetilde{n}_{AO}}{2\left(n_A + n_{AB} + n_B + n_O\right)}, \tag{9.54}$$

$$p_B^{[t+1]} = \frac{2\widetilde{n}_{BB} + n_{AB} + \widetilde{n}_{BO}}{2\left(n_A + n_{AB} + n_B + n_O\right)}. \tag{9.55}$$

The unobserved counts at iteration t are imputed via their expected values, given $\mathbf{n}$ and $\left(p_A^{[t]}, p_B^{[t]}\right)$. The unobserved counts are distributed binomially (see Example 2.15 in Chapter 2) as follows:

$$n_{AA} \sim Bi\left(\frac{p_A^2}{p_A^2 + 2p_A\left(1 - p_A - p_B\right)}, n_A\right),$$

$$n_{AO} \sim Bi\left(\frac{2p_A\left(1 - p_A - p_B\right)}{p_A^2 + 2p_A\left(1 - p_A - p_B\right)}, n_A\right),$$

$$n_{BB} \sim Bi\left(\frac{p_B^2}{p_B^2 + 2p_B\left(1 - p_A - p_B\right)}, n_B\right),$$

and

$$n_{BO} \sim Bi\left(\frac{2p_B\left(1 - p_A - p_B\right)}{p_B^2 + 2p_B\left(1 - p_A - p_B\right)}, n_B\right).$$

Hence, expectations can be computed immediately. For example,

$$\widetilde{n}_{AA} = n_A \frac{p_A^{2[t]}}{p_A^{2[t]} + 2p_A^{[t]}\left(1 - p_A^{[t]} - p_B^{[t]}\right)},$$

and similarly for the other components of the missing data. Using some starting values for the gene frequencies, the missing counts $\widetilde{n}_{ij}$ are imputed and the next round of gene frequency values are computed from (9.54) and (9.55). In the case of the present example, starting with $p_A^{[0]} = p_B^{[0]} = 0.2$ leads to $\widehat{p}_A = 0.209130654$ and to $\widehat{p}_B = 0.080801008$ (with $\widehat{p}_0 = 1 - \widehat{p}_A - \widehat{p}_B$) after 9 EM iterations.

To obtain an estimate of the asymptotic variance we apply the method of Oakes (1999). This requires obtaining the 2×2 matrix of second derivatives evaluated at $p_A, p_A^{[t]} = \widehat{p}_A$ and at $p_B, p_B^{[t]} = \widehat{p}_B$, which yields

$$\left[\begin{array}{cc} \frac{\partial^2 Q\left(p_A, p_B | p_A^{[t]}, p_B^{[t]}\right)}{\partial p_A \partial p_A} & \frac{\partial^2 Q\left(p_A, p_B | p_A^{[t]}, p_B^{[t]}\right)}{\partial p_A \partial p_B} \\ \frac{\partial^2 Q\left(p_A, p_B | p_A^{[t]}, p_B^{[t]}\right)}{\partial p_B \partial p_A} & \frac{\partial^2 Q\left(p_A, p_B | p_A^{[t]}, p_B^{[t]}\right)}{\partial p_B \partial p_B} \end{array}\right] \Bigg|_{\substack{p_A, p_A^{[t]} = \widehat{p}_A \\ p_B, p_B^{[t]} = \widehat{p}_B}}$$

$$= -\left[\begin{array}{cc} 26,349.8 & 5993.29 \\ 5993.29 & 58667.1 \end{array}\right] \tag{9.56}$$

and

$$\left[\begin{array}{cc} \frac{\partial^2 Q\left(p_A,p_B|p_A^{[t]},p_B^{[t]}\right)}{\partial p_A \partial p_A^{[t]}} & \frac{\partial^2 Q\left(p_A,p_B|p_A^{[t]},p_B^{[t]}\right)}{\partial p_A \partial p_B^{[t]}} \\ \frac{\partial^2 Q\left(p_A,p_B|p_A^{[t]},p_B^{[t]}\right)}{\partial p_B \partial p_A^{[t]}} & \frac{\partial^2 Q\left(p_A,p_B|p_A^{[t]},p_B^{[t]}\right)}{\partial p_B \partial p_B^{[t]}} \end{array} \right] \Bigg|_{\substack{p_A,p_A^{[t]}=\widehat{p}_A \\ p_B,p_B^{[t]}=\widehat{p}_B}}$$
$$= \left[\begin{array}{cc} 3134.28 & 962.148 \\ 962.148 & 2657.74 \end{array} \right]. \tag{9.57}$$

Using (9.36), the estimate of the observed information matrix is:

$$-\left\{ \left[\begin{array}{cc} -26,349.8 & -5,993.29 \\ -5,993.29 & -58,667.1 \end{array} \right] + \left[\begin{array}{cc} 3,134.28 & 962.148 \\ 962.148 & 2,657.74 \end{array} \right] \right\}$$
$$= \left[\begin{array}{cc} 23,215.5 & 5,031.14 \\ 5,031.14 & 56,009.4 \end{array} \right],$$

where the second term in the first line corresponds to the negative of the missing information. The estimate of the asymptotic variance of $(\widehat{p}_A, \widehat{p}_B)'$ is:

$$Var\left(\widehat{p}_A, \widehat{p}_B|\mathbf{n}\right) = 10^{-6} \left[\begin{array}{cc} 43.930 & -3.946 \\ -3.946 & 18.209 \end{array} \right],$$

in agreement with the estimate obtained in Example 4.7 based on Newton–Raphson. ■

Example 9.3 *Maximum likelihood estimation in the mixed linear model* In this example the iterative EM equations for the ML estimation of fixed effects and of variance components in a univariate Gaussian mixed linear model with 2 variance components are derived. The general EM algorithm, with its distinct E- and M-steps, and the form of EM discussed in Section 9.6, are illustrated. Both approaches, of course, yield identical results.

The model considered here was introduced in Example 1.18 in Chapter 1. The data $\mathbf{y}$ (vector of dimension $n \times 1$) are assumed to be a realization from

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2\right),$$

and the unobserved vector of the additive genetic values $(q \times 1)$ is assumed to follow the multivariate normal distribution

$$\mathbf{a}|\mathbf{A}\sigma_a^2 \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right).$$

The vector of fixed effects $\boldsymbol{\beta}$ has order $p \times 1$; $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices, and the unknown variance components are the scalars σ_a^2 and σ_e^2. The matrix $\mathbf{A}$ is known; it describes the covariance structure of $\mathbf{a}$ and depends on the additive genetic relationships among individuals in the

pedigree. Here the focus of inference is $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \sigma_a^2, \sigma_e^2\right)'$. The observed data likelihood is

$$L\left(\boldsymbol{\theta}|\mathbf{y}\right) = \int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2\right) p\left(\mathbf{a}|\mathbf{A}\sigma_a^2\right) d\mathbf{a}$$
$$\propto \left|\mathbf{V}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right], \tag{9.58}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{I}\sigma_e^2$ is the unconditional variance–covariance matrix of the observed data $\mathbf{y}$. Rather than working with (9.58), the ML estimate of $\boldsymbol{\theta}$ is obtained using the EM algorithm.

General EM Algorithm

Regarding the random effects $\mathbf{a}$ as the missing data, the complete-data likelihood is

$$L\left(\boldsymbol{\theta}, \mathbf{a}|\mathbf{y}\right) = \left|\mathbf{I}\sigma_e^2\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right)\right]$$
$$\times \left|\mathbf{A}\sigma_a^2\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right], \tag{9.59}$$

and the corresponding log-likelihood is

$$\ln p\left(\boldsymbol{\theta}, \mathbf{a}|\mathbf{y}\right) = \text{constant} - \frac{n}{2}\ln\sigma_e^2 - \frac{q}{2}\ln\sigma_a^2 - \frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}$$
$$-\frac{1}{2\sigma_e^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right). \tag{9.60}$$

The E-step is

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) = \int \ln p\left(\boldsymbol{\theta}, \mathbf{a}|\mathbf{y}\right) p\left(\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{a}$$
$$= -\frac{n}{2}\ln\sigma_e^2 - \frac{q}{2}\ln\sigma_a^2 - \frac{1}{2\sigma_a^2}E_{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left[\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right]$$
$$-\frac{1}{2\sigma_e^2}E_{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right).$$

Let:

$$E_{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left[\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right] = \widetilde{\mathbf{a}}^{[t]},$$

and

$$Var_{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left[\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right] = \widetilde{\mathbf{V}}_a^{[t]}.$$

Then using results for expectation of quadratic forms (Searle, 1971),

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) = -\frac{n}{2}\ln\sigma_e^2 - \frac{q}{2}\ln\sigma_a^2 - \frac{1}{2\sigma_a^2}\left[\widetilde{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widetilde{\mathbf{a}}^{[t]} + tr\left(\mathbf{A}^{-1}\widetilde{\mathbf{V}}_a^{[t]}\right)\right]$$
$$-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right) + tr\left(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right)\right].$$

The M-step consists of setting the following equations equal to zero:

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_e^2}\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right),$$

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \sigma_a^2} = -\frac{q}{2\sigma_a^2} + \frac{1}{2\left(\sigma_a^2\right)^2}\left[\widetilde{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widetilde{\mathbf{a}}^{[t]} + tr\left(\mathbf{A}^{-1}\widetilde{\mathbf{V}}_a^{[t]}\right)\right],$$

and

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial \sigma_e^2} = -\frac{n}{2\sigma_e^2}$$
$$+\frac{1}{2\left(\sigma_e^2\right)^2}\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right) + tr\left(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right)\right].$$

Solving for $\boldsymbol{\theta}$ one obtains the iterative system

$$\boldsymbol{\beta}^{[t+1]} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right), \tag{9.61}$$

$$\sigma_a^{2[t+1]} = \frac{1}{q}\left[\widetilde{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widetilde{\mathbf{a}}^{[t]} + tr\left(\mathbf{A}^{-1}\widetilde{\mathbf{V}}_a^{[t]}\right)\right], \tag{9.62}$$

and

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right) + tr\left(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right)\right]. \tag{9.63}$$

More explicit expressions for $\widetilde{\mathbf{a}}^{[t]}$ and $\widetilde{\mathbf{V}}_a^{[t]}$ are given in a section at the end of this example.

Exponential Family Version of EM

The complete-data likelihood (9.59) can be put in the form (9.20) up to proportionality, by making use of the following:

$$\mathbf{x} = \left[\mathbf{y}', \mathbf{a}'\right]',$$

$$b\left(\mathbf{x}\right) = 1,$$

$$\begin{aligned}\boldsymbol{\theta} &= \left(\theta_1, \theta_2, \boldsymbol{\theta}_3\right)' \\ &= \left[\frac{1}{\sigma_a^2}, \frac{1}{\sigma_e^2}, \frac{\boldsymbol{\beta}}{\sigma_e^2}\right]',\end{aligned}$$

$$
\begin{aligned}
a(\boldsymbol{\theta}) &= \left(\sigma_a^2\right)^{\frac{q}{2}} \left(\sigma_e^2\right)^{\frac{n}{2}} \exp\left[\frac{1}{2\sigma_e^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] \\
&= \theta_1^{-\frac{q}{2}}\theta_2^{-\frac{n}{2}} \exp\left[\frac{\boldsymbol{\theta}_3'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}_3}{2\theta_2}\right], \qquad (9.64a)
\end{aligned}
$$

$$
\mathbf{t}(\mathbf{x}) = \begin{bmatrix} -\frac{1}{2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \\ -\frac{1}{2}(\mathbf{y}-\mathbf{Z}\mathbf{a})'(\mathbf{y}-\mathbf{Z}\mathbf{a}) \\ \mathbf{X}'(\mathbf{y}-\mathbf{Z}\mathbf{a}) \end{bmatrix}.
$$

The implementation of the EM algorithm consists of solving for $\boldsymbol{\theta}^{[t+1]}$ the system of equations defined by (9.28). The right-hand side of (9.28) is the unconditional expectation of the vector of sufficient statistics $\mathbf{t}(\mathbf{x})$, which is equal to (9.23). From (9.64a),

$$
\ln a(\boldsymbol{\theta}) = -\frac{q}{2}\ln\theta_1 - \frac{n}{2}\ln\theta_2 + \frac{1}{2\theta_2}\boldsymbol{\theta}_3'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}_3.
$$

The required unconditional expectations are given by the following partial derivatives

$$
\frac{\partial \ln a(\boldsymbol{\theta})}{\partial\theta_1} = -\frac{q}{2\theta_1},
$$

$$
\frac{\partial \ln a(\boldsymbol{\theta})}{\partial\theta_2} = -\frac{n}{2\theta_2} - \frac{1}{2\theta_2^2}\boldsymbol{\theta}_3'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}_3,
$$

and

$$
\frac{\partial \ln a(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_3} = \frac{1}{\theta_2}\mathbf{X}'\mathbf{X}\boldsymbol{\theta}_3.
$$

The conditional expectation of $\mathbf{t}(\mathbf{x})$ (given $\mathbf{y}$ and $\boldsymbol{\theta}^{[t]}$) has the components

$$
\underset{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}{E}\left[-\frac{1}{2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right] = -\frac{1}{2}\left[\widetilde{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widetilde{\mathbf{a}}^{[t]} + tr\left(\mathbf{A}^{-1}\widetilde{\mathbf{V}}_a^{[t]}\right)\right],
$$

$$
\begin{aligned}
&\underset{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}{E}\left[-\frac{1}{2}(\mathbf{y}-\mathbf{Z}\mathbf{a})'(\mathbf{y}-\mathbf{Z}\mathbf{a})\,|\boldsymbol{\theta}^{[t]},\mathbf{y}\right] \\
= \;& -\frac{1}{2}\left[\left(\mathbf{y}-\mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)'\left(\mathbf{y}-\mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right) + tr\left(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right)\right],
\end{aligned}
$$

$$
\underset{\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}{E}\left[\mathbf{X}'(\mathbf{y}-\mathbf{Z}\mathbf{a})\,|\boldsymbol{\theta}^{[t]},\mathbf{y}\right] = \mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right).
$$

Writing the equations $\partial \ln a(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ in terms of $\left(\boldsymbol{\beta}, \sigma_a^2, \sigma_e^2\right)$, and equating to the conditional expectations, yields

$$
\sigma_a^{2[t+1]} = \frac{1}{q}\left[\widetilde{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widetilde{\mathbf{a}}^{[t]} + tr\left(\mathbf{A}^{-1}\widetilde{\mathbf{V}}_a^{[t]}\right)\right], \qquad (9.65)
$$

$$n\sigma_e^{2[t+1]} + \boldsymbol{\beta}^{[t+1]\prime}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{[t+1]}$$
$$= \left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)' \left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right) + tr\left(\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right), \tag{9.66}$$

and

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{[t+1]} = \mathbf{X}'\left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right). \tag{9.67}$$

From (9.67),

$$\boldsymbol{\beta}^{[t+1]} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right). \tag{9.68}$$

Substituting (9.68) in (9.66), and using

$$\left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)' \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right] \left(\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)$$
$$= \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right),$$

one obtains

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\widetilde{\mathbf{a}}^{[t]}\right)\right]$$
$$+ \frac{1}{n} tr\left[\mathbf{Z}'\mathbf{Z}\widetilde{\mathbf{V}}_a^{[t]}\right].$$

The iterative system arrived at is identical to that defined by (9.61), (9.62) and (9.63).

Algebraic Notes

The starting point for the computation of the mean vector and variance–covariance matrix of $\left[\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right]$ is the joint distribution $\left[\mathbf{a}, \mathbf{y}|\boldsymbol{\theta}^{[t]}\right]$:

$$\begin{matrix}\mathbf{y}\\ \mathbf{a}\end{matrix}\ \Bigg|\ \boldsymbol{\theta}^{[t]} \sim N\left(\begin{bmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{V} & \mathbf{Z}\mathbf{A}\sigma_a^2\\ \mathbf{A}\mathbf{Z}'\sigma_a^2 & \mathbf{A}\sigma_a^2\end{bmatrix}\right),$$

where $\mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{I}\sigma_e^2$. From properties of the multivariate normal distribution it follows that

$$\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y} \sim N\left(E\left(\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right), Var\left(\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right)\right),$$

where

$$E\left(\mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) = \mathbf{A}\mathbf{Z}'\mathbf{V}^{-1[t]}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t]}\right)\sigma_a^{2[t]}.$$

Substituting in this expression

$$\mathbf{V}^{-1[t]} = \frac{1}{\sigma_e^{2[t]}}\mathbf{I} - \frac{1}{\sigma_e^{2[t]}}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k^{[t]}\right)^{-1}\mathbf{Z}',$$

gives

$$
\begin{aligned}
&E\left(\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right)\\
&=\mathbf{A}\mathbf{Z}'\left[\frac{1}{\sigma_e^{2[t]}}\mathbf{I}-\frac{1}{\sigma_e^{2[t]}}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\mathbf{Z}'\right]\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t]}\right)\sigma_a^{2[t]}\\
&=\left[\mathbf{A}\frac{1}{k^{[t]}}-\frac{1}{k^{[t]}}\mathbf{A}\mathbf{Z}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\right]\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t]}\right)\\
&=\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[t]}\right),
\end{aligned}
\tag{9.69}
$$

where $k^{[t]}=\sigma_e^{2[t]}/\sigma_a^{2[t]}$. The last line follows because

$$
\mathbf{A}\frac{1}{k^{[t]}}-\frac{1}{k^{[t]}}\mathbf{A}\mathbf{Z}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}=\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}.
$$

This can be verified by post-multiplying the left-hand side by the inverse of the right-hand side, which retrieves $\mathbf{I}$.

Now we show that

$$
Var\left(\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right)=\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\sigma_e^{2[t]}. \tag{9.70}
$$

Again, using properties of the multivariate normal distribution,

$$
\begin{aligned}
&Var\left(\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}\right)=\mathbf{A}\sigma_a^{2[t]}-\mathbf{A}\mathbf{Z}'\sigma_a^{2[t]}\mathbf{V}^{-1[t]}\mathbf{Z}\mathbf{A}\sigma_a^{2[t]}\\
&=\mathbf{A}\sigma_a^{2[t]}-\left[\mathbf{A}\frac{1}{k^{[t]}}-\frac{1}{k^{[t]}}\mathbf{A}\mathbf{Z}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\right]\mathbf{Z}'\mathbf{Z}\mathbf{A}\sigma_a^{2[t]}\\
&=\mathbf{A}\sigma_a^{2[t]}-\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{A}\sigma_a^{2[t]}\\
&=\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\sigma_e^{2[t]}.
\end{aligned}
$$

The last line can be verified by premultiplying the third line by

$$
\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right),
$$

which gives $\mathbf{I}\sigma_e^{2[t]}$.

An alternative derivation from a Bayesian perspective (which implies assigning a flat, improper prior to the location vector $\boldsymbol{\beta}$) is to use as point of departure

$$
\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{a}\end{bmatrix}\Bigg|\,\sigma_a^{2[t]},\sigma_e^{2[t]},\mathbf{y}\sim N\left(\begin{bmatrix}\widehat{\boldsymbol{\beta}}^{[t]}\\ \widehat{\mathbf{a}}^{[t]}\end{bmatrix},\begin{bmatrix}\mathbf{C}^{11} & \mathbf{C}^{12}\\ \mathbf{C}^{21} & \mathbf{C}^{22}\end{bmatrix}^{[t]}\sigma_e^{2[t]}\right). \tag{9.71}
$$

Using similar algebra as in Example 1.18 from Chapter 1, one can show that the mean vector of $\left[\mathbf{a}|\boldsymbol{\beta},\sigma_a^{2[t]},\sigma_e^{2[t]},\mathbf{y}\right]$ is equal to

$$
\left(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k^{[t]}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right),
$$

and its covariance matrix is equal to (9.70). ■

Example 9.4 *Restricted maximum likelihood estimation in the mixed linear model*

The model is as in the preceding example, but now the focus of inference is $\boldsymbol{\theta} = \left(\sigma_a^2, \sigma_e^2\right)$, with $\boldsymbol{\beta}$ viewed as a nuisance parameter. A Bayesian perspective will be adopted, and the mode of the posterior distribution with density $p\left(\sigma_a^2, \sigma_e^2|\mathbf{y}\right)$ is chosen as point estimator. Assigning improper uniform prior distributions to each of $\left(\sigma_a^2, \sigma_e^2\right)$ and to $\boldsymbol{\beta}$, then

$$p\left(\sigma_a^2, \sigma_e^2|\mathbf{y}\right) \propto \int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2\right) p\left(\mathbf{a}|\mathbf{A}, \sigma_a^2\right) d\mathbf{a}\, d\boldsymbol{\beta}.$$

In this setting the mode of the posterior distribution of the variance components is identical to the REML estimator, (Harville, 1974). Joint maximization of this expression is difficult. However, it is relatively easy to structure an EM algorithm, where the missing data are now $\mathbf{z} = \left(\boldsymbol{\beta}', \mathbf{a}'\right)'$. The complete-data posterior distribution $p\left(\sigma_a^2, \sigma_e^2, \mathbf{z}|\mathbf{y}\right)$ is identical to (9.60) and the E-step is now

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) &= \int \ln p\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{a}|\mathbf{y}\right) p\left(\boldsymbol{\beta}, \mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right) d\mathbf{a}\, d\boldsymbol{\beta} \\ &= -\frac{n}{2}\ln\sigma_e^2 - \frac{q}{2}\ln\sigma_a^2 - \frac{1}{2\sigma_a^2} E_{\boldsymbol{\beta},\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left[\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right] \\ &-\frac{1}{2\sigma_e^2} E_{\boldsymbol{\beta},\mathbf{a}|\boldsymbol{\theta}^{[t]},\mathbf{y}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right). \end{aligned} \tag{9.72}$$

This, on using (9.71), takes the form

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right) &= -\frac{n}{2}\ln\sigma_e^2 - \frac{q}{2}\ln\sigma_a^2 \\ &-\frac{1}{2\sigma_a^2}\left[\widehat{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widehat{\mathbf{a}} + tr\left(\mathbf{A}^{-1}\mathbf{C}^{22[t]}\right)\sigma_e^{2[t]}\right] \\ &-\frac{1}{2\sigma_e^2}\left[\widehat{\mathbf{e}}^{[t]\prime}\widehat{\mathbf{e}}^{[t]} + tr\left[\left[\mathbf{X}, \mathbf{Z}\right]\mathbf{C}^{-1[t]}\left[\mathbf{X}, \mathbf{Z}\right]'\right]\sigma_e^{2[t]}\right], \end{aligned}$$

where $\widehat{\mathbf{e}}^{[t]} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{[t]} - \mathbf{Z}\widehat{\mathbf{a}}^{[t]}\right)$ and

$$\mathbf{C}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}.$$

Since the missing data are now $\mathbf{z} = \left(\boldsymbol{\beta}', \mathbf{a}'\right)'$, expectations in (9.72) are taken with respect to $\left[\boldsymbol{\beta}, \mathbf{a}|\boldsymbol{\theta}^{[t]}, \mathbf{y}\right]$. The M-step is

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\sigma_a^2} = -\frac{q}{2\sigma_a^2} + \frac{1}{2\left(\sigma_a^2\right)^2}\left[\widehat{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widehat{\mathbf{a}} + tr\left(\mathbf{A}^{-1}\mathbf{C}^{22[t]}\right)\sigma_e^{2[t]}\right],$$

and

$$\frac{\partial Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}\right)}{\partial\sigma_e^2} = -\frac{n}{2\sigma_e^2}$$
$$+\frac{1}{2\left(\sigma_e^2\right)^2}\left[\widehat{\mathbf{e}}^{[t]\prime}\widehat{\mathbf{e}}^{[t]} + tr\left[[\mathbf{X},\mathbf{Z}]\,\mathbf{C}^{-1[t]}\,[\mathbf{X},\mathbf{Z}]'\right]\sigma_e^{2[t]}\right].$$

Setting to zero yields the iterative system

$$\sigma_a^{2[t+1]} = \frac{\widehat{\mathbf{a}}^{[t]\prime}\mathbf{A}^{-1}\widehat{\mathbf{a}} + tr\left(\mathbf{A}^{-1}\mathbf{C}^{22[t]}\right)\sigma_e^{2[t]}}{q}, \tag{9.73}$$

$$\sigma_e^{2[t+1]} = \frac{\widehat{\mathbf{e}}^{[t]\prime}\widehat{\mathbf{e}}^{[t]} + tr\left[[\mathbf{X},\mathbf{Z}]\,\mathbf{C}^{-1[t]}\,[\mathbf{X},\mathbf{Z}]'\right]\sigma_e^{2[t]}}{n}. \tag{9.74}$$

Contrary to ML estimation, REML estimation or inference via the posterior mode of $\left[\sigma_a^2, \sigma_e^2|\mathbf{y}\right]$ requires inverting the entire coefficient matrix $\mathbf{C}$. ■

The preceding examples illustrate the versatility of the EM algorithm. Viewed more generally, the algorithm can be interpreted as a data augmentation technique (Tanner and Wong, 1987), with the imputations made using conditional expectations. Additional examples of varying degrees of complexity can be found in Little and Rubin (1987) and in McLachlan and Krishnan (1997).

This page intentionally left blank

Part III

Markov Chain Monte Carlo Methods

This page intentionally left blank

10

An Overview of Discrete Markov Chains

10.1 Introduction

The theory of Markov chains governs the behavior of the Markov chain Monte Carlo (MCMC) methods that are discussed in Chapter 11 and onwards. The purpose of this chapter is to present an overview of some elements of this theory so that the reader can obtain a feeling for the mechanisms underlying MCMC. Feller (1970), Cox and Miller (1965), Karlin and Taylor (1975), Meyn and Tweedie (1993), Norris (1997), or Robert and Casella (1999) should be consulted for a more detailed and formal treatment of the subject. A recent contribution to the literature, that discusses the finer theoretical and practical issues in a clear style with a view towards MCMC applications, is Liu (2001). For the sake of simplicity, this chapter considers only chains with finite state spaces.

The aspects of Markov chain theory discussed here are the following ones. First, the fundamental ingredients of a Markov chain are introduced. These consist of the initial probability distribution of the states of the chain and the matrix of transition probabilities. The two together govern the evolution of the Markov chain. One wishes to know if, after a number of transitions, the Markov chain converges to some equilibrium distribution, independently of the initial probability distribution, and if this equilibrium distribution is unique. Two important properties for discrete, finite state Markov chains, establish the existence of a unique equilibrium distribution: aperiodicity and irreducibility. Another important property especially from the point of view of MCMC is that of reversibility. As shown in Chapter

11, transition probability matrices or transition kernels (in the case of continuous Markov chains) can be constructed more or less easily using the condition of reversibility as point of departure. The transition kernels constructed in this way, guarantee that aperiodic and irreducible chains have a unique equilibrium distribution. Convergence to this distribution, however, takes place only asymptotically. Further, the rate of convergence is an important aspect of the behavior of MCMC, a subject which is discussed and illustrated in a closing section of this chapter. Other examples illustrating convergence can be found in Chapter 12.

10.2 Definitions

Following Grossman and Turner (1974), suppose that a mouse is placed in a box divided into three intercommunicating compartments labelled 1, 2, and 3. Technically the three compartments are the possible states of the system at any time, and the set $S = \{1, 2, 3\}$ is called the state space. One can define the random variable "presence of the mouse in a given compartment". This random variable can take one of three possible values. A movement from compartment i to j (which includes the case where $i = j$, indicating no movement) is referred to as a transition of the system from the ith to the jth state. This transition will occur with some conditional probability $p(i, j)$, called the transition probability of moving from compartment i to j. This probability is conditional on the mouse being in compartment i. Consider a sequence of random variables defining the presence of the mouse in the compartments over a period of time. If the probability that the mouse moves to a given compartment depends on which compartment it finds itself immediately before the move, and not on which compartments the mouse had visited in the past, then the sequence of random variables defines a Markov chain. A Markov chain deals with the study of the possible transitions between the states of a system via probabilistic methods. An important goal is to characterize the probability distribution after n transitions. This is the distribution of the proportion of times that the mouse is expected to be in each compartment after n transitions. It is of interest to know whether for large n, this distribution stabilizes, irrespective of the initial distribution. A formal definition of a Markov chain follows.

A finite state, discrete Markov chain is a sequence of random variables X_n, $(n = 0, 1, 2, \ldots,)$ where the X_n take values in the finite set $S = \{0, 1, \ldots, N-1\}$ and are called the states of the Markov chain. The subscripts n can be interpreted as stages or time periods. If $X_n = i$, the process is said to be in state i at time n. A Markov chain must satisfy the

following Markov property

$$\begin{aligned} &\Pr\left(X_n = j | X_{n-1} = i, X_{n-2} = k, \ldots, X_0 = m\right) \\ &= \Pr\left(X_n = j | X_{n-1} = i\right) = p\left(i, j\right), \quad i, j, k, \ldots, m \in S. \end{aligned} \tag{10.1}$$

This property may be interpreted as stating that, for a Markov chain, the conditional distribution at time n, given all the past states $X_0 = m, \ldots, X_{n-1} = i$, only depends on the immediately preceding state, $X_{n-1} = i$. A sequence of independent random variables is a trivial example of a Markov chain.

The evolution of the chain is described by its transition probability (10.1). The element $p\left(i, j\right)$ represents the probability that the chain at time n is in state j, given that at time $n-1$ it was in state i. This is a conditional probability, where j is stochastic and i is fixed. In the notation of distribution theory, this would normally be written $p\left(j|i\right)$. However, the standard Markov chain notation is adhered to here and in the next chapter (for example, as in Cox and Miller, 1965). The transition probabilities can be arranged in a matrix $\mathbf{P} = \{p\left(i, j\right)\}$, where i is a row suffix and j a column suffix. This is the $N \times N$ transition probability matrix of the Markov chain. Only transition probability matrices that are independent of time (n) are considered here. These are known as time homogeneous chains: the probability of a transition from a given state to another depends on the two states and not on time.

The $(i+1)$th row of $\mathbf{P}$ is the probability distribution of the values of X_n under the condition that $X_{n-1} = i$. Every entry of $\mathbf{P}$ satisfies $p\left(i, j\right) \geq 0$, and every row of $\mathbf{P}$ satisfies $\sum_j p\left(i, j\right) = 1$.

10.3 State of the System after n-Steps

As stated above, the matrix $\mathbf{P}$ describes the probability of transitions of the chain in one time period. Consider a chain with state space $S = \{0, 1, 2\}$. The matrix of transition probabilities, of order 3×3 takes the form

$$\mathbf{P} = \begin{bmatrix} p\left(0,0\right) & p\left(0,1\right) & p\left(0,2\right) \\ p\left(1,0\right) & p\left(1,1\right) & p\left(1,2\right) \\ p\left(2,0\right) & p\left(2,1\right) & p\left(2,2\right) \end{bmatrix}.$$

The probability that at stage (or time) $2+m$ the random variable will be in state 2, given that it was in stage 1 at time m can be written as

$$\begin{aligned} p^{(2)}(1,2) &= \Pr(X_{2+m}=2|X_m=1) \\ &= \sum_{j=0}^{2} \Pr(X_{2+m}=2, X_{1+m}=j|X_m=1) \\ &= \sum_{j=0}^{2} \Pr(X_{1+m}=j|X_m=1)\Pr(X_{2+m}=2|X_{1+m}=j) \\ &= p(1,0)\,p(0,2)+p(1,1)\,p(1,2)+p(1,2)\,p(2,2). \end{aligned} \tag{10.2}$$

The equality in the third line follows from the Markov property. Equation (10.2) makes it explicit that to compute the probability of moving from state 1 to 2 in two transitions, requires summation over the probabilities of going through all possible intermediate states before the system reaches state 2. The last line in (10.2) can be recognized as the product of row 2 of $\mathbf{P}$ (associated with state 1) and column 3 of $\mathbf{P}$ (associated with state 2).

The preceding argument can be generalized to arrive at two important results. The first one concerns the probability of moving from i to j in n transitions. This is given by

$$p^{(n)}(i,j) = \Pr(X_n=j|X_0=i) = \Pr(X_{n+k}=j|X_k=i). \tag{10.3}$$

The element $p^{(n)}(i,j)$ is the $(i+1, j+1)$ entry of $\mathbf{P}^n$. The second result is described by the Chapman–Kolmogorov equations

$$\begin{aligned} p^{(m+n)}(i,j) &= \Pr(X_{m+n}=j|X_0=i) \\ &= \sum_{k=0}^{N-1} \Pr(X_{m+n}=j, X_m=k|X_0=i) \\ &= \sum_{k=0}^{N-1} \Pr(X_{m+n}=j|X_m=k)\Pr(X_m=k|X_0=i) \\ &= \sum_{k=0}^{N-1} p^{(m)}(i,k)\,p^{(n)}(k,j). \end{aligned} \tag{10.4}$$

The matrix analogue to (10.4) is

$$\mathbf{P}^{m+n} = \mathbf{P}^m\mathbf{P}^n.$$

This result relates long-term behavior to short-term behavior and describes how X_n depends on the starting value X_0.

10.4 Long-Term Behavior of the Markov Chain

Let $\boldsymbol{\pi}'^{(n)}$ be the N-dimensional row vector denoting the probability distribution of X_n. The i^{th} component of $\boldsymbol{\pi}^{(n)}$ is:

$$\pi^{(n)}(i) = \Pr(X_n = i), \quad i \in S.$$

Clearly, when $n = 0$, $\boldsymbol{\pi}^{(0)}$ defines the initial probability distribution of the chain. Now,

$$\begin{aligned}\Pr(X_n = j) &= \sum_{i=0}^{N-1} \Pr(X_n = j | X_{n-1} = i) \Pr(X_{n-1} = i) \\ &= \sum_{i=0}^{N-1} p(i,j) \Pr(X_{n-1} = i), \ j = 0, 1, \ldots, N-1. \end{aligned} \tag{10.5}$$

The left-hand side is the $(j+1)$th element of $\boldsymbol{\pi}^{(n)}$, and the right-hand side is the product of the row vector $\boldsymbol{\pi}'^{(n-1)}$ with column $(j+1)$ of the transition matrix $\mathbf{P}$. Therefore a matrix generalization of (10.5) is given by

$$\boldsymbol{\pi}'^{(n)} = \boldsymbol{\pi}'^{(n-1)}\mathbf{P}. \tag{10.6}$$

Since $\boldsymbol{\pi}'^{(1)} = \boldsymbol{\pi}'^{(0)}\mathbf{P}$, $\boldsymbol{\pi}'^{(2)} = \boldsymbol{\pi}'^{(1)}\mathbf{P} = \boldsymbol{\pi}'^{(0)}\mathbf{P}\mathbf{P} = \boldsymbol{\pi}'^{(0)}\mathbf{P}^2$, and so on, it follows that the probability distribution of the chain at time n is given by

$$\boldsymbol{\pi}'^{(n)} = \boldsymbol{\pi}'^{(0)}\mathbf{P}^n. \tag{10.7}$$

Thus, we reach the important conclusion that the random evolution of the Markov chain is completely specified in terms of the distribution of the initial state and the transition probability matrix $\mathbf{P}$. This means that given the initial probability distribution and the transition probability matrix, it is possible to describe the behavior of the process at any specified time period n.

10.5 Stationary Distribution

A question of fundamental importance is whether the chain converges to a limiting distribution, independent of any legal starting distribution. As shown below, this requires that $\mathbf{P}^n$ converges to some invariant matrix, and that, at the limit, it has identical rows. Suppose that $\boldsymbol{\pi}$ is a limiting probability vector such that, from (10.7),

$$\boldsymbol{\pi}' = \lim_{n \to \infty} \boldsymbol{\pi}'^{(0)}\mathbf{P}^n.$$

Then,

$$\begin{aligned}\boldsymbol{\pi}' &= \lim_{n\to\infty} \boldsymbol{\pi}'^{(0)}\mathbf{P}^{n+1} \\ &= \left(\lim_{n\to\infty} \boldsymbol{\pi}'^{(0)}\mathbf{P}^{n}\right)\mathbf{P} \\ &= \boldsymbol{\pi}'\mathbf{P}. \end{aligned} \tag{10.8}$$

The distribution $\boldsymbol{\pi}$ is said to be a stationary distribution (also known as the invariant or equilibrium distribution) if it satisfies (10.8). An interpretation of (10.8) is that if a chain has reached a stage where $\boldsymbol{\pi}$ is the stationary distribution, it retains it in subsequent moves. Expression (10.8) can also be written as the system of equations

$$\pi(j) = \sum_{i=0}^{N-1} \pi(i)\, p(i,j). \tag{10.9}$$

For the case of finite state Markov chains under consideration here, stationary distributions always exist (Grimmet and Stirzaker, 1992). The issue in general is convergence and uniqueness. This problem is taken up at the end of this chapter.

Example 10.1 *A three-state space Markov chain*
Consider a Markov chain consisting of three states, $0, 1, 2$, with the following transition probability matrix:

$$\begin{aligned}\mathbf{P} &= \begin{bmatrix} p(0,0) & p(0,1) & p(0,2) \\ p(1,0) & p(1,1) & p(1,2) \\ p(2,0) & p(2,1) & p(2,2) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}. \end{aligned}$$

Row 2 say, with elements $p(1,0) = \frac{1}{4}$, $p(1,1) = \frac{1}{2}$, and $p(1,2) = \frac{1}{4}$ represents the probability distribution of the values of X_{n+1} given $X_n = 1$. Element $p(1,2)$ say, is the probability that the system moves from state 1 to state 2 in one transition. It can be verified that

$$\mathbf{P}^4 = \begin{bmatrix} 0.2760 & 0.4653 & 0.2587 \\ 0.2326 & 0.4485 & 0.3189 \\ 0.1725 & 0.4251 & 0.4024 \end{bmatrix}$$

and $p^4(1,2) = 0.3189$ is the probability of moving from state 1 to state 2 in four transitions. If the starting probability distribution of the chain is

$$\boldsymbol{\pi}'^{(0)} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix},$$

then we have, to four decimal places,

$$\begin{aligned}
\boldsymbol{\pi}'^{(1)} &= \begin{pmatrix} 0.2500 & 0.5000 & 0.2500 \end{pmatrix}, \\
\boldsymbol{\pi}'^{(2)} &= \begin{pmatrix} 0.2500 & 0.4583 & 0.2917 \end{pmatrix}, \\
\boldsymbol{\pi}'^{(4)} &= \begin{pmatrix} 0.2396 & 0.4514 & 0.3090 \end{pmatrix}, \\
\boldsymbol{\pi}'^{(5)} &= \begin{pmatrix} 0.2326 & 0.4485 & 0.3189 \end{pmatrix}.
\end{aligned}$$

Eventually the system converges to the stationary distribution

$$\boldsymbol{\pi}' = \begin{pmatrix} 0.2222 & 0.4444 & 0.3333 \end{pmatrix}, \tag{10.10}$$

and, for large n,

$$\mathbf{P}^n = \begin{bmatrix} \boldsymbol{\pi}' \\ \boldsymbol{\pi}' \\ \boldsymbol{\pi}' \end{bmatrix}.$$

The same stationary distribution (10.10) is arrived at, regardless of the starting value of $\boldsymbol{\pi}^{(0)}$. In fact, for this example, the distribution (10.10) is unique. This distribution can be derived from (10.8) or (10.9). Let the stationary distribution be

$$\boldsymbol{\pi}' = \begin{pmatrix} \pi(0) & \pi(1) & \pi(2) \end{pmatrix},$$

with $\pi(2) = 1 - \pi(1) - \pi(0)$. From (10.8), the system of equations to be solved for $\pi(0)$ and $\pi(1)$ is

$$\frac{\pi(0)}{2} + \frac{\pi(1)}{4} = \pi(0),$$

$$\frac{\pi(0)}{2} + \frac{\pi(1)}{2} + \frac{1 - \pi(1) - \pi(0)}{3} = \pi(1).$$

The unique solution is $\pi(0) = 2/9$, $\pi(1) = 4/9$, and $\pi(2) = 1 - \pi(1) - \pi(0) = 3/9$. ■

10.6 Aperiodicity and Irreducibility

In the example above the Markov chain converges to a stationary distribution, and it was stated that this distribution is unique. For finite state Markov chains, convergence and uniqueness require the chain to be aperiodic and irreducible.

Consider a three state Markov chain and that the only possible transitions are $1 \to 2$, $2 \to 3$, and $3 \to 1$. That is, the states repeat themselves every 3 movements of the chain. If the process starts at state 2, say, this state will be revisited in a periodic fashion at times $3, 6, 9, \ldots$. The greatest

common divisor of these integers is 3; it is then said that this state has period equal to 3.

Formally, the period of state j is defined as the greatest common divisor of all integers $n \geq 1$ for which $p^{(n)}(j,j) > 0$. If the period of state j of the chain is d say, it means that $p^{(n)}(j,j) = 0$ whenever n is not divisible by d, and d is the largest integer with this property. An aperiodic state has period 1, or alternatively, a state j is aperiodic if $p^{(n)}(j,j) > 0$ for all sufficiently large n. A sufficient condition for a state to have period 1 is

$$\Pr(X_n = j | X_0 = j) \text{ and } \Pr(X_{n+1} = j | X_0 = j) > 0 \tag{10.11}$$

for some $n \geq 0$ and some state $j = 0, 1, \ldots, N-1$. Clearly, aperiodicity holds when $p(j,j) = \Pr(X_n = j | X_{n-1} = j) > 0$ for all j.

An important property of aperiodicity is that if states j and i communicate and state j is aperiodic, then state i is also aperiodic. A chain is aperiodic if all states have period 1. Now, if all states of a Markov chain communicate, such that every state is reachable from every other state in a finite number of transitions, the Markov chain is called irreducible. All states of an irreducible chain have the same period.

More formally, irreducibility means that for every pair of states (i,j), $p^{(k)}(i,j) = \Pr(X_{n+k} = j | X_n = i) > 0$ for some $k \geq 0$ (the value of k may be different for different i, j; more formally we should write $k(i,j)$).

A chain that is irreducible with period d has a transition probability matrix with d eigenvalues with absolute value 1. This property has important implications for convergence of the chain, as is illustrated in the example below.

Finite state Markov chains that are aperiodic and irreducible have the property that, for some $n \geq 0$, $\mathbf{P}^n$ has all entries positive. Such a Markov chain is called ergodic. In the final section of this chapter, it is shown that an ergodic Markov chain converges to a unique stationary distribution $\boldsymbol{\pi}$, for all legal initial probability distributions. This stationary distribution is the unique solution to (10.8). Four cases of Markov chains that are not ergodic are presented in the following example. The transition probability matrices of these Markov chains are such that, for large n, $\mathbf{P}^n$ has non-positive entries.

Example 10.2 *Four Markov chains that are not ergodic*
Consider a Markov chain with the following transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Taking powers of this matrix discloses quickly that $\mathbf{P}^n$ looks different depending on whether n is even or odd. For large n, if n is even,

$$\mathbf{P}^n \simeq \begin{bmatrix} 0.25 & 0 & 0.50 & 0 & 0.25 \\ 0 & 0.50 & 0 & 0.50 & 0 \\ 0.25 & 0 & 0.50 & 0 & 0.25 \\ 0 & 0.50 & 0 & 0.50 & 0 \\ 0.25 & 0 & 0.50 & 0 & 0.25 \end{bmatrix},$$

and if n is odd,

$$\mathbf{P}^n \simeq \begin{bmatrix} 0 & 0.50 & 0 & 0.50 & 0 \\ 0.25 & 0 & 0.50 & 0 & 0.25 \\ 0 & 0.50 & 0 & 0.50 & 0 \\ 0.25 & 0 & 0.50 & 0 & 0.25 \\ 0 & 0.50 & 0 & 0.50 & 0 \end{bmatrix}.$$

The Markov chain defined by this transition probability is irreducible (for some n, $p^{(n)}(i,j) > 0$ for all i,j) and periodic with period $d = 2$ (starting in state 1 say, $p^n(1,1) > 0$ at times $n = 2,4,6,8,\ldots$; the greatest common divisor of the values that n can take is 2). The unique stationary distribution for this chain (the only solution to (10.8)) can be shown to be

$$\begin{aligned} \boldsymbol{\pi} &= \lim_{n\to\infty} \frac{1}{2}\left(\boldsymbol{\pi}^{(n)} + \boldsymbol{\pi}^{(n+1)}\right) \\ &= \begin{bmatrix} 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \end{bmatrix}'. \end{aligned}$$

Even though a unique stationary distribution exists, the chain does not converge to it. The equilibrium probability at state i, $\pi(i)$, rather than representing the limit of $p^{(n)}(j,i)$, represents the average amount of time that is spent in state i.

One can verify that the eigenvalues of $\mathbf{P}$ are $-1, 0, 1, -1/\sqrt{2}, 1/\sqrt{2}$; that is, there are $d = 2$ eigenvalues with absolute value 1. We return to this point at the end of the chapter.

As a second case, consider the chain with the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{10.12}$$

Rather than a unique stationary distribution, the system has five stationary distributions. Thus, for large n,

$$\mathbf{P}^n \simeq \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 & 0.25 \\ 0.50 & 0 & 0 & 0 & 0.50 \\ 0.25 & 0 & 0 & 0 & 0.75 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The rows of $\mathbf{P}^n$ represent the five stationary distributions, and each of these satisfy (10.8).

As a third case, consider the following reducible chain defined by the transition probability

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{5}{6} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{24} & \frac{16}{24} & \frac{5}{24} \\ 0 & 0 & 0 & \frac{1}{6} & \frac{5}{6} \end{bmatrix}.$$

For large n

$$\mathbf{P}^n \simeq \begin{bmatrix} 0.25 & 0.75 & 0 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.182 & 0.364 & 0.455 \\ 0 & 0 & 0.182 & 0.364 & 0.455 \\ 0 & 0 & 0.182 & 0.364 & 0.455 \end{bmatrix}.$$

The chain splits into two noncommunicating subchains; each of these converges to an equilibrium distribution, but one cannot move from state spaces $\{0, 1\}$ to state spaces $\{2, 3, 4\}$.

The fourth case illustrates an irreducible (all the states of the chain communicate), periodic chain, with period $d = 3$,

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

The only possible transitions are from state 1, to state 2, to state 3, to state 1, and so on. That is, the chain returns to a given state at times $n = 3, 6, 9, \ldots$; the greatest common divisor of these numbers is 3. The stationary distribution obtained solving (10.8) is

$$\boldsymbol{\pi} = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]'. \tag{10.13}$$

However, the chain does not converge to (10.13). One can verify that $\mathbf{P}$ has $d = 3$ eigenvalues with absolute value 1. The consequences of this will become apparent in the last section of this chapter. ■

10.7 Reversible Markov Chains

Consider an ergodic Markov chain with state space S that converges to an invariant distribution π. Let $x \in S$ denote the current state of the system, and let $y \in S$ denote the state at the next step. Let $p(x, y)$ be the probability of a transition from x to y and let $p(y, x)$ denote the probability of a transition in the opposite direction. A Markov chain is said to be reversible if it satisfies the condition:

$$\pi(x)\, p(x, y) = \pi(y)\, p(y, x), \quad \text{for all } x, y \in S, \tag{10.14}$$

which is known as the detailed balance equation. An ergodic chain in equilibrium, satisfying (10.14) has π as its unique stationary distribution. This can be confirmed by summing both sides of (10.14) over y to yield the equilibrium condition (10.9):

$$\pi(x) = \sum_{y \in S} \pi(y)\, p(y, x).$$

The left-hand side of (10.14) can be expressed as

$$\begin{aligned} \pi(x)\, p(x, y) &= \Pr(X_n = x) \Pr(X_{n+1} = y | X_n = x) \\ &= \Pr(X_n = x, X_{n+1} = y), \quad \text{for all } x, y \in S. \end{aligned} \tag{10.15}$$

This makes explicit that detailed balance is a statement involving a joint probability. Therefore, the reversibility condition can also be written as

$$\Pr(X_n = x, X_{n+1} = y) = \Pr(X_n = y, X_{n+1} = x), \quad \text{for all } x, y \in S. \tag{10.16}$$

The reversibility condition plays an important role in the construction of MCMC algorithms, because it is often easy to generate a Markov chain having the desired stationary distribution using (10.14) as a point of departure. This is discussed in Chapter 11, where the role of reversibility in deriving appropriate transition kernels is highlighted. Proving that π is the unique stationary distribution from nonreversible chains can be difficult. Below is an example where this is not the case.

Example 10.3 *An irreducible nonreversible Markov chain*

Consider the three-state chain on $S = \{0, 1, 2\}$ with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}.$$

It is easy to verify that for large n,

$$\mathbf{P}^n = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Therefore, the unique invariant distribution is

$$\boldsymbol{\pi} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}'.$$

However, the chain is not reversible. For example,

$$\begin{aligned} \pi(0)\, p(0,1) &= \frac{1}{3}(0.2) \\ &\neq \pi(1)\, p(1,0) = \frac{1}{3}(0.1). \end{aligned}$$

■

Example 10.4 *A multinomial distribution*
Consider the 2×2 table studied by Casella and George (1992):

$$\begin{aligned} \Pr(X=0, Y=0) &= f_{x,y}(0,0) = p_1, \\ \Pr(X=1, Y=0) &= f_{x,y}(1,0) = p_2, \\ \Pr(X=0, Y=1) &= f_{x,y}(0,1) = p_3, \\ \Pr(X=1, Y=1) &= f_{x,y}(1,1) = p_4, \end{aligned}$$

with all $p_i \geq 0$ and $\sum_{i=1}^{4} p_i = 1$. The marginal distributions of X and Y are Bernoulli, with success probabilities $(p_2 + p_4)$ and $(p_3 + p_4)$, respectively,

$$\begin{aligned} f_x(0) &= p_1 + p_3, \\ f_x(1) &= p_2 + p_4, \end{aligned}$$

and

$$\begin{aligned} f_y(0) &= p_1 + p_2, \\ f_y(1) &= p_3 + p_4. \end{aligned}$$

The conditional distributions are easily derived. These are

$$\begin{aligned} \mathbf{P}_{x|y} &= \begin{bmatrix} f_{x|y}(0|0) & f_{x|y}(1|0) \\ f_{x|y}(0|1) & f_{x|y}(1|1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned}\mathbf{P}_{y|x} &= \begin{bmatrix} f_{y|x}(0|0) & f_{y|x}(1|0) \\ f_{y|x}(0|1) & f_{y|x}(1|1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix}.\end{aligned}$$

A number of Markov chains can be generated from this model. For example, the product $\mathbf{P}_{x|y}\mathbf{P}_{y|x}$ generates the transition probability matrix $\mathbf{P}_{x|x}$:

$$\mathbf{P}_{x|x} = \mathbf{P}_{x|y}\mathbf{P}_{y|x}$$

with elements

$$\begin{aligned}\Pr(X_{n+1} = i|X_n = j) = \sum_{k=0}^{1} &\Pr(X_{n+1} = i|Y_{n+1} = k) \\ &\times \Pr(Y_{n+1} = k|X_n = j).\end{aligned} \tag{10.17}$$

For some start value of the marginal distribution of X, this transition probability matrix generates the sequence X_n, $(n = 1, 2, \ldots,)$ which is a Markov chain. From (10.17) one can readily obtain $\mathbf{P}_{x|x}$ whose elements are:

$$\Pr(X_1 = 0|X_0 = 0) = \frac{p_1}{p_1+p_2}\frac{p_1}{p_1+p_3} + \frac{p_3}{p_3+p_4}\frac{p_3}{p_1+p_3},$$

$$\Pr(X_1 = 1|X_0 = 0) = \frac{p_2}{p_1+p_2}\frac{p_1}{p_1+p_3} + \frac{p_4}{p_3+p_4}\frac{p_3}{p_1+p_3},$$

$$\Pr(X_1 = 0|X_0 = 1) = \frac{p_1}{p_1+p_2}\frac{p_2}{p_2+p_4} + \frac{p_3}{p_3+p_4}\frac{p_4}{p_2+p_4},$$

$$\Pr(X_1 = 1|X_0 = 1) = \frac{p_2}{p_1+p_2}\frac{p_2}{p_2+p_4} + \frac{p_4}{p_3+p_4}\frac{p_4}{p_2+p_4}.$$

It can be confirmed that

$$\Pr(X_1 = 0|X_0 = 0) + \Pr(X_1 = 1|X_0 = 0) = 1$$

and that

$$\Pr(X_1 = 0|X_0 = 1) + \Pr(X_1 = 1|X_0 = 1) = 1.$$

Since $\mathbf{P}_{x|x}$ is a probability matrix and the elements of $\mathbf{P}_{x|x}^n$ are all positive, the Markov chain has a unique stationary distribution that satisfies $\boldsymbol{\pi}' = \boldsymbol{\pi}'\mathbf{P}_{x|x}$. Simple calculations yield

$$\begin{aligned}(p_1+p_3)\Pr(X_1 = 0|X_0 = 0) + (p_2+p_4)\Pr(X_1 = 0|X_0 = 1) \\ = (p_1+p_3)\end{aligned}$$

and

$$(p_1+p_3)\Pr(X_1=1|X_0=0)+(p_2+p_4)\Pr(X_1=1|X_0=1)$$
$$=(p_2+p_4),$$

confirming that $\boldsymbol{\pi}=\begin{bmatrix} p_1+p_3 & p_2+p_4 \end{bmatrix}'$ is the stationary distribution of the Markov chain.

The above Markov chain satisfies the reversibility condition (10.16). Thus,

$$\begin{aligned}\Pr(X_n=1,X_{n-1}=0)&=\Pr(X_n=1|X_{n-1}=0)\Pr(X_{n-1}=0)\\&=\Pr(X_n=0,X_{n-1}=1)\\&=\frac{p_1p_2}{p_1+p_2}+\frac{p_3p_4}{p_3+p_4}.\end{aligned}$$

From the same multinomial model, another Markov chain can be generated by the 4×4 transition probability matrix $\mathbf{P}_{x_ny_n|x_{n-1}y_{n-1}}$ with elements defined as follows:

$$\begin{aligned}&\Pr(X_n=k,Y_n=l|X_{n-1}=i,Y_{n-1}=j)\\&=\Pr(X_n=k|X_{n-1}=i,Y_{n-1}=j)\\&\times\Pr(Y_n=l|X_n=k,X_{n-1}=i,Y_{n-1}=j),\quad(i,j)(k,l)\in S.\end{aligned}\tag{10.18}$$

Instead of using (10.18), consider generating a Markov chain using the following transition probability:

$$\Pr(X_n=k|Y_{n-1}=j)\Pr(Y_n=l|X_n=k),\quad(i,j)(k,l)\in S.\tag{10.19}$$

This creates the Markov chain (X_n,Y_n), $(n=1,2,\ldots)$ whose state space is $S=\{0,1\}^2$ and which consists of correlated Bernoulli random variables. Notice that the chain is formed by a sequence of conditional distributions: first X_n is updated from its conditional distribution, given Y_{n-1}, and this is followed by updating Y_n from its conditional distribution, given the value of X_n from the previous step.

It is easy to verify that this Markov chain has a unique stationary distribution given by $\boldsymbol{\pi}'=(p_1,p_2,p_3,p_4)$, that is the only solution to (10.8). Thus, the chain formed by the sequence of conditional distributions has the joint distribution $\boldsymbol{\pi}$ as its unique stationary distribution. This is an important observation that will be elaborated further in the next chapter.

The Markov chain defined by (10.19) does not satisfy the reversibility condition. To verify this, consider, for example,

$$\begin{aligned}
&\Pr(X_n = 1, Y_n = 1, X_{n-1} = 0, Y_{n-1} = 0) \\
&= \Pr(X_n = 1, Y_n = 1 | X_{n-1} = 0, Y_{n-1} = 0) \\
&\quad \times \Pr(X_{n-1} = 0, Y_{n-1} = 0) \\
&= \Pr(X_n = 1 | Y_{n-1} = 0) \Pr(Y_n = 1 | X_n = 1) \\
&\quad \times \Pr(X_{n-1} = 0, Y_{n-1} = 0) \\
&= p_1 \left(\frac{p_2}{p_1 + p_2} \frac{p_4}{p_2 + p_4} \right). \qquad (10.20)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\Pr(X_n = 0, Y_n = 0, X_{n-1} = 1, Y_{n-1} = 1) \\
&= \Pr(X_n = 0, Y_n = 0 | X_{n-1} = 1, Y_{n-1} = 1) \\
&\quad \times \Pr(X_{n-1} = 1, Y_{n-1} = 1) \\
&= \Pr(X_n = 0 | Y_{n-1} = 1) \Pr(Y_n = 0 | X_n = 0) \\
&\quad \times \Pr(X_{n-1} = 1, Y_{n-1} = 1) \\
&= p_4 \left(\frac{p_1}{p_1 + p_3} \frac{p_3}{p_3 + p_4} \right). \qquad (10.21)
\end{aligned}$$

Since in general, (10.20) $\neq$ (10.21), the ergodic Markov chain defined by (10.19) is not reversible.

The mechanism defined by the transition probability (10.19) updates the variables (X, Y) one at a time in a systematic manner. While (10.19) does not generate a reversible Markov chain, notice that each individual update represents a transition probability that defines a reversible Markov chain. For example

$$\begin{aligned}
\Pr(X_n = 0, Y_{n-1} = 1) &= \Pr(X_n = 0 | Y_{n-1} = 1) \Pr(Y_{n-1} = 1) \\
&= \frac{p_2}{p_1 + p_2} (p_1 + p_2) = p_2,
\end{aligned}$$

which is equal to the time reversed transition

$$\begin{aligned}
\Pr(X_{n-1} = 1, Y_n = 0) &= \Pr(Y_n = 0 | X_{n-1} = 1) \Pr(X_{n-1} = 1) \\
&= \frac{p_2}{p_2 + p_4} (p_2 + p_4) = p_2,
\end{aligned}$$

and this holds for all possible values of (X, Y). As discussed in the next chapter, the Markov chain defined by (10.19) is an example of the systematic scan, single-site Gibbs sampler. ■

10.8 Limiting Behavior of Discrete, Finite State-Spaces, Ergodic Markov Chains

In Section 10.6 it was stated that ergodic Markov chains converge to a unique equilibrium distribution $\boldsymbol{\pi}$ that satisfies

$$\boldsymbol{\pi}' = \boldsymbol{\pi}'\mathbf{P}. \tag{10.22}$$

This implies that $\boldsymbol{\pi}'$ is a left eigenvector with eigenvalue 1. (From linear algebra, recall that if $\boldsymbol{\pi}'$ is a left eigenvector of $\mathbf{P}$, if it satisfies $\boldsymbol{\pi}'\mathbf{P} = \lambda\boldsymbol{\pi}'$, where λ is the associated eigenvalue. Setting $\lambda = 1$ retrieves (10.22), so that an equilibrium distribution must be a left eigenvector with eigenvalue 1. On the other hand, a right eigenvector satisfies $\mathbf{P}\boldsymbol{\pi} = \lambda\boldsymbol{\pi}$). Since the evolution of the Markov chain is governed by its transition probability, insight into the limiting behavior of the Markov chain amounts to studying the properties of $\mathbf{P}^n$ as $n \to \infty$, which is the topic of this final section.

Below, use is made of the Perron-Frobenius theorem from linear algebra (Karlin and Taylor, 1975), which implies that ergodic matrices have one simple eigenvalue equal to 1, and all other eigenvalues have absolute value less than 1. It follows that $\boldsymbol{\pi}$ in (10.22) is the unique stationary distribution of the Markov chain. However, the proof of convergence and rate of convergence of the ergodic Markov chain to $\boldsymbol{\pi}$ requires a little more work.

First suppose that matrix $\mathbf{P}$, of dimension $m \times m$, has distinct eigenvalues $\lambda_i, i = 1, 2, \ldots, m$. Let $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m)$ be the matrix whose columns $\mathbf{c}_i$ represent the corresponding m linearly independent right eigenvectors. Then it is a standard result of matrix theory (Karlin and Taylor, 1975) that

$$\mathbf{P} = \mathbf{C}\mathbf{D}\mathbf{C}^{-1}, \tag{10.23}$$

where the rows of $\mathbf{C}^{-1}$ are the left eigenvectors of $\mathbf{P}$ and $\mathbf{D}$ is a diagonal matrix with diagonal elements consisting of the eigenvalues λ_i. That is,

$$\mathbf{D} = diag\left\{\lambda_1, \lambda_2, \ldots, \lambda_m\right\}.$$

The square of $\mathbf{P}$ is given by

$$\begin{aligned} \mathbf{P}^2 &= \mathbf{C}\mathbf{D}\mathbf{C}^{-1}\mathbf{C}\mathbf{D}\mathbf{C}^{-1} \\ &= \mathbf{C}\mathbf{D}^2\mathbf{C}^{-1}, \end{aligned}$$

and, in general,

$$\mathbf{P}^n = \mathbf{C}\mathbf{D}^n\mathbf{C}^{-1}. \tag{10.24}$$

Consider (10.23). Since $\mathbf{D}$ is diagonal it can be written in the form:

$$\mathbf{D} = \lambda_1\mathbf{I}_1 + \lambda_2\mathbf{I}_2 + \cdots + \lambda_m\mathbf{I}_m,$$

where $\mathbf{I}_i$ is a square matrix with all elements equal to zero except for the term on the diagonal corresponding to the intersection of the i^{th} row and i^{th} column, which is equal to 1. Substituting in (10.23) yields

$$\begin{aligned}\mathbf{P} &= \lambda_1 \mathbf{C}\mathbf{I}_1\mathbf{C}^{-1} + \lambda_2 \mathbf{C}\mathbf{I}_2\mathbf{C}^{-1} + \cdots + \lambda_m \mathbf{C}\mathbf{I}_m\mathbf{C}^{-1} \\ &= \lambda_1 \mathbf{Q}_1 + \lambda_2 \mathbf{Q}_2 + \cdots + \lambda_m \mathbf{Q}_m, \end{aligned} \tag{10.25}$$

where $\mathbf{Q}_i = \mathbf{C}\mathbf{I}_i\mathbf{C}^{-1}$ and has the following properties:

(a) $\mathbf{Q}_i\mathbf{Q}_i = \mathbf{C}\mathbf{I}_i\mathbf{C}^{-1}\mathbf{C}\mathbf{I}_i\mathbf{C}^{-1} = \mathbf{Q}_i, \quad i = 1, 2, \ldots, m;$
(b) $\mathbf{Q}_i\mathbf{Q}_j = \mathbf{C}\mathbf{I}_i\mathbf{C}^{-1}\mathbf{C}\mathbf{I}_j\mathbf{C}^{-1} = \mathbf{C}\mathbf{I}_i\mathbf{I}_j\mathbf{C}^{-1} = \mathbf{0}, \quad i \neq j.$

Using these properties, it follows that (10.24) can be written as

$$\mathbf{P}^n = \lambda_1^n \mathbf{Q}_1 + \lambda_2^n \mathbf{Q}_2 + \cdots + \lambda_m^n \mathbf{Q}_m. \tag{10.26}$$

Without loss of generality, assume that the first term on the right-hand side corresponds to the largest eigenvalue in absolute terms. At this point we assume that $\mathbf{P}$ is aperiodic and irreducible and therefore ergodic. Here we invoke the Perron-Frobenius theorem which allows to set in (10.26), $\lambda_1 = 1$ and $|\lambda_i| < 1$, $i > 1$. Therefore, for large n, $\mathbf{P}^n$ converges to $\mathbf{Q}_1$ which is unique; that is,

$$\lim_{n\to\infty} \mathbf{P}^n = \mathbf{Q}_1,$$

since all terms in (10.26), with the exception of the first one, tend to 0 as n tends to infinity. The rate of convergence to $\mathbf{Q}_1$ depends on the value of λ_i with the largest modulus, other than $\lambda_1 = 1$.

We now show that $\mathbf{Q}_1 = \mathbf{1}\boldsymbol{\pi}'$, where $\mathbf{1}$ is a column vector of $1'$s of dimension m. A little manipulation yields

$$\begin{aligned}\mathbf{Q}_1 &= \mathbf{C}\mathbf{I}_1\mathbf{C}^{-1} \\ &= \begin{bmatrix} c_{11} & 0 & \ldots & 0 \\ c_{21} & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & 0 \\ c_{m1} & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} c^{11} & c^{12} & \ldots & c^{1m} \\ c^{21} & c^{22} & \ldots & c^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c^{m1} & c^{m2} & \ldots & c^{mm} \end{bmatrix}, \end{aligned} \tag{10.27}$$

where c_{i1} $(i = 1, 2, \ldots, m)$ is the ith element of $\mathbf{c}_1$ (the right eigenvector associated with the eigenvalue equal to 1) and c^{ji} is the ith element of the jth row of matrix $\mathbf{C}^{-1}$. The first row of $\mathbf{C}^{-1}$ is the left eigenvector with eigenvalue equal to 1. In connection with (10.22), it was argued that this first row of $\mathbf{C}^{-1}$ is equal to the stationary distribution $\boldsymbol{\pi}$. Now, expanding (10.27) yields

$$\mathbf{Q}_1 = \begin{bmatrix} c_{11}(& c^{11} & c^{12} & \ldots & c^{1m} &) \\ c_{21}(& c^{11} & c^{12} & \ldots & c^{1m} &) \\ & & \vdots & & & \\ c_{m1}(& c^{11} & c^{12} & \ldots & c^{1m} &) \end{bmatrix}.$$

Since the elements of the rows of $\mathbf{Q}_1$ define the stationary distribution, they must add to 1. It follows that

$$c_{11} = c_{21} = \cdots = c_{m1} = w_1, \text{ say.}$$

Hence, $\mathbf{Q}_1$ has identical rows $\widetilde{\boldsymbol{\pi}}'$ say, where $\widetilde{\boldsymbol{\pi}}$ is some equilibrium distribution equal to

$$\widetilde{\boldsymbol{\pi}}' = w_1 \begin{pmatrix} c^{11} & c^{12} & \dots & c^{1m} \end{pmatrix} = w_1 \boldsymbol{\pi}'.$$

We know that the equilibrium distribution is the left eigenvector with eigenvalue 1, with $\sum_{j=1}^{m} c^{1j} = 1$; therefore $w_1 = 1$ and $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}$ is the unique equilibrium distribution.

In the development above it was assumed that $\mathbf{P}$ could be written in the form (10.23), where $\mathbf{D}$ is diagonal. When this is not possible, one can find a matrix $\mathbf{B}$ via a Jordan decomposition that has the form

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \mathbf{M} & \\ 0 & & & \end{bmatrix}$$

such that $\mathbf{P}^n = \mathbf{C}\mathbf{B}^n\mathbf{C}^{-1}$ where $\mathbf{M}^n \to \mathbf{0}$ (Cox and Miller, 1965). Then, in the same way as in the previous case,

$$\lim_{n\to\infty} \mathbf{P}^n = \lim_{n\to\infty} \mathbf{C}\mathbf{B}^n\mathbf{C}^{-1} = \mathbf{C} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \mathbf{0} & \\ 0 & & & \end{bmatrix} \mathbf{C}^{-1} = \mathbf{1}\boldsymbol{\pi}'.$$

Example 10.5 *A three-state ergodic Markov chain*
Consider the stochastic, ergodic matrix in Example 10.1. The matrix has three distinct eigenvalues

$$\lambda_1 = 1, \quad \lambda_2 = \frac{1}{12}\left(4 - \sqrt{10}\right), \quad \lambda_3 = \frac{1}{12}\left(4 + \sqrt{10}\right),$$

and the corresponding right eigenvectors are

$$\begin{aligned} \mathbf{c}_1' &= [1, 1, 1], \\ \mathbf{c}_2' &= \left[\frac{1}{2}\left(1 + \sqrt{10}\right), \frac{1}{4}\left(-4 - \sqrt{10}\right), 1\right], \\ \mathbf{c}_3' &= \left[\frac{1}{2}\left(1 - \sqrt{10}\right), \frac{1}{4}\left(-4 + \sqrt{10}\right), 1\right]. \end{aligned}$$

Forming the diagonal matrix $\mathbf{D} = diag\{\lambda_1, \lambda_2, \lambda_3\}$ with the three eigenvalues, and forming matrix $\mathbf{C}$ with columns corresponding to the three

eigenvectors, it can be verified readily that, up to four decimal places,

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} = \mathbf{CDC}^{-1}$$

$$= \begin{bmatrix} 1 & 2.0811 & -1.0811 \\ 1 & -1.7906 & -0.2094 \\ 1 & 1 & 1 \end{bmatrix} \mathbf{D} \begin{bmatrix} 0.2222 & 0.4444 & 0.3333 \\ 0.1699 & -0.2925 & 0.1225 \\ -0.3922 & -0.1519 & 0.5442 \end{bmatrix}.$$

Further, expression (10.26) is of the following form:

$$\mathbf{P}^n = 1^n \begin{bmatrix} 0.2222 & 0.4444 & 0.3333 \\ 0.2222 & 0.4444 & 0.3333 \\ 0.2222 & 0.4444 & 0.3333 \end{bmatrix}$$
$$+ \left(\frac{1}{12}\left(4 - \sqrt{10}\right)\right)^n \mathbf{Q}_2 + \left(\frac{1}{12}\left(4 + \sqrt{10}\right)\right)^n \mathbf{Q}_3,$$

which shows that the two transient terms in the second line tend rapidly to zero as n increases. ■

Example 10.6 *A periodic, irreducible Markov chain*
Consider the chain introduced in Example 10.2 with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

This matrix has period $d = 2$. The five distinct eigenvalues are

$$\lambda_1 = -1, \quad \lambda_2 = 0, \quad \lambda_3 = 1, \quad \lambda_4 = -\frac{1}{\sqrt{2}}, \quad \lambda_5 = \frac{1}{\sqrt{2}}.$$

The matrix $\mathbf{C}$ whose columns are the corresponding right eigenvectors is

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 \\ -1 & 0 & 1 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 1 & -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Forming the diagonal matrix $\mathbf{D}$ whose diagonal elements are the five eigenvalues $\lambda_1, \ldots, \lambda_5$, one can verify result (10.23).

For this example, (10.26) takes the following form:

$$\mathbf{P}^n = (-1)^n \begin{bmatrix} 0.125 & -0.25 & 0.25 & -0.25 & 0.125 \\ -0.125 & 0.25 & -0.25 & 0.25 & -0.125 \\ 0.125 & -0.25 & 0.25 & -0.25 & 0.125 \\ -0.125 & 0.25 & -0.25 & 0.25 & -0.125 \\ 0.125 & -0.25 & 0.25 & -0.25 & 0.125 \end{bmatrix}$$
$$+ 1^n \begin{bmatrix} 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \\ 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \\ 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \\ 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \\ 0.125 & 0.25 & 0.25 & 0.25 & 0.125 \end{bmatrix}$$
$$+ \left(-\frac{1}{\sqrt{2}}\right)^n \mathbf{Q}_4 + \left(\frac{1}{\sqrt{2}}\right)^n \mathbf{Q}_5,$$

which shows that, for large n, the eigenvectors associated with the eigenvalues with absolute value less than 1 become irrelevant. Also, as illustrated before, the value of $\mathbf{P}^n$ varies depending on whether n is even or odd. The unique stationary distribution for this Markov chain exists, but the chain fails to converge to it. ■

The next chapter discusses MCMC methods. Here, one creates a Markov chain using random number generators. This particular chain is constructed in such a way that, eventually, the numbers drawn can be interpreted as samples from the stationary distribution $\boldsymbol{\pi}$, which is often a posterior distribution. The key to the right construction process is the choice of transition kernel which governs the behavior of the system. Many transition kernels can be used for a particular problem, and, typically, they are all derived by imposing the condition of reversibility. Then provided that the Markov chain is ergodic, and only then, a reversible system has $\boldsymbol{\pi}$ as its unique stationary distribution. Convergence to this distribution however, takes place asymptotically. Many factors, such as the choice of transition kernel, the degree of correlation of the parameters of the model in their posterior distributions, and the structure of the data available, can interfere with a smooth trip towards the final goal. This general area is discussed an illustrated in the chapters ahead.

11
Markov Chain Monte Carlo

11.1 Introduction

Markov chain Monte Carlo (MCMC) has become a very important computational tool in Bayesian statistics, since it allows inferences to be drawn from complex posterior distributions where analytical or numerical integration techniques cannot be applied. The idea underlying these methods is to generate a Markov chain via iterative Monte Carlo simulation that has, at least in an asymptotic sense, the desired posterior distribution as its equilibrium or stationary distribution. An important paper in this area is Tierney (1994).

The classic papers of Metropolis et al. (1953) and of Hastings (1970) gave rise to a very general MCMC method, namely the Metropolis–Hastings algorithm, of which the Gibbs sampler, which was introduced in statistics and given its name by Geman and Geman (1984), is a special case. In Hastings (1970), the algorithm is used for the simulation of discrete equilibrium distributions on a space of fixed dimension. In a statistical setting, fixed dimensionality implies that the number of parameters of the model is a known value. The reversible jump MCMC algorithm introduced in Green (1995) is a generalization of Metropolis–Hastings; it has been applied for simulation of equilibrium distributions on spaces of varying dimension. The number of parameters of a model is inferred via its marginal posterior distribution.

In this chapter, the focus is on these MCMC algorithms. Much of the material is drawn from the tutorial paper of Waagepetersen and Sorensen

(2001). Markov chains with continuous state spaces are considered here; that is, the random variables are assumed to be distributed continuously. Many results for discrete state spaces discussed in Chapter 10 hold also, but some technical modifications are needed for Markov chains with continuous state spaces. For example, convergence to a unique stationary distribution requires the continuous Markov chain to be irreducible and aperiodic, as for discrete chains. In addition, a continuous Markov chains must be "Harris recurrent". This is a measure theoretical technicality required to avoid the possibility of starting points from which convergence is not assured. This property shall not be discussed here, and the reader is referred to Meyn and Tweedie (1993), to Tierney (1994), and to Robert and Casella (1999).

This chapter is organized as follows. The following section introduces notation and terminology, since the continuous state space situation requires some additional details. Subsequently, an overview of MCMC is presented, including a description of the Metropolis–Hastings, Gibbs sampling, Langevin-Hastings and reversible jump methods. The chapter ends with a short description of the data augmentation strategy, a technique that may lead to simpler computational expressions in the process of fitting a probability model via MCMC.

11.2 Preliminaries

11.2.1 Notation

The notation in this chapter is the standard one found in the literature on MCMC, and it deviates somewhat from that used in the rest of this book. Here no notational distinction is made between scalar and vector random variables, but this should not lead to ambiguities. In this chapter, the probability of the event E is denoted by $P(E)$ rather than by $\Pr(E)$.

Suppose that $Z = (Z_1, \ldots, Z_d)$ is a real random vector of dimension $d \geq 1$. Resuming the notation introduced in (1.1), we shall say that Z has density f on $\mathbb{R}^d$ if the probability that Z belongs to a subset A of $\mathbb{R}^d$ is given by the d-fold integral

$$P(Z \in A) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I((z_1, \ldots, z_d) \in A) f(z_1, \ldots, z_d)\, dz_1 \cdots dz_d,$$

where the indicator function $I((z_1, \ldots, z_d) \in A)$ is one if $(z_1, \ldots, z_d)$ is in A, and zero otherwise. If, for example, $A = [a_1, b_1] \times [a_2, b_2]$, then the integral is

$$\int \int I(x \in [a_1, b_1], y \in [a_2, b_2]) f(x, y)\, dx\, dy = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, dx\, dy.$$

Integrals will usually be written in abbreviated form

$$P(Z \in A) = \int I(z \in A) f(z)\, dz = \int_A f(z)\, dz$$

whenever the dimension of the integral is clear from the context.

11.2.2 Transition Kernels

In the case of Markov chains with continuous state spaces, it does not make sense to write $p(i,j) = \Pr(X_n = j | X_{n-1} = i)$, since this probability is always null. Further, the fact that X is continuously distributed precludes construction of transition probability matrices and transition kernels are used instead. Assume that X is a p-dimensional stochastic vector with density f on $\mathbb{R}^p$ and assume that A is a subset in $\mathbb{R}^p$. Then the transition kernel $\mathsf{P}(\cdot,\cdot)$ specifies the conditional probability that X_{n+1} falls in A given that $X_n = x$ and will be written as

$$\mathsf{P}(x, A) = P(X_{n+1} \in A | X_n = x). \tag{11.1}$$

This notation (departing from what has been used before in the book) is required for reasons mentioned below.

11.2.3 Varying Dimensionality

Let Y be a stochastic vector in $\mathbb{R}^q$. In general, the joint probability that $X \in A$ and $Y \in B$ will be written as

$$P(X \in A, Y \in B) = \int_A \mathsf{P}(x, B)\, f(x)\, dx \tag{11.2}$$

for all subsets $A \subseteq \mathbb{R}^p$ and $B \subseteq \mathbb{R}^q$. If Y has density in $\mathbb{R}^q$, then (X, Y) has joint density h on $\mathbb{R}^{p+q}$. In this case, (11.2) is the well-known identity

$$P(X \in A, Y \in B) = \int_A \int_B h(x, y)\, dx\, dy. \tag{11.3}$$

However, as discussed in the sections describing single-site Metropolis–Hastings updates and reversible jump samplers, there are situations when the q-dimensional stochastic vector Y has density on $\mathbb{R}^{q'}$, $q' < q$. In this case, (X, Y) does not have density h on $\mathbb{R}^{p+q}$ and, consequently, the joint probability $P(X \in A, Y \in B)$ cannot be calculated from (11.3) and one needs to refer to (11.2). As an example of the latter, consider computing for the two-dimensional vector $Y = (Y_1, Y_2)'$ and the scalar variable X:

$$P(Y \in B | X = x) \tag{11.4}$$

Suppose that given $X = x$, Y is defined by the deterministic mapping g

$$(Y_1, Y_2) = g(x, U) = (x + U, x - U),$$

where U is a stochastic one-dimensional random variable with density q on $\mathbb{R}$. Given x, Y does not have density on $\mathbb{R}^2$ since it can only take values on the line $y_1 = 2x - y_2$; that is, once y_2 is known, the value of y_1 is determined completely. Then (11.4) must be computed from the one-dimensional integral

$$P(Y \in B | X = x) = \mathsf{P}(x, B) = \int I((x + U, x - U) \in B)\, q(u)\, du$$

and $P(X \in A, Y \in B)$, $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}^2$, is computed from (11.2):

$$P(X \in A, Y \in B) = \int_A \int I((x + U, x - U) \in B)\, q(u)\, f(x)\, du\, dx.$$

If $B = [a_1, b_1] \times [a_2, b_2]$, this becomes

$$\begin{aligned} & P(X \in A, Y \in B) \\ = & \int_A \int I(x + U \in [a_1, b_1], x - U \in [a_2, b_2])\, q(u)\, f(x)\, du\, dx. \end{aligned}$$

On the other hand, if $[Y_1, Y_2, X]$ had density h on $\mathbb{R}^3$, one could use the standard formula

$$P(X \in A, Y \in B) = \int_A \int_{a_1}^{b_1} \int_{a_2}^{b^2} h(x, y_1, y_2)\, dx\, dy_1\, dy_2.$$

11.3 An Overview of Markov Chain Monte Carlo

The MCMC algorithms to be described below are devices for constructing a Markov chain that has the desired equilibrium distribution as its limiting invariant distribution. In the previous chapter on Markov chains, the transition probability matrix $\mathbf{P}$ was known, and one wished to characterize the equilibrium distribution. The nature of the problem with MCMC methods is the opposite one: the equilibrium distribution is known (usually up to proportionality), but the transition kernel is unknown. How does one construct a transition kernel that generates a Markov chain with a desired stationary distribution?

Let X denote a real stochastic vector of unknown parameters or unobservable variables associated with some model, and assume it has a distribution with density π on $\mathbb{R}^d$. The density π could represent a posterior density. This density has typically a complex form, such that the necessary

expectations with respect to π cannot be evaluated analytically or by using standard techniques for numerical integration. In particular, π may only be known up to an unknown normalizing constant. Direct simulation of π may be difficult, but as shown below, it is usually quite easy to construct a Markov chain whose invariant distribution has density given by π. The Markov chain $X_1, X_2, \ldots$ is specified in terms of the distribution for the initial state X_1 and the transition kernel $\mathsf{P}(\cdot, \cdot)$ which specifies the conditional distribution of X_{i+1} given the previous state X_i. If the value of the current state is $X_i = x$, then the probability that X_{i+1} is in a set $A \subseteq \mathbb{R}^d$ is given by (11.1). If the generated Markov chain is irreducible with invariant distribution π, it can be used for Monte Carlo estimation of various expectations $E(h(X))$ with respect to π (Tierney, 1994; Meyn and Tweedie, 1993). That is, for any function h on $\mathbb{R}^d$ with finite expectation $E(h(X))$,

$$E(h(X)) = \int h(x)\,\pi(x)\,dx = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} h(X_i). \tag{11.5}$$

Thus, $E(h(X))$ can be approximated by

$$E(h(X)) \approx \frac{\sum_{i=1}^{N} h(X_i)}{N}, \tag{11.6}$$

the sample average, for some large N. Here h could be the indicator function of a set $A \subseteq \mathbb{R}^d$, so that $E(h(X))$ equals the probability

$$P(X \in A) = E(I(X \in A)),$$

which is approximated by $\sum_{i=1}^{N} I(X_i \in A)/N$. (Convergence of the sample average to $E(h(X))$ for all starting values, also requires the assumption of Harris recurrence). Result (11.5) does not rest upon the condition of aperiodicity, even though this is required for convergence of the Markov chain to π.

As discussed in the previous chapter, the density π is invariant for the Markov chain if the transition kernel $\mathsf{P}(\cdot, \cdot)$ of the Markov chain preserves π, i.e., if $X_i \sim \pi$ implies $X_{i+1} \sim \pi$ or, in terms of $\mathsf{P}(\cdot, \cdot)$ and π, if

$$\int_{\mathbb{R}^d} \mathsf{P}(x, B)\,\pi(x)\,dx = \int_B \pi(x)\,dx. \tag{11.7}$$

Both terms of this expression are equal to the marginal probability that $X \in B$, provided that π is the invariant density. In order to verify that π is the invariant density using (11.7) is a difficult task, since this involves integration with respect to π. The infeasibility of doing this was the reason for using MCMC in the first place. However, choosing a transition kernel which imposes the stronger condition of reversibility with respect to π is

sufficient to guarantee that π is invariant for the Markov chain. In the previous chapter it was indicated that reversibility holds if (X_i, X_{i+1}) has the same distribution as the time-reversed subchain (X_{i+1}, X_i) whenever X_i has density π, i.e., if

$$\begin{aligned} P\left(X_{i+1} \in A, X_i \in B\right) &= \int_B \mathsf{P}\left(x, A\right) \pi\left(x\right) dx \\ = P\left(X_{i+1} \in B, X_i \in A\right) &= \int_A \mathsf{P}\left(x, B\right) \pi\left(x\right) dx \end{aligned} \tag{11.8}$$

for subsets $A, B \subseteq \mathbb{R}^d$. Taking $A = \mathbb{R}^d$, the integral in the first line becomes $\int_B \mathsf{P}\left(x, \mathbb{R}^d\right) \pi\left(x\right) dx = \int_B \pi\left(x\right) dx$, since $\mathsf{P}\left(x, \mathbb{R}^d\right) = 1$. Therefore the reversibility condition (11.8) implies (11.7). The Metropolis–Hastings algorithm or, more generally, the reversible jump MCMC, offers a practical recipe for constructing a reversible Markov chain with the desired invariant distribution by an appropriate choice of a transition kernel.

11.4 The Metropolis–Hastings Algorithm

11.4.1 An Informal Derivation

An informal and intuitive derivation of the Metropolis–Hastings algorithm will be given first. This is followed by a more formal approach which eases the way into reversible jump MCMC.

Since direct sampling from the target distribution may not be possible, the Metropolis–Hastings algorithm starts by generating candidate draws from a so-called proposal distribution. These draws are then "corrected" so that they behave, asymptotically, as random observations from the desired equilibrium or target distribution. The Markov chain produced by the Metropolis–Hastings algorithm at each stage is thus constructed in two steps: a proposal step and an acceptance step. These two steps are associated with the proposal distribution and with the acceptance probability which are the two ingredients of the Metropolis–Hastings transition kernel. Assume that at stage n the state of the chain is $X_n = x$. The random variable X may be a scalar or a vector. The next state of the chain is chosen by first sampling a candidate point $Y_{n+1} = y$ from a proposal distribution with p.d.f. $q\left(x, \cdot\right)$. The density $q\left(x, \cdot\right)$ may (or may not) depend on the current point x. If it does, $q\left(x, y\right)$ is a conditional p.d.f. of y, given x, which in other parts of this text has been referred to as $p\left(y|x\right)$. For example, $q\left(x, \cdot\right)$ may be a multivariate normal distribution with mean x and fixed covariance matrix. The candidate point y is then accepted with probability $a\left(x, y\right)$ to be derived below. If y is accepted, then the next state becomes $X_{n+1} = y$. If the candidate point is rejected, the chain does not move, i.e., $X_{n+1} = x$. The algorithm is extremely simple:

INITIALIZE X_0
DO $i = 1, n$
SAMPLE Y FROM $q(x, \cdot)$
SAMPLE U FROM $Un(0, 1)$
IF $(U \leq a(X, Y))$ SET $X_i = y$
OTHERWISE SET $X_i = x$
ENDDO

We turn to the informal derivation of the acceptance probability $a(x, y)$, drawing from the tutorial of Chib and Greenberg (1995). For a joint updating Metropolis–Hastings algorithm in equilibrium, the random vector (X_n, Y_{n+1}), consisting of the current Markov chain state and the proposal, has joint density g, given by

$$g(x, y) = q(x, y)\, \pi(x), \tag{11.9}$$

where π is the equilibrium density and q is the proposal density of Y_{n+1}, given that X_n has value x. If q satisfies the reversibility condition such that

$$q(x, y)\, \pi(x) = q(y, x)\, \pi(y) \tag{11.10}$$

for all (x, y), then the proposal density is the correct transition kernel of the Metropolis–Hastings chain. Most likely it will be the case that, for some (x, y),

$$q(x, y)\, \pi(x) > q(y, x)\, \pi(y), \tag{11.11}$$

say. In order to achieve equality and thus ensuring reversibility, one can introduce a probability $a(x, y) < 1$ on the left-hand side, such that transitions from x to y $(x \neq y)$ are made according to $q(x, y)\, a(x, y)$. Setting $a(y, x) = 1$ on the right-hand side yields

$$\begin{aligned} q(x, y)\, \pi(x)\, a(x, y) &= q(y, x)\, \pi(y)\, a(y, x) \\ &= q(y, x)\, \pi(y) \end{aligned}$$

from which the acceptance probability becomes

$$a(x, y) = \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}. \tag{11.12}$$

If inequality (11.11) is reversed, a probability $a(y, x)$ is introduced appropriately and derived as above, after setting $a(x, y) = 1$. The probabilities a guarantee that detailed balance is satisfied. These arguments imply that the probability of acceptance must be

$$a(x, y) = \min\left(1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right), \quad \pi(x)\, q(x, y) > 0. \tag{11.13}$$

Notice that if (11.10) holds, $a(x, y) = 1$ and the candidate draw y is accepted. This is equivalent to sampling the candidate point from the equilibrium distribution.

A key observation is that the posterior distribution π must be known up to proportionality only, since the normalizing constant cancels in the ratio $\pi(y)/\pi(x)$. Finally, it is clear from (11.13) that when symmetrical proposal densities are considered with $q(y,x) = q(x,y)$, the acceptance probability reduces to

$$a(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right), \quad \pi(x) > 0,$$

which is the case originally considered by Metropolis et al. (1953).

A few comments about implementation are in order. The acceptance probability is clearly defined provided $\pi(x)\, q(x,y) > 0$. If the chain starts with a value x_0 such that $\pi(x_0) > 0$, then $\pi(x_n) > 0$ for every n, since values of the proposal Y for which $\pi(y) = 0$ lead to an acceptance probability equal to zero, and are therefore rejected. The success of the method depends on striking some "right" rate of acceptance. Parameterization and choice of the proposal density play a fundamental role here. Acceptance ratios in the neighborhood of 1 imply very similar values between previous and proposed states and the chain will move very slowly (unless, of course, the proposal distribution is the equilibrium distribution, leading to an acceptance ratio of 1!). On the other hand, if the proposed displacement is too large and falls where the posterior has no support, this will lead to a high rejection rate. Here, the chain will remain in the same state for many iterations. No general optimization rules are available. However, Roberts et al. (1997) obtained optimal acceptance rates of 23.4% for high-dimensional problems under quite general conditions. Guidance on the proper choice of proposal densities can be found, for example, in Chib and Greenberg (1995).

11.4.2 A More Formal Derivation

A more formal derivation of the acceptance probability (11.13) is given here; this will be useful in understanding reversible jump MCMC later on. A distinction is made between two implementations of the Metropolis–Hastings algorithm. These are the simultaneous and the single-site updating schemes. In the former scheme all the random variables of the model are updated jointly, whereas in the latter, random variables are updated one at a time. In the single-site updating strategy, and in common with reversible jump MCMC, the proposal distribution and the target distribution have densities on spaces of different dimension. The informal derivation of the previous section was based on the simultaneous updating algorithm.

Metropolis–Hastings Simultaneous Updating Algorithm

As before, let X_n denote the $n$the state of a Metropolis–Hastings chain $X_1, X_2, \ldots$ and let Y_{n+1} denote the proposal for the next state of the chain. The random vector (X_n, Y_{n+1}) has joint density g on $\mathbb{R}^{2d}$ given by (11.9), where π is the d-dimensional target density and $q(x,\cdot)$ is the $d-$dimensional

proposal density of Y_{n+1}, given that X has the value $x \in \mathbb{R}^d$. In this section the acceptance probability of the simultaneous updating Metropolis–Hastings algorithm is derived subject to the reversibility condition

$$P\left(X_n \in A, X_{n+1} \in B\right) = P\left(X_n \in B, X_{n+1} \in A\right) \tag{11.14}$$

for all $A, B \subseteq \mathbb{R}^d$. The left-hand side of (11.14) can be written as

$$P\left(X_n \in A, X_{n+1} \in B\right) = \int_A P\left(X_{n+1} \in B | X_n = x\right) \pi\left(x\right) dx. \tag{11.15}$$

For any $B \subseteq \mathbb{R}^d$ define the proposal distribution as

$$\mathsf{Q}\left(x, B\right) = P\left(Y_{n+1} \in B | X_n = x\right) = \int I\left(y \in B\right) q\left(x, y\right) dy \tag{11.16}$$

which is the conditional probability that Y_{n+1} belongs in a set B, given that $X_n = x$. Also define

$$\begin{aligned} \mathsf{Q}^a\left(x, B\right) &= P\left(Y_{n+1} \in B \text{ and } Y_{n+1} \text{ is accepted} | X_n = x\right) \\ &= \int I\left(y \in B\right) q\left(x, y\right) a\left(x, y\right) dy \end{aligned} \tag{11.17}$$

as the conditional probability that Y_{n+1} belongs in a set B and Y_{n+1} is accepted, given that $X_n = x$ and define further

$$s\left(x\right) = P\left(Y_{n+1} \text{ is rejected} | X_n = x\right) \tag{11.18}$$

as the conditional probability of rejecting the proposal, given that $X_n = x$. Then the transition kernel $P\left(X_{n+1} \in B | X_n = x\right)$ can be written as

$$P\left(X_{n+1} \in B | X_n = x\right) = \mathsf{Q}^a\left(x, B\right) + s\left(x\right) I\left(x \in B\right). \tag{11.19}$$

This is by virtue of the law of total probability, since there are two ways in which $X_{n+1} \in B$. One is generating a proposal Y_{n+1} that belongs in B and that this candidate is accepted; the other one is rejecting the proposal, so that the new state $X_{n+1} = X_n = x$, and that $x \in B$. Hence (11.15) becomes equal to

$$\begin{aligned} &P\left(X_n \in A, X_{n+1} \in B\right) \\ &= \int_A \mathsf{Q}^a\left(x, B\right) \pi\left(x\right) dx + \int_A s\left(x\right) I\left(x \in B\right) \pi\left(x\right) dx. \end{aligned} \tag{11.20}$$

By virtue of symmetry, the right-hand side of (11.14) is given by

$$\begin{aligned} &P\left(X_n \in B, X_{n+1} \in A\right) \\ &= \int_B \mathsf{Q}^a\left(x', A\right) \pi\left(x'\right) dx' + \int_B s\left(x'\right) I\left(x' \in A\right) \pi\left(x'\right) dx'. \end{aligned} \tag{11.21}$$

The second term on the right-hand side of (11.20) can be written as

$$\int_A s(x) I(x \in B) \pi(x) dx = \int s(x) I(x \in B \cap A) \pi(x) dx \qquad (11.22)$$

and the second term on the right-hand side of (11.21) can similarly be written as

$$\int_B s(x') I(x' \in A) \pi(x') dx' = \int s(x') I(x' \in B \cap A) \pi(x') dx'. \quad (11.23)$$

Clearly, (11.22) = (11.23) (notice that x and x' are dummy variables of integration), and therefore the reversibility condition is satisfied if

$$\int_A \mathsf{Q}^a(x, B) \pi(x) dx = \int_B \mathsf{Q}^a(x', A) \pi(x') dx'. \qquad (11.24)$$

Using definition (11.17) for $\mathsf{Q}^a(x, B)$, the left-hand side of (11.24) can be written more explicitly as

$$\begin{aligned} \int_A \mathsf{Q}^a(x, B) \pi(x) dx &= \int_A \int I(y \in B) q(x, y) a(x, y) \pi(x) dx dy \\ &= \int \int I(x \in A, y \in B) q(x, y) a(x, y) \pi(x) dx dy, \end{aligned} \qquad (11.25)$$

and, similarly, for the right-hand side of (11.24):

$$\begin{aligned} &\int_B \mathsf{Q}^a(x', A) \pi(x') dx' \\ &= \int \int I(x' \in B, y' \in A) q(x', y') a(x', y') \pi(x') dx' dy'. \end{aligned} \qquad (11.26)$$

In order to write (11.25) and (11.26) explicitly as functions of the same variables, one can make the change of variables $(y = x')$ and $(x = y')$ – see the Note below. Since the Jacobian of the transformation is 1, the right-hand side of (11.26) now becomes

$$\int \int I(y \in B, x \in A) q(y, x) a(y, x) \pi(y) dy dx.$$

Substituting in (11.24) yields the following expression for the reversibility condition:

$$\begin{aligned} &\int \int I(x \in A, y \in B) q(x, y) a(x, y) \pi(x) dx dy \\ = &\int \int I(y \in B, x \in A) q(y, x) a(y, x) \pi(y) dy dx. \end{aligned}$$

Equality is satisfied if

$$q(x, y) a(x, y) \pi(x) = q(y, x) a(y, x) \pi(y).$$

Choosing the acceptance probability as large as possible (i.e., setting $a(y, x)$ equal to 1) subject to detailed balance, as suggested by Peskun (1973) yields

$$a(x, y) = \min\left(1, \frac{q(y, x)\,\pi(y)}{q(x, y)\,\pi(x)}\right). \tag{11.27}$$

Note In moving from $X_n = x$ to $X_{n+1} = x'$, a proposal with realized value y is generated from $q(x, \cdot)$. If the proposal is accepted, $x' = y$. In the opposite move, from $X_n = x'$ to $X_{n+1} = x$, the proposal with realized value y' is generated from $q(x', \cdot)$. If the proposal is accepted, $x = y'$.

Metropolis–Hastings Single-Site Updating Algorithm

In the single-site updating algorithm, only one component of $X_n \in \mathbb{R}^d$ is updated at a time. Then, given that $X_n = x$, Y_{n+1} equals x except at the $i$the component, where x_i is replaced by a random variable Z_i generated from a one-dimensional proposal density $q_i(x, \cdot)$ that may or may not depend on x or on a subset of x. Since

$$Y_{n+1} \in B \Leftrightarrow (x_1, \ldots, x_{i-1}, Z_i, x_{i+1}, \ldots, x_d) \in B,$$

then the probability that Y_{n+1} belongs in $B \subseteq \mathbb{R}^d$, given $X_n = x$, is given by the proposal distribution

$$\begin{aligned} \mathsf{Q}(x, B) &= P(Y_{n+1} \in B | X_n = x) \\ &= \int I((x_1, \ldots, x_{i-1}, z_i, x_{i+1}, \ldots, x_d) \in B)\, q_i(x, z_i)\, dz_i \end{aligned} \tag{11.28}$$

which is a one-dimensional integral. Notice that target density π lives on $\mathbb{R}^d$, while the proposal density $q_i(x, \cdot)$ lives on $\mathbb{R}$.

Consider the move from a realized value equal to

$$x = (x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$$

and to a realized value equal to

$$x' = (x_1, \ldots, x_{i-1}, z_i, x_{i+1}, \ldots, x_d).$$

Note that in the notation used here, contrary to the case in the previous section, x' differs from x in one element only. The probability that X_{n+1} belongs in $B \subseteq \mathbb{R}^d$, given $X_n = x$, is given by

$$P(X_{n+1} \in B | X_n = x) = \mathsf{Q}^a(x, B) + s(x)\, I(x \in B), \tag{11.29}$$

where $\mathsf{Q}^a(x, B) = P(Y_{n+1} \in B \text{ and } Y_{n+1} \text{ is accepted} | X_n = x)$ is here equal to

$$\mathsf{Q}^a(x, B) = \int I(x' \in B)\, a(x, x')\, q_i(x, z_i)\, dz_i. \tag{11.30}$$

The second term in (11.29) is $s(x) = P(Y_{n+1} \text{ is rejected} | X_n = x)$.

The left-hand side of the reversibility equation (11.14) can be written as follows:

$$P(X_n \in A, X_{n+1} \in B) = \int_A \mathsf{Q}^a(x, B)\, \pi(x)\, dx + \int_A s(x)\, I(x \in B)\, \pi(x)\, dx.$$

Similarly, for the move in the opposite direction, the right-hand side of (11.14) is given by

$$P(X_n \in B, X_{n+1} \in A) = \int_B \mathsf{Q}^a(x', A)\, \pi(x')\, dx' + \int_B s(x')\, I(x' \in A)\, \pi(x')\, dx',$$

where X_n has the realized value x'. As in the previous section, the second terms in the right-hand side of these expressions can be shown to be equal. Therefore, reversibility is satisfied if

$$\int_A \mathsf{Q}^a(x, B)\, \pi(x)\, dx = \int_B \mathsf{Q}^a(x', A)\, \pi(x')\, dx', \tag{11.31}$$

where

$$\begin{aligned} \mathsf{Q}^a(x', A) &= P(Y_{n+1} \in A \text{ and } Y_{n+1} \text{ is accepted} | X_n = x') \\ &= \int I(x \in A)\, q_i(x', x_i)\, a(x', x)\, dx_i. \end{aligned} \tag{11.32}$$

Substituting (11.30) in the left-hand side of (11.31) yields

$$\begin{aligned} &\int_A \int I(x' \in B)\, q_i(x, z_i)\, a(x, x')\, \pi(x)\, dz_i\, dx \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} I(x' \in B, x \in A)\, q_i(x, z_i)\, a(x, x')\, \pi(x)\, dz_i\, dx. \end{aligned} \tag{11.33}$$

Similarly, substituting (11.32) in the right-hand side of (11.31) yields:

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}} I(x \in A, x' \in B)\, q_i(x', x_i)\, a(x', x)\, \pi(x')\, dx'\, dx_i. \tag{11.34}$$

A condition for equality of (11.33) and (11.34) is that

$$q_i(x, z_i)\, a(x, x')\, \pi(x) = q_i(x', x_i)\, a(x', x)\, \pi(x').$$

As in (11.27), using the criterion due to Peskun (1973) yields the acceptance probability

$$a(x, x') = \min\left(1, \frac{q_i(x', x_i)\, \pi(x')}{q_i(x, z_i)\, \pi(x)}\right). \tag{11.35}$$

The arguments above also hold when the updating variable, rather than being a scalar, is a vector and a subset of x.

11.5 The Gibbs Sampler

The Gibbs sampler is a very popular MCMC algorithm because of its computational simplicity. As shown below, it is a special case of the Metropolis–Hastings algorithm. In order to see this, in the move from x to x', let the proposal be generated from

$$q_i(x, z_i) = \pi(z_i|x_{-i}), \tag{11.36}$$

where x_{-i} is equal to x with its ith component deleted, that is, $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. Similarly, in the opposite move, let the proposal be a draw from

$$q_i(x', x_i) = \pi(x_i|x_{-i}). \tag{11.37}$$

In a Bayesian context, the right-hand sides of (11.36) and (11.37) are known as the fully conditional posterior distributions. Since

$$\pi(z_i|x_{-i}) = \pi(x') / \pi(x_{-i})$$

and

$$\pi(x_i|x_{-i}) = \pi(x) / \pi(x_{-i}),$$

substituting in (11.35) yields

$$\frac{q_i(x', x_i)\,\pi(x')}{q_i(x, z_i)\,\pi(x)} = \frac{\pi(x)\,\pi(x')\,\pi(x_{-i})}{\pi(x)\,\pi(x')\,\pi(x_{-i})} = 1.$$

Therefore, a Metropolis–Hastings proposal generated from the appropriate fully conditional distribution (a conditional posterior distribution in the Bayesian setting), is accepted always. This scheme is known as Gibbs sampling.

The transition kernel of the Gibbs sampler for the updating of all elements of x involves the product

$$\pi(z_1|x_2, x_3, \ldots, x_d)\,\pi(z_2|z_1, x_3, \ldots, x_d) \ldots \pi(z_d|z_1, z_2, \ldots, z_{d-1}).$$

The transition kernel of the Gibbs sampler preserves π. To see this and for notational simplicity, let $d = 3$. Then, in terms of (11.7):

$$\mathsf{P}(x, B) = \int I(z_1, z_2, z_3 \in B)\,\pi(z_1|x_2, x_3)\,\pi(z_2|z_1, x_3) \times \pi(z_3|z_1, z_2)\,dz_1\,dz_2\,dz_3,$$

where it is clear from the context that the integral is three-dimensional. Substituting in the left-hand side of (11.7) and integrating over the distribution of x_1, x_2, and x_3 yields

$$\begin{aligned} &\int \mathsf{P}(x, B)\,\pi(x_1, x_2, x_3)\,dx_1\,dx_2\,dx_3 \\ &= \int I(z_1, z_2, z_3 \in B)\,\pi(z_1, z_2, z_3)\,dz_1\,dz_2\,dz_3 \\ &= P(X_1, X_2, X_3 \in B). \end{aligned}$$

Equation (11.7) is also satisfied if the transition kernel is defined with respect to only one of the elements of x. To verify this, and in terms of this three-dimensional example, the transition kernel for the updating of the first element of x is now

$$\mathsf{P}(x, B) = \int I(z_1, x_2, x_3 \in B)\, \pi(z_1 | x_2, x_3)\, dz_1$$

and the left-hand side of equation (11.7) is

$$\int \mathsf{P}(x, B)\, \pi(x)\, dx$$
$$= \int\int I(z_1, x_2, x_3 \in B)\, \pi(z_1 | x_2, x_3)\, \pi(x_1, x_2, x_3)\, dz_1\, dx_1\, dx_2\, dx_3.$$

Integrating over the distribution of x_1 yields

$$\int I(z_1, x_2, x_3 \in B)\, \pi(z_1, x_2, x_3)\, dz_1\, dx_2\, dx_3$$
$$= P(X_1, X_2, X_3 \in B).$$

11.5.1 Fully Conditional Posterior Distributions

Consider the vector of parameters $(\theta_1, \theta_2, \ldots, \theta_p)$ whose posterior distribution is proportional to $p(\theta_1, \ldots, \theta_{i-1}, \theta_i, \theta_{i+1}, \ldots, \theta_p | \mathbf{y})$. Let

$$\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$$

be the vector of dimension $(p - r)$, $p > r$, $r \geq 1$, which is equal to $\boldsymbol{\theta}$ with its ith component, $\boldsymbol{\theta}_i$, deleted, and where r is the number of elements in $\boldsymbol{\theta}_i$. The fully conditional posterior distribution of $\boldsymbol{\theta}_i$ is

$$p(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \mathbf{y}) = \frac{p(\theta_1, \ldots, \theta_{i-1}, \boldsymbol{\theta}_i, \theta_{i+1}, \ldots, \theta_p | \mathbf{y})}{\int p(\theta_1, \ldots, \theta_{i-1}, \boldsymbol{\theta}_i, \theta_{i+1}, \ldots, \theta_p | \mathbf{y})\, d\boldsymbol{\theta}_i}$$
$$\propto p(\theta_1, \ldots, \theta_{i-1}, \boldsymbol{\theta}_i, \theta_{i+1}, \ldots, \theta_p | \mathbf{y}). \tag{11.38}$$

In many applications, $r = 1$ and parameters are updated one at a time. In general, single-site updating leads to moves along each coordinate, whereas updating several components in a block allows for more general moves. Joint updating, which incorporates information on the correlation structure among the components in the joint conditional posterior distribution, can result in faster convergence when correlations are strong (Liu et al., 1994).

11.5.2 The Gibbs Sampling Algorithm

Consider the vector of parameters of a model $(\theta_1, \theta_2, \ldots, \theta_p)$, with posterior density $p(\theta_1, \theta_2, \ldots, \theta_p | \mathbf{y})$, known up to proportionality. Assume that the

user supplies "legal" starting values

$$\left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}\right),$$

in the sense that $p\left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}|\mathbf{y}\right) > 0$. The implementation of the Gibbs sampler consists of iterating through the loop:

$$\begin{aligned}
&\text{draw } \theta_1^{(1)} \text{ from } p\left(\theta_1|\theta_2^{(0)}, \ldots, \theta_p^{(0)}, \mathbf{y}\right),\\
&\text{draw } \theta_2^{(1)} \text{ from } p\left(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}, \mathbf{y}\right),\\
&\text{draw } \theta_3^{(1)} \text{ from } p\left(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_p^{(0)}, \mathbf{y}\right),\\
&\vdots\\
&\text{draw } \theta_p^{(1)} \text{ from } p\left(\theta_p|\theta_1^{(1)}, \ldots, \theta_{p-1}^{(1)}, \mathbf{y}\right),\\
&\text{draw } \theta_1^{(2)} \text{ from } p\left(\theta_1|\theta_2^{(1)}, \ldots, \theta_p^{(1)}, \mathbf{y}\right),\\
&\vdots\\
&\text{and so on.}
\end{aligned}$$

After an initial period during which samples are dependent on the starting value (burn in period), the draws $\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_p^{(i)}$, for sufficiently large i, are regarded as samples from the normalized posterior distribution with density

$$p\left(\theta_1, \theta_2, \ldots, \theta_p|\mathbf{y}\right) / \int p\left(\theta_1, \theta_2, \ldots, \theta_p|\mathbf{y}\right) d\theta_1 \ldots d\theta_p.$$

The coordinate $\theta_j^{(i)}$ is regarded as a draw from its marginal posterior distribution with density

$$p\left(\theta_j|\mathbf{y}\right) / \int p\left(\theta_j|\mathbf{y}\right) d\theta_j.$$

The Fully Conditional Distributions Determine the Joint Distribution

The Gibbs sampler produces draws from a joint distribution by sampling successively from all fully conditional posterior distributions. This implies that the form of the fully conditional distributions determines uniquely the form of the joint distribution. The main idea is sketched below for the two-dimensional case. The general case is known as the Hammersley–Clifford in the spatial statistics literature (Besag, 1974).

Consider the identity

$$p\left(x, y\right) = p\left(y|x\right) p\left(x\right) = p\left(x|y\right) p\left(y\right). \tag{11.39}$$

From (11.39), it follows that

$$p(y) = \frac{p(y|x)}{p(x|y)} p(x) \propto \frac{p(y|x)}{p(x|y)}. \tag{11.40}$$

The normalized marginal density is

$$p(y) = \frac{p(y|x)/p(x|y)}{\int p(y|x)/p(x|y)\, dy}.$$

Substituting in (11.39) yields

$$p(x,y) = \frac{p(y|x)}{\int p(y|x)/p(x|y)\, dy}.$$

This shows that $p(x,y)$ can be expressed in terms of the conditional distributions (even though it may not be possible to write the joint distribution explicitly). This result is based on the implicit assumption that the joint distribution $[X,Y]$ exists.

Example 11.1 *A single observation from a bivariate normal distribution with known covariance matrix*

As a trivial example, consider data that consist of a single observation $\mathbf{y} = (y_1, y_2)$ from a bivariate normal distribution with unknown mean

$$\boldsymbol{\theta} = (\theta_1, \theta_2)$$

and known covariance matrix

$$\mathbf{V} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Thus,

$$\mathbf{y}|\boldsymbol{\theta}, \mathbf{V} \sim N(\boldsymbol{\theta}, \mathbf{V}).$$

One wishes to obtain draws from $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{V})$. Let the prior distribution for $\boldsymbol{\theta}$ be proportional to a constant, independent of $\boldsymbol{\theta}$. Then the posterior density of $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{V}) &\propto p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{V}) \\ &\propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{V}), \end{aligned} \tag{11.41}$$

a bivariate normal distribution. The stochastic element in this distribution is the vector $\boldsymbol{\theta}$, for fixed data $\mathbf{y}$ and covariance matrix $\mathbf{V}$. When regarded as a function of $\boldsymbol{\theta}$, (11.41) is the density of a normal distribution with mean $\mathbf{y}$ and variance $\mathbf{V}$. That is

$$\boldsymbol{\theta}|\mathbf{y}, \mathbf{V} \sim N_2(\mathbf{y}, \mathbf{V}). \tag{11.42}$$

Implementation of the Gibbs sampler in a single-site updating scheme requires drawing successively from

$$\theta_1|\theta_2, \mathbf{y}, \mathbf{V} \sim N\left(y_1 + \rho\left(\theta_2 - y_2\right), 1 - \rho^2\right) \tag{11.43}$$

and from

$$\theta_2|\theta_1, \mathbf{y}, \mathbf{V} \sim N\left(y_2 + \rho\left(\theta_1 - y_1\right), 1 - \rho^2\right). \tag{11.44}$$

After a number of rounds of iteration (the burn-in period) the system converges to the stationary distribution (11.42). At this point, the samples obtained from (11.43) are Monte Carlo draws from $p\left(\theta_1|\mathbf{y}, \mathbf{V}\right)$ and those from (11.44) are Monte Carlo draws from $p\left(\theta_2|\mathbf{y}, \mathbf{V}\right)$, the densities of the respective marginal posterior distributions. The samples (θ_1, θ_2) are correlated draws from $p\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{V}\right)$. This correlation among the samples slows down convergence and increases the Monte Carlo sampling error of estimates of features of the posterior distribution. On the other hand, joint updating involves sampling from (11.42); in this example, the system converges in one round. In the joint updating implementation, the pairs (θ_1, θ_2) are independent draws from $p\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{V}\right)$. This is equivalent to direct sampling from the target distribution of interest. Obviously, one would not employ MCMC in such a situation. ■

Example 11.2 *A hierarchical Bayesian model*
Consider the two-parameter Bayesian model

$$y_i|\mu, \sigma^2 \sim N\left(\mu, \sigma^2\right), \quad i = 1, \ldots, n, \ -\infty < \mu < \infty, \ \sigma^2 > 0,$$

where μ and σ^2 are the unknown mean and variance, respectively. These are assumed to be a priori independent, with prior distributions equal to

$$\mu \sim N\left(0, 1\right)$$

and

$$\sigma^2|S, v \sim vS\chi_v^{-2}.$$

That is, the mean is assigned a normal $(0, 1)$ prior, and the variance a scaled inverted chi-square distribution with (assumed known) parameters S and v. Under conditional independence, the likelihood is

$$\begin{aligned} p\left(\mathbf{y}|\mu, \sigma^2\right) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right] \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum\limits_{i=1}^{n}\left(y_i - \mu\right)^2}{2\sigma^2}\right], \end{aligned}$$

and the joint posterior density is given by

$$\begin{aligned} p\left(\mu,\sigma^2|\mathbf{y}\right) &\propto p\left(\mu\right)p\left(\sigma^2\right)p\left(\mathbf{y}|\mu,\sigma^2\right) \\ &\propto \exp\left[-\frac{\mu^2}{2}\right]\left(\sigma^2\right)^{-\left(\frac{v}{2}+1\right)}\exp\left[-\frac{vS}{2\sigma^2}\right]\left(\sigma^2\right)^{-\frac{n}{2}} \\ &\quad\times\exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}\right]. \end{aligned} \tag{11.45}$$

Implementing the Gibbs sampler requires knowledge of the fully conditional posterior distributions with densities $p\left(\mu|\sigma^2,\mathbf{y}\right)$ and $p\left(\sigma^2|\mu,\mathbf{y}\right)$ – omitting the conditioning on hyperparameters S and v.
The derivation of the fully conditional posterior distribution of μ requires extracting terms which are function of μ from the joint posterior density (11.45). This leads to

$$\begin{aligned} p\left(\mu|\sigma^2,\mathbf{y}\right) &\propto \exp\left[-\frac{\mu^2}{2}\right]\exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{\mu^2}{2}\right]\exp\left[-\frac{\sum_{i=1}^{n}\left[\left(y_i-\widehat{\mu}\right)+\left(\widehat{\mu}-\mu\right)\right]^2}{2\sigma^2}\right], \end{aligned} \tag{11.46}$$

where $\widehat{\mu}=\sum_{i=1}^{n}y_i/n$. Expanding the square in the second term and noting that

$$\sum_i\left(y_i-\widehat{\mu}\right)\left(\widehat{\mu}-\mu\right)=\left(\widehat{\mu}-\mu\right)\sum_i\left(y_i-\widehat{\mu}\right)=0,$$

expression (11.46) reduces to

$$p\left(\mu|\sigma^2,\mathbf{y}\right)\propto\exp\left[-\frac{\sigma^2\mu^2+n\left(\mu-\widehat{\mu}\right)^2}{2\sigma^2}\right]. \tag{11.47}$$

Using the identity (i.e., Box and Tiao, 1973, page 74):

$$A\left(z-a\right)^2+B\left(z-b\right)^2=\left(A+B\right)\left(z-c\right)^2+\frac{AB}{A+B}\left(a-b\right)^2$$

with $c=\left(Aa+Bb\right)/\left(A+B\right)$, and associating σ^2 with A, n with B, μ with z, 0 with a and $\widehat{\mu}$ with b, then (11.47) can be rewritten as

$$p\left(\mu|\sigma^2,\mathbf{y}\right)\propto\exp\left[-\frac{\left(\sigma^2+n\right)\left(\mu-m\right)^2}{2\sigma^2}\right], \tag{11.48}$$

where $m = n\widehat{\mu}/\left(\sigma^2 + n\right)$. By inspection, (11.48) is recognized as the kernel of the density of a normal distribution with mean m and variance $\sigma^2/\left(\sigma^2 + n\right)$. Therefore,

$$\mu|\sigma^2, \mathbf{y} \sim N\left(\frac{n\widehat{\mu}}{(\sigma^2+n)}, \frac{\sigma^2}{(\sigma^2+n)}\right). \tag{11.49}$$

In order to derive $p\left(\sigma^2|\mu, \mathbf{y}\right)$, terms including σ^2 are extracted from the joint posterior density (11.45); this leads to

$$\begin{aligned} p\left(\sigma^2|\mu, \mathbf{y}\right) &\propto \left(\sigma^2\right)^{-\left(\frac{v}{2}+1\right)} \exp\left[-\frac{vS}{2\sigma^2}\right] \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2}\right] \\ &= \left(\sigma^2\right)^{-\left(\frac{\widetilde{v}}{2}+1\right)} \exp\left[-\frac{\widetilde{v}\widetilde{S}}{2\sigma^2}\right], \end{aligned}$$

where $\widetilde{S} = \left(vS + \sum_{i=1}^{n}\left(y_i-\mu\right)^2\right)/\widetilde{v}$, and $\widetilde{v} = v + n$. This is the kernel of the density of a scaled inverted chi-square distribution with parameters $\widetilde{v}$ and $\widetilde{S}$. In brief,

$$\sigma^2|\mu, \mathbf{y} \sim \widetilde{v}\widetilde{S}\chi_{\widetilde{v}}^{-2}. \tag{11.50}$$

Generation of a sample from (11.50) requires drawing a chi-square deviate with $\widetilde{v}$ degrees of freedom, inverting this value, and multiplying it by $\left(vS + \sum_{i=1}^{n}\left(y_i-\mu\right)^2\right)$. One cycle of the Gibbs sampling algorithm consists of drawing from (11.49), computing the quantities $\left(vS + \sum_{i=1}^{n}\left(y_i-\mu\right)^2\right)$ using the realized value in place of μ, and finally drawing from (11.50). At convergence, $\mu^{(i)}$ and $\sigma^{2(i)}$ are elements of the ith sample from the marginal distributions $[\mu|\mathbf{y}]$ and $\left[\sigma^2|\mathbf{y}\right]$, respectively. ■

Example 11.3 *A bivariate normal model with unknown covariance matrix*

Let $\mathbf{y}_i = \left(y_{i1}, y_{i2}\right)'$, $(i = 1, \ldots, n)$, represent n independent samples from the bivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and variance defined by the 2×2 matrix $\mathbf{V}$. That is,

$$\begin{aligned} p\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}\right) &= \left|2\pi\mathbf{V}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)'\mathbf{V}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\right] \\ &= (2\pi)^{-n}\left|\mathbf{V}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{S}\right)\right], \end{aligned} \tag{11.51}$$

where

$$\mathbf{S} = \begin{bmatrix} \sum_i \left(y_{i1}-\mu_1\right)^2 & \sum_i \left(y_{i1}-\mu_1\right)\left(y_{i2}-\mu_2\right) \\ \sum_i \left(y_{i1}-\mu_1\right)\left(y_{i2}-\mu_2\right) & \sum_i \left(y_{i2}-\mu_2\right)^2 \end{bmatrix}.$$

Take a two-dimensional uniform distribution as prior for $\boldsymbol{\mu}$, and let the prior for $\mathbf{V}$ be the scaled inverted Wishart distribution with density

$$p\left(\mathbf{V}|\mathbf{V}_0, v\right) \propto |\mathbf{V}|^{-\frac{1}{2}(v+3)} \exp\left[-\frac{1}{2} tr\left(\mathbf{V}^{-1}\mathbf{V}_0^{-1}\right)\right],$$

where $\mathbf{V}_0$ and v are hyperparameters assumed known. The joint posterior density of the parameters of this model (suppressing the dependence on $\mathbf{V}_0$ and v in the notation) is

$$\begin{aligned} p\left(\boldsymbol{\mu}, \mathbf{V}|\mathbf{y}\right) &\propto p\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}\right) p\left(\mathbf{V}\right) \\ &= (2\pi)^{-n} |\mathbf{V}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{V}^{-1}\mathbf{S}\right)\right] |\mathbf{V}|^{-\frac{1}{2}(v+3)} \\ &\quad \times \exp\left[-\frac{1}{2} tr\left(\mathbf{V}^{-1}\mathbf{V}_0^{-1}\right)\right]. \end{aligned} \tag{11.52}$$

Extracting the terms in $\boldsymbol{\mu}$ from (11.52) yields as conditional posterior density

$$\begin{aligned} p\left(\boldsymbol{\mu}|\mathbf{V}, \mathbf{y}\right) &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \boldsymbol{\mu}\right)' \mathbf{V}^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}\right)\right] \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left[\left(\mathbf{y}_i - \overline{\mathbf{y}}\right) + \left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)\right]' \mathbf{V}^{-1}\left[\left(\mathbf{y}_i - \overline{\mathbf{y}}\right) + \left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)\right]\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)' \mathbf{V}^{-1}\left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)\right] \\ &= \exp\left[-\frac{1}{2} n\left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)' \mathbf{V}^{-1}\left(\overline{\mathbf{y}} - \boldsymbol{\mu}\right)\right], \end{aligned}$$

where $\overline{\mathbf{y}}' = \left(\overline{y}_1, \overline{y}_2\right) = \left(n^{-1}\sum_{i=1}^{n} y_{i1}, n^{-1}\sum_{i=1}^{n} y_{i2}\right)$. Therefore, from inspection, the fully conditional posterior distribution of the vector $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu}|\mathbf{V}, \mathbf{y} \sim N\left(\overline{\mathbf{y}}, n^{-1}\mathbf{V}\right). \tag{11.53}$$

The fully conditional distribution of the covariance matrix is derived from (11.52) as well. The terms in $\mathbf{V}$ are

$$\begin{aligned} p\left(\mathbf{V}|\boldsymbol{\mu}, \mathbf{y}\right) &\propto |\mathbf{V}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{V}^{-1}\mathbf{S}\right)\right] |\mathbf{V}|^{-\frac{1}{2}(v+3)} \\ &\quad \times \exp\left[-\frac{1}{2} tr\left(\mathbf{V}^{-1}\mathbf{V}_0^{-1}\right)\right] \\ &= |\mathbf{V}|^{-\frac{1}{2}(v+3+n)} \exp\left[-\frac{1}{2} tr\left[\mathbf{V}^{-1}\left(\mathbf{S} + \mathbf{V}_0^{-1}\right)\right]\right], \end{aligned}$$

which is the kernel of a scaled inverted Wishart distribution with parameters $v+n$ and $\left(\mathbf{S}+\mathbf{V}_0^{-1}\right)^{-1}$. Thus,

$$\mathbf{V}|\boldsymbol{\mu},\mathbf{y} \sim IW\left(\left(\mathbf{S}+\mathbf{V}_0^{-1}\right)^{-1}, v+n\right). \tag{11.54}$$

■

11.6 Langevin–Hastings Algorithm

A general algorithm that, in principle, allows joint updates for the complete parameter vector of a model ($\boldsymbol{\varphi}$, say, of length r), is based on the Langevin–Hastings algorithm (Besag, 1994). The idea is to generate a proposal vector $\boldsymbol{\varphi}^{(t+1)}$ at cycle $t+1$, from a candidate generating density defined by the normal process

$$N\left(\boldsymbol{\varphi}^{(t)}+\frac{\gamma}{2}\frac{\partial}{\partial\boldsymbol{\varphi}^{(t)}}\ln p\left(\boldsymbol{\varphi}^{(t)}|\mathbf{y}\right), \mathbf{I}\gamma\right), \tag{11.55}$$

where γ is a scalar tuning parameter, $\mathbf{I}$ is the $r\times r$ identity matrix and $\ln p(\boldsymbol{\varphi}|\mathbf{y})$ is the log-posterior density, usually known up to proportionality. Note that the vector $\boldsymbol{\varphi}$ includes all the parameters of the model. For example in a Gaussian process, this includes location and dispersion parameters. The proposal is then accepted using the Metropolis–Hastings acceptance ratio (11.27). The use of the gradient of the log-target density in the proposal distribution can lead to much better convergence properties than, for example, the simple random walk Metropolis–Hastings proposal kernel $N\left(\boldsymbol{\varphi}^{(t)}, \mathbf{I}\gamma\right)$. The gradient in (11.55) is supposed to improve moves toward the mode of $p(\boldsymbol{\varphi}|\mathbf{y})$. Further developments and improvements of this approach were presented by Stramer and Tweedie (1998).

11.7 Reversible Jump MCMC

One motivation for reversible jump MCMC is to provide a more general recipe than Metropolis–Hastings to simulate from posterior distributions on spaces of varying dimension; these arise naturally when the number of parameters of the object of inference is not fixed. An example is inferences concerning the number of quantitative trait loci (QTL) affecting a trait, using genetic markers. This number can be treated as a random variable. The Markov chain is allowed to jump across states of different dimension, and each state is characterized by a particular number of QTL. The distribution of the proportion of times that the chain has spent across the various states is a Monte Carlo estimate of the marginal posterior distribution of

the number of QTL affecting the trait. This is illustrated in Chapter 16. Another classical example is the number of densities appearing in a mixture distribution, a problem studied by Richardson and Green (1997).

Reversible jump though, is not limited to simulation of invariant distributions that have densities on spaces of different dimensions. The algorithm can also be used when the models posed have the same number of parameters. An example of this situation is given at the end of this chapter involving a gamma and a lognormal model. Of course, if the number of competing models is small, other approaches may be computationally more efficient than reversible jump.

The reversible jump algorithm was introduced by Green (1995), who illustrates possible applications of reversible jump with several examples. The paper is rather technical and a reader unfamiliar with measure theory will find it difficult to fully grasp the details. Waagepetersen and Sorensen (2001) present a simple, self-contained derivation of reversible jump in a tutorial style and avoiding measure theoretical details. This section is based on the latter paper. The derivation of the acceptance probability to moves between spaces of possibly different dimension presented here, follows essentially the same steps as for the Metropolis–Hastings acceptance probability: using the reversibility condition as the point of departure, first, the transition kernel is expressed in terms of a proposal distribution and an acceptance probability. Second, a change of variable is performed that allows both sides of the detailed balance equation to be expressed in terms of the same parameters. The change of variable is made possible by a monotone transformation; in the case of the reversible jump, this requires that the dimension matching condition is fulfilled, as explained below. Identifying the conditions for equality between the probability of opposite moves and, thus, for detailed balance to hold, leads to the final step in the derivation of the acceptance probability.

Below, a step-by-step derivation of the reversible jump acceptance probability is presented. The notation is a little involved since there is a need to account for the fact that parameters change as the Markov chain jumps from one model to the next. Two examples are presented at the end of this chapter and a further application of reversible jump can be found in Chapter 16 on QTL analysis. In one of the examples, two models with a different number of parameters are considered. In the other, the number of parameters is the same in both models, illustrating the generality of the algorithm for model choice.

11.7.1 The Invariant Distribution

Here π denotes the joint probability distribution of (M, Z), where $M \in \{1, 2, \ldots, I\}$ is a "model indicator" and Z is a real stochastic vector, possibly of varying dimension (I represents either a finite integer or ∞). The models can be different because of their parametric form but the number

of parameters do not need to differ. The vector Z takes values in the set C defined as the union $C = \cup_{m=1}^{I} C_m$ of spaces $C_m = \mathbb{R}^{n_m}$, $(n_m > 1)$. Given $M = m$, Z can only take values in C_m, so that π is specified by $p_m = P(M = m)$ and densities $f(\cdot|M = m)$ on C_m, $(m = 1, 2, \ldots)$. Thus, for $A_m \subseteq C_m$, the joint probability distribution of (M, Z) is

$$\begin{aligned} P(M = m, Z \in A_m) &= P(M = m)\, P(Z \in A_m | M = m) \\ &= p_m \int_{A_m} f(z|M = m)\, dz. \end{aligned} \tag{11.56}$$

The density $f(\cdot|M = m)$ is denoted f_m hereinafter.

If a number of competing models are posed, then p_m may represent the posterior probability of model m, and given $M = m$, f_m is the posterior density of the n_m-dimensional vector Z of parameters associated with model m. In this case

$$p_m f_m = c^{-1} \widetilde{p}_m h(z|m)\, l(y|m, z), \tag{11.57}$$

where $\widetilde{p}_m$ is the prior probability of model m, $h(z|m)$ is the prior density of Z given $M = m$, $l(y|m, z)$ is the likelihood of the data y given $(M, Z) = (m, z)$, and c is the overall (typically unknown) normalizing constant

$$c = \sum_{m=1}^{I} \widetilde{p}_m \int_{C_m} h(z|m)\, l(y|m, z)\, dz. \tag{11.58}$$

11.7.2 Generating the Proposal

In the joint updating Metropolis–Hastings algorithm, the candidate point $Y_{n+1} \in \mathbb{R}^d$ for the new state $X_{n+1} \in \mathbb{R}^d$, is generated from the d-dimensional proposal density $q(x, \cdot)$. In the single-site updating, the d-dimensional candidate point

$$Y_{n+1} = (x_1, x_2, \ldots, x_{i-1}, Z_i, x_{i+1}, \ldots, x_d)$$

is generated by drawing the random variable Z_i from the one-dimensional proposal density $q_i(x, \cdot)$. With some abuse of notation, Y_{n+1} above can be written

$$Y_{n+1} = g(x_1, x_2, \ldots, x_{i-1}, Z_i, x_{i+1}, \ldots, x_d),$$

where the function g is the identity mapping.

In the context of reversible jump, each state X_i of the chain contains two components, i.e., $X_i = (M_i, Z_i)$, where M_i is the model indicator and where Z_i is a stochastic vector in C_{M_i}. Suppose that (m, z) is the value of the current state X_n of the Markov chain and a move to the value (m', z') is considered for the next state X_{n+1}. A proposal $Y_{n+1} = \left(Y_{n+1}^{\text{ind}}, Y_{n+1}^{\text{par}}\right)$ for X_{n+1} is generated as described below, where superscripts ind and par are

labels for the proposal of the model indicator M_{n+1} and for the vector Z_{n+1}, respectively. With user-defined probability $p_{mm'}$, $\left(\sum_{m'=1}^{I} p_{mm'} = 1\right)$, the proposal Y_{n+1}^{ind} is set equal to m', and given $Y_{n+1}^{\text{ind}} = m'$, the proposal Y_{n+1}^{par} is generated in $C_{m'}$. A very general mechanism is to construct the proposal Y_{n+1}^{par} by applying a deterministic mapping $g_{1mm'}$ to the previous value z and to a random component U. This mechanism can be formulated as

$$Y_{n+1}^{\text{par}} = g_{1mm'}(z, U), \tag{11.59}$$

where U is a random vector which has density $q_{mm'}(z, \cdot)$ on $\mathbb{R}^{n_{mm'}}$. The proposal Y_{n+1} is finally accepted with an acceptance probability

$$a_{mm'}\left(z, Y_{n+1}^{\text{par}}\right),$$

which is derived below.

When considering a move from a state (m, z) to

$$(m', z') = (m', g_{1mm'}(z, u)),$$

and a move in the opposite direction from (m', z') to

$$(m, z) = (m, g_{1m'm}(z', u')),$$

the vectors (z, u) and (z', u') must be of equal dimension. That is, the dimension matching condition

$$n_m + n_{mm'} = n_{m'} + n_{m'm} \tag{11.60}$$

needs to be fulfilled. Further, it will be assumed that there exist functions $g_{2mm'}$ and $g_{2m'm}$ such that the mapping $g_{mm'}$, given by

$$(z', u') = g_{mm'}(z, u) = (g_{1mm'}(z, u), g_{2mm'}(z, u)), \tag{11.61}$$

is one-to-one with

$$\begin{aligned}(z, u) &= g_{mm'}^{-1}(z', u') \\ &= g_{m'm}(z', u') = (g_{1m'm}(z', u'), g_{2m'm}(z', u'))\end{aligned} \tag{11.62}$$

and that $g_{mm'}$ is differentiable. The transformations (11.61) and (11.62) are possible because the mapping $g_{mm'}$ is one-to-one with $g_{m'm}$; a necessary condition for the existence of the one-to-one mapping is that the dimension matching (11.60) holds.

11.7.3 Specifying the Reversibility Condition

Assuming $X_n = (M_n, Z_n) \sim \pi$, the condition of reversibility is

$$\begin{aligned}&P(M_n = m, Z_n \in A_m, M_{n+1} = m', Z_{n+1} \in B_{m'}) \\ &= P(M_n = m', Z_n \in B_{m'}, M_{n+1} = m, Z_{n+1} \in A_m)\end{aligned} \tag{11.63}$$

for all $m, m' \in \{1, 2, \ldots, I\}$, and all subsets A_m and $B_{m'}$ of C_m and $C_{m'}$, respectively. In analogy with (11.15), the left-hand side of (11.63) is

$$\begin{aligned} &P\left(M_n = m, Z_n \in A_m, M_{n+1} = m', Z_{n+1} \in B_{m'}\right) \\ &= \int_{A_m} P\left(M_{n+1} = m', Z_{n+1} \in B_{m'} | X_n = (m, z)\right) p_m f_m(z)\, dz, \end{aligned} \tag{11.64}$$

where $P\left(M_{n+1} = m', Z_{n+1} \in B_{m'} | X_n = (m, z)\right)$ is the transition kernel. As in (11.17), let

$$\begin{aligned} &\mathsf{Q}^a_{mm'}(z, B_{m'}) \\ &= P\left(Y^{\text{ind}}_{n+1} = m', Y^{\text{par}}_{n+1} \in B_{m'} \text{ and } Y_{n+1} \text{ is accepted} | X_n = (m, z)\right) \end{aligned}$$

be the joint probability of generating the proposal Y_{n+1} with $Y^{\text{ind}}_{n+1} = m'$ and Y^{par}_{n+1} in $B_{m'}$ and accepting the proposal, given that the current state of the Markov chain is $X_n = (m, z)$. Also, as in (11.18), let

$$s_m(z) = P\left(Y_{n+1} \text{ is rejected} | X_n = (m, z)\right)$$

be the probability of rejecting the proposal. Then the transition kernel can be written as

$$\begin{aligned} &P\left(M_{n+1} = m', Z_{n+1} \in B_{m'} | X_n = (m, z)\right) \\ = \;& \mathsf{Q}^a_{mm'}(z, B_{m'}) + s_m(z)\, I\left(m = m', z \in B_{m'}\right). \end{aligned}$$

Substituting in (11.64), the left-hand side of (11.63) equals

$$\begin{aligned} &p_m \int_{A_m} \mathsf{Q}^a_{mm'}(z, B_{m'})\, f_m(z)\, dz \\ &+ p_m \int_{A_m} s_m(z)\, I\left(m = m', z \in B_{m'}\right) f_m(z)\, dz \\ &= p_m \int_{A_m} \mathsf{Q}^a_{mm'}(z, B_{m'})\, f_m(z)\, dz \\ &+ p_m \int s_m(z)\, I\left(m = m', z \in A_m \cap B_{m'}\right) f_m(z)\, dz, \end{aligned} \tag{11.65}$$

where

$$\begin{aligned} &p_m \int_{A_m} \mathsf{Q}^a_{mm'}(z, B_{m'})\, f_m(z)\, dz \\ &= P\left(M_n = m, Z_n \in A_m, Y^{ind}_{n+1} = m', Y^{par}_{n+1} \in B_{m'}, Y_{n+1} \text{ is accepted}\right). \end{aligned}$$

By symmetry, the right-hand side of (11.63) equals

$$\begin{aligned} &p_{m'} \int_{B_{m'}} \mathsf{Q}^a_{m'm}(z', A_m)\, f_{m'}(z')\, dz' \\ &+ p_{m'} \int s_{m'}(z')\, I\left(m = m', z' \in B_{m'} \cap A_m\right) f_{m'}(z')\, dz'. \end{aligned} \tag{11.66}$$

The second terms in (11.65) and (11.66) are equal both in the case when $m \neq m'$ (in which case they are zero, because the indicator function takes the value zero), and when $m = m'$ (in which case the move is within the same model, and both expressions are identical). Therefore a sufficient condition for reversibility to hold is, for all m and m',

$$\begin{aligned} & p_m \int_{A_m} \mathsf{Q}^a_{mm'}(z, B_{m'}) f_m(z)\, dz \\ &= p_{m'} \int_{B_{m'}} \mathsf{Q}^a_{m'm}(z', A_m) f_{m'}(z')\, dz'. \end{aligned} \tag{11.67}$$

11.7.4 Derivation of the Acceptance Probability

Equation (11.67) is now written more explicitly. Since,
(a) Y^{ind}_{n+1} is set equal to m' with probability $p_{mm'}$;
(b) Y^{par}_{n+1} is generated in $C_{m'}$ and belongs in $B_{m'}$ implies $Y^{\text{par}}_{n+1} \in B_{m'} \Leftrightarrow g_{1mm'}(z, U) = z' \in B_{m'}$;
(c) Y_{n+1} is accepted with probability $a_{mm'}(z, g_{1mm'}(z, U)) = a_{mm'}(z, z')$, and $U \sim q_{mm'}(z, \cdot)$.

It follows that

$$\mathsf{Q}^a_{mm'}(z, B_{m'}) = p_{mm'} \int I(z' \in B_{m'})\, a_{mm'}(z, z')\, q_{mm'}(z, u)\, du. \tag{11.68}$$

The left-hand side of (11.67) is therefore

$$\begin{aligned} & p_m \int_{A_m} \mathsf{Q}^a_{mm'}(z, B_{m'}) f_m(z)\, dz \\ &= p_m \int_{A_m} \int I(z' \in B_{m'})\, p_{mm'} a_{mm'}(z, z')\, q_{mm'}(z, u)\, f_m(z)\, dz\, du \\ &= p_m \int \int I(z \in A_m, z' \in B_{m'})\, p_{mm'} \\ &\quad \times a_{mm'}(z, z')\, q_{mm'}(z, u)\, f_m(z)\, dz\, du, \end{aligned} \tag{11.69}$$

and, by symmetry, the right-hand side is

$$\begin{aligned} & p_{m'} \int_{B_{m'}} \mathsf{Q}^a_{m'm}(z', A_m) f_{m'}(z')\, dz' \\ &= p_{m'} \int \int I(z' \in B_{m'}, z \in A_m)\, p_{m'm} \\ &\quad \times a_{m'm}(z', z)\, q_{m'm}(z', u')\, f_{m'}(z')\, dz'\, du'. \end{aligned} \tag{11.70}$$

To study the conditions that satisfy reversibility and therefore equality of (11.69) and (11.70), both equations will now be expressed as functions of the same variables. This is possible due to the dimension matching assumption

and relationships (11.61) and (11.62). Using the fact that (see equations (2.36) and (2.38) of Chapter 2)

$$dz'du' = |\det\left(g'_{mm'}\left(z,u\right)\right)|\, dz\, du, \tag{11.71}$$

where

$$g'_{mm'}\left(z,u\right) = \frac{\partial g_{mm'}\left(z,u\right)}{\partial\left(z,u\right)} = \begin{bmatrix} \frac{\partial g_{1mm'}(z,u)}{\partial dz} & \frac{\partial g_{2mm'}(z,u)}{\partial z} \\ \frac{\partial g_{1mm'}(z,u)}{\partial du} & \frac{\partial g_{2mm'}(z,u)}{\partial u} \end{bmatrix},$$

equation (11.70) can be written as

$$\begin{aligned} &P\left(M_n = m', Z_n \in B_{m'}, M_{n+1} = m, Z_{n+1} \in A_m\right) \\ &= \int\int I\left(z' \in B_{m'}, z \in A_m\right) p_{m'm} a_{m'm}\left(z', z\right) \\ &\times q_{m'm}\left(z', u'\right) p_{m'} f_{m'}\left(z'\right) |\det\left(g'_{mm'}\left(z,u\right)\right)|\, dz\, du. \end{aligned} \tag{11.72}$$

By inspection, it is clear that equality between (11.69) and (11.72) is satisfied if

$$\begin{aligned} &p_{mm'} a_{mm'}\left(z, z'\right) q_{mm'}\left(z,u\right) p_m f_m\left(z\right) \\ &= p_{m'm} a_{m'm}\left(z', z\right) q_{m'm}\left(z', u'\right) p_{m'} f_{m'}\left(z'\right) |\det\left(g'_{mm'}\left(z,u\right)\right)|. \end{aligned}$$

Choosing the acceptance probability as large as possible, subject to the detailed balance condition as suggested by Peskun (1973), yields

$$\begin{aligned} &a_{mm'}\left(z, z'\right) \\ &= \min\left(1, \frac{p_{m'm} q_{m'm}\left(z', u'\right) p_{m'} f_{m'}\left(z'\right)}{p_{mm'} q_{mm'}\left(z,u\right) p_m f_m\left(z\right)} \left|\det\left(\frac{\partial g_{mm'}\left(z,u\right)}{\partial\left(z,u\right)}\right)\right|\right) \end{aligned} \tag{11.73}$$

whenever $p_{mm'} q_{mm'}\left(z,u\right) p_m f_m\left(z\right) > 0$ and where

$$\left(z', u'\right) = g_{mm'}\left(z,u\right).$$

In practice, $p_{mm'} q_{mm'}\left(z,u\right) p_m f_m\left(z\right) = 0$ only happens if the Markov chain is initialized in a state (m,z) for which $p_m f_m\left(z\right) = 0$. The acceptance probability for a move from z' to z is given by the inverse of (11.73).

11.7.5 Deterministic Proposals

Sometimes it may be simpler to apply deterministic proposals for a move from C_m to $C_{m'}$, i.e., to let $Y_{n+1} = g_{1mm'}\left(z\right)$, and still use a stochastic proposal for the move in the opposite direction. In this case, the dimension matching condition equals

$$n_m = n_{m'} + n_{m'm} \tag{11.74}$$

since $n_{mm'} = 0$. Equations (11.61) and (11.62) become

$$(z', u') = g_{mm'}(z) = (g_{1mm'}(z), g_{2mm'}(z)) \tag{11.75}$$

and

$$(z) = g_{mm'}^{-1}(z', u') = g_{m'm}(z', u') = g_{1m'm}(z', u'). \tag{11.76}$$

That is, the change from state z to state z', defined by (11.75), does not involve the generation of a stochastic variable U; the move is deterministic. The move in the opposite direction, defined by (11.76), requires U'; this move is stochastic. The reversibility condition has the same form as in (11.67), but (11.68) is now given by

$$\begin{aligned} \mathsf{Q}_{mm'}^{a}(z, B_{m'}) &= p_{mm'} I(g_{1mm'}(z) \in B_{m'}) a_{mm'}(z, g_{1mm'}(z)) \\ &= p_{mm'} I(z' \in B_{m'}) a_{mm'}(z, z'). \end{aligned} \tag{11.77}$$

Substituting (11.77) in the left-hand side of (11.67):

$$\begin{aligned} & p_m \int_{A_m} \mathsf{Q}_{mm'}^{a}(z, B_{m'}) f_m(z) dz \\ &= \int I(z \in A_m, z' \in B_{m'}) p_{mm'} \\ &\quad \times a_{mm'}(z, z') p_m f_m(z) dz. \end{aligned} \tag{11.78}$$

With a stochastic proposal for the opposite move, the right-hand side of (11.67) is unchanged and is given by (11.70). The equivalent to (11.71) is now

$$dz' du' = |\det(g'_{mm'}(z))| dz, \tag{11.79}$$

where

$$g'_{mm'}(z) = \frac{\partial g_{mm'}(z)}{\partial z}.$$

Substituting (11.79) in (11.70), and using (11.78), yields the following expression for the detailed balance equation

$$\begin{aligned} & \int I(z \in A_m, z' \in B_{m'}) p_{mm'} a_{mm'}(z, z') p_m f_m(z) dz \\ &= \int I(z' \in B_{m'}, z \in A_m) p_{m'm} a_{m'm}(z', z) \\ &\quad \times q_{m'm}(z', u') p_{m'} f_{m'}(z') |\det(g'_{mm'}(z))| dz, \end{aligned} \tag{11.80}$$

where $u' = g_{2mm'}(z)$, a function of z. Using the same approach as before, the acceptance probability is given now by

$$a(z, z') = \min\left(1, \frac{p_{m'm} q_{m'm}(z', u') p_{m'} f_{m'}(z')}{p_{mm'} p_m f_m(z)} |\det(g'_{mm'}(z))|\right). \tag{11.81}$$

11.7.6 Generating Proposals via the Identity Mapping

In the development presented so far, the proposal in the move from (m, z) to (m', z') is generated via the deterministic mapping (11.16). An alternative to this approach is to let $g_{1mm'}(z, U)$ be the identity mapping and to set $U = Z'$, which results in

$$Y_{n+1}^{par} = Z'.$$

The random variable Z' is generated from the density $q_{mm'}(z, \cdot)$ on $\mathbb{R}^{n_{m'}}$, which may depend on the current value z. The expression equivalent to (11.68) is now

$$\mathsf{Q}_{mm'}^{a}(z, B_{m'}) = p_{mm'} \int I(z' \in B_{m'})\, a_{mm'}(z, z')\, q_{mm'}(z, z')\, dz'. \quad (11.82)$$

Then it is easy to show that under this strategy, the acceptance probability is given by

$$a_{mm'}(z, z') = \min\left(1, \frac{p_{m'm} q_{m'm}(z', z)\, p_{m'} f_{m'}(z')}{p_{mm'} q_{mm'}(z, z')\, p_m f_m(z)}\right), \quad (11.83)$$

which does not include a Jacobian term because of the use of the identity mapping.

This strategy (which we label the FF strategy) was suggested by S. Fernández and R. Fernando (Rohan Fernando (2001), personal communication). The form of (11.83) is similar to (11.13); however, in (11.83), there is an extra term $p_{m'm}/p_{mm'}$ and, further, $q_{mm'}$ and $q_{m'm}$ are densities on $\mathbb{R}^{n_{m'}}$ and on $\mathbb{R}^{n_m}$, respectively. In (11.13), $q(x, y)$ and $q(y, x)$ are both densities on $\mathbb{R}^d$ where d is the dimension of Y and X.

Example 11.4 *Comparing differences between two treatments*

The reversible jump algorithm is showed in detail with a trivial example. Consider a model $(M = 1)$ where the data are assumed to be an i.i.d. realization from

$$y_i | M = 1, t, \sigma^2 \sim N\left(t, \sigma^2\right), \quad i = 1, \ldots, N, \quad (11.84)$$

or from the alternative model $(M = 2)$

$$y_{ij} | M = 2, t_i, \sigma^2 \sim N\left(t_i, \sigma^2\right), \quad i = 1, 2,\ j = 1, \ldots, n, \quad (11.85)$$

where $N = 2n$. In (11.84), t is an overall mean and σ^2 is the variance of the distribution $\left[y_i | M = 1, t, \sigma^2\right]$. The sampling model (11.85) postulates instead that there are two "treatments" t_1 and t_2, and that observations have variance σ^2. To complete the Bayesian structure, prior distributions

$$\Pr(M = 1)\, p\left(t, \sigma^2 | M = 1\right)$$

and

$$\Pr(M = 2)\, p\left(t_1, t_2, \sigma^2 | M = 2\right)$$

are assigned to the parameters of both sampling models. The problem consists of discriminating between these two models. This is done here using reversible jump MCMC, despite the fact that by an appropriate choice of prior distributions, closed forms for the relevant posterior distributions and for the Bayes factor for these models are available (see e.g., O'Hagan, 1994). It is convenient to introduce the stochastic variable T:

$$T = \begin{cases} t, & M = 1, \\ (t_1, t_2), & M = 2. \end{cases}$$

The posterior distribution, which has the form in (11.56), can then be written

$$p\left(M = i, T, \sigma^2 | \mathbf{y}\right) \propto \Pr\left(M = i\right) p\left(T, \sigma^2 | M = i\right) p\left(\mathbf{y} | M = i, T, \sigma^2\right). \tag{11.86}$$

First, reversible jump is implemented using stochastic proposals in both directions. Assume that the current state of the Markov chain is $(m = 1, z)$ and a move to $(m' = 2, z')$, where $z = \left(t, \sigma^2\right)$ and $z' = \left(t_1, t_2, \sigma^2\right)$, is proposed with probability $p_{mm'}$. This probability is chosen by the user, subject to $p_{11} + p_{12} = 1$. With stochastic proposals in both directions, the dimension-matching condition (11.60) is satisfied if two stochastic variables $(u = v_1, v_2)$ are generated in the move from m to m', and one stochastic variable $(u' = v)$ is generated in the move from m' to m. In this case, $n_m = 2$ (associated with t, σ^2), $n_{mm'} = 2$ (associated with v_1, v_2), $n_{m'} = 3$ (associated with t_1, t_2, σ^2), and $n_{m'm} = 1$ (associated with v). The mapping is

$$\begin{aligned} (z', u') &= \left(t_1, t_2, \sigma^2, v\right) \\ &= g_{mm'}\left(z, u\right) \\ &= \left(g_{1mm'}\left(z, u\right), g_{2mm'}\left(z, u\right)\right). \end{aligned}$$

A reasonable choice could be

$$\begin{bmatrix} t_1 \\ t_2 \\ \sigma^2 \\ v \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} t \\ \sigma^2 \\ v_1 \\ v_2 \end{bmatrix}.$$

That is, $g_{1mm'}\left(z, u\right) = \left(t_1, t_2, \sigma^2\right) = \left(t + v_1, t + v_2, \sigma^2\right)$ and

$$g_{2mm'}\left(z, u\right) = v = \frac{1}{2}\left(v_1 + v_2\right).$$

The absolute value of the Jacobian of the transformation in the move from m to m' is

$$\left|\det\left[\frac{\partial g_{mm'}\left(z, u\right)}{\partial\left(z, u\right)}\right]\right| = \left|\det\left[\frac{\partial\left(t + v_1, t + v_2, \sigma^2, \frac{1}{2}\left(v_1 + v_2\right)\right)}{\partial\left(t, \sigma^2, v_1, v_2\right)}\right]\right| = 1.$$

If v is generated from $q_v(z,\cdot)$ and v_1, v_2 from $q_{v_1v_2}(z',\cdot)$, the expression for the acceptance probability (11.73) is

$$a_{mm'}(z,z') = \min\left(1, \frac{p_{m'm}q_v(z,v)\,p_{m'}f_{m'}}{p_{mm'}q_{v_1v_2}(z',v_1,v_2)\,p_mf_m}\right),$$

where, in terms of (11.73), the posterior distribution of $\left(M,T,\sigma^2\right)$ is

$$p_{m'}f_{m'} = c^{-1}\Pr(M=2)\,p\left(t_1,t_2,\sigma^2|M=2\right)p\left(\mathbf{y}|M=2,t_1,t_2,\sigma^2\right) \tag{11.87}$$

and

$$p_mf_m = c^{-1}\Pr(M=1)\,p\left(t,\sigma^2|M=1\right)p\left(\mathbf{y}|M=1,t,\sigma^2\right). \tag{11.88}$$

Notice that the constant c^{-1} cancels in the acceptance ratio.

Second, reversible jump is implemented using a stochastic proposal in one of the moves and a deterministic proposal in the other. As before, consider the move from $m=1$ to $m'=2$. Now the dimension-matching condition is

$$n_m + n_{mm'} = n_{m'}$$

because $n_{m'm}=0$. For the present example, $n_m=2$ (associated with t,σ^2), $n_{mm'}=1$ (associated with u), and $n_{m'}=3$ (associated with t_1,t_2, σ^2). The move from m to m' is based on a stochastic proposal, since it requires the generation of the random variable U from $q(z,\cdot)$. The mapping is

$$\begin{aligned} z' &= \left(t_1,t_2,\sigma^2\right) \\ &= g_{mm'}(z,u) \\ &= g_{1mm'}(z,u). \end{aligned}$$

A reasonable choice is

$$\begin{bmatrix} t_1 \\ t_2 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} t \\ u \\ \sigma^2 \end{bmatrix}. \tag{11.89}$$

That is, $g_{1mm'}(z,u) = \left(t_1,t_2,\sigma^2\right) = \left(t+u,t-u,\sigma^2\right)$. The absolute value of the Jacobian of the transformation is

$$\left|\det\left(\frac{\partial g_{mm'}(z,u)}{\partial(z,u)}\right)\right| = \left|\det\left(\frac{\partial\left(t+u,t-u,\sigma^2\right)}{\partial\left(t,\sigma^2,u\right)}\right)\right| = 2$$

and the acceptance probability for the jump from m to m' is

$$a_{mm'}(z,z') = \min\left(1, \frac{p_{m'm}p_{m'}f_{m'}}{p_{mm'}q(z,u)\,p_mf_m}2\right),$$

where $p_{m'} f_{m'}$ is equal to (11.87) and $p_m f_m$ is equal to (11.88).
The move from $m' = 3$ to $m = 2$ is deterministic. Inverting (11.89) yields

$$\begin{bmatrix} t \\ u \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \sigma^2 \end{bmatrix}.$$

That is,

$$\begin{aligned} (z, u) &= \left(t, \sigma^2, u\right) \\ &= g_{m'm}(z) \\ &= \left(g_{1m'm}(z), g_{2m'm}(z)\right), \end{aligned}$$

where

$$g_{1m'm}(z) = \left(t, \sigma^2\right) = \left(\frac{1}{2}(t_1 + t_2), \sigma^2\right)$$

and

$$g_{2m'm}(z) = u = \frac{1}{2}(t_1 - t_2).$$

The absolute value of the Jacobian of the transformation is

$$\left|\det\left(\frac{\partial g_{m'm}(z)}{\partial z}\right)\right| = \left|\det\left(\frac{\partial\left(\frac{1}{2}(t_1 + t_2), \sigma^2, \frac{1}{2}(t_1 - t_2)\right)}{\partial(t_1, t_2, \sigma^2)}\right)\right| = \frac{1}{2}.$$

The acceptance probability for the jump from m' to m is given by

$$a_{m'm}(z', z) = \min\left(1, \frac{p_{mm'} q(z, u)\, p_m f_m}{p_{m'm} p_{m'} f_{m'}} \frac{1}{2}\right),$$

which is equal to the inverse of $a_{mm'}(z, z')$, as expected.

Finally, the problem is approached via the FF strategy. Consider the same move from $m = 1$ to $m' = 2$ with now, $z = t$ and the update including $z' = (t_1, t_2)$ only, as σ^2 is common to Models 1 and 2. Then the acceptance probability is simply

$$a_{mm'}(z, z') = \min\left(1, \frac{p_{m'm} q_{m'm}(t)\, p_{m'} f_{m'}}{p_{mm'} q_{mm'}(t_1, t_2)\, p_m f_m}\right),$$

where $q_{m'm}(\cdot)$ is a density on $\mathbb{R}$ (e.g., a normal density with mean and variance $(t_1 + t_2)/2$ and σ^2, respectively) and $q_{mm'}$ is a density on $\mathbb{R}^2$ (e.g., a bivariate normal density with mean (t, t) and well-tuned covariance).
Extension to an unknown number of covariates (treatments) is obvious, and requires incorporating a prior distribution to this number. In this case, reversible jump offers a recipe for computing the posterior probability for the number of covariates in the regression model. That is, the number of covariates is treated as a random variable, to be inferred from the data at hand. ■

Example 11.5 *Choosing between a gamma and a lognormal model*
In this example, reversible jump is applied to obtain a MCMC-based posterior probability of two models with the same number of parameters, the gamma and the lognormal model. Such a comparison cannot be performed via the traditional Neyman–Pearson maximum likelihood ratio test described in Chapter 4. This is so because these models do not generate the required nested structure. In order to perform the test within the frequentist paradigm a modification of the Neyman–Pearson maximum likelihood ratio test is required (Cox, 1961, 1962).
A gamma distributed random variable has p.d.f.

$$g_1(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1}\exp\left[-y/\beta\right], \quad 0 < y < \infty,\ \alpha > 0, \beta > 0.$$

The first two moments of this distribution are

$$\begin{aligned} E(y) &= \alpha\beta, \\ E\left(y^2\right) &= \beta^2\alpha(\alpha+1). \end{aligned}$$

A lognormally distributed random variable has p.d.f.

$$\begin{aligned} g_2\left(y|\mu,\sigma^2\right) &= \frac{1}{y\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(\ln y-\mu)^2\right], \\ 0 &< y < \infty,\ -\infty < \mu < \infty,\ \sigma > 0. \end{aligned}$$

The first two moments of this distribution are

$$\begin{aligned} E(y) &= \exp\left[\mu + \frac{\sigma^2}{2}\right], \\ E\left(y^2\right) &= \exp\left[2\mu + 2\sigma^2\right]. \end{aligned}$$

Consider Model 1 as the gamma model, and Model 2 as the lognormal model. Define the stochastic indicator $M \in \{1,2\}$ for Models 1 and 2 respectively.
Let the posterior probability associated with Model 1 be

$$f_1(\alpha,\beta,M=1|y) \propto g_1(y|\alpha,\beta,M=1)h_1(\alpha,\beta|M=1)\Pr(M=1), \tag{11.90}$$

where g_1 is the gamma density, h_1 is a prior for the gamma density parameters α and β, and $\Pr(M=1)$ is the a priori probability of Model 1. Also, let the posterior probability associated with Model 2 be

$$f_2(\mu,\sigma^2,M=2|y) \propto g_2(y|\mu,\sigma^2,M=2)h_2(\mu,\sigma^2|M=2)\Pr(M=2), \tag{11.91}$$

where g_2 is the lognormal density, h_2 is a prior for the lognormal density parameters μ and σ^2, and $\Pr(M=2)$ is the a priori probability of Model

2. Suppose that the current state of the Markov chain is (m, z), $m = 1$, $z = (\alpha, \beta)$, and that a move is to be made to the lognormal model, i.e., to a state (m', z'), $m' = 2$, $z' = \left(\mu, \sigma^2\right)$. One way to propose values for the parameters μ and σ^2 might be to equate the first- and second-order moments under the current gamma model and the proposed lognormal model and, subsequently, add/multiply some noise, U. More precisely, solve $\exp(\tilde{\mu}+\tilde{\sigma}^2/2) = \alpha\beta$ and $\exp(2\tilde{\mu}+2\tilde{\sigma}^2) = \beta^2\alpha(\alpha+1)$ with respect to $\tilde{\mu}$ and $\tilde{\sigma}^2$, and let the proposals be $\mu = \tilde{\mu}+U_1$ and $\sigma^2 = \tilde{\sigma}^2 U_2$, where $U = (U_1, U_2)$ is generated from $q_{mm'}$. In this case, we have

$$\begin{aligned} \mu &= \ln\left(\frac{\alpha\beta}{\sqrt{1+1/\alpha}}\right) + U_1, \\ \sigma^2 &= \ln(1+1/\alpha)\, U_2, \\ U_1' &= U_1, \\ U_2' &= U_2. \end{aligned} \tag{11.92}$$

That is,

$$(z, u) = (\alpha, \beta, u_1, u_2)$$

and

$$\begin{aligned} (z', U') &= \left(\mu, \sigma^2, U_1', U_2'\right) \\ &= g_{mm'}(\alpha, \beta, U_1, U_2), \\ &= g_{1mm'}\left(\alpha, \beta, U_1, U_2\right),\, g_{2mm'}\left(\alpha, \beta, U_1, U_2\right), \\ &= \left(\log(\alpha\beta/\sqrt{1+1/\alpha}) + U_1,\, \log(1+1/\alpha)U_2\right),\, \left(U_1,\, U_2\right), \end{aligned}$$

where

$$\begin{aligned} g_{1mm'}\left(\alpha, \beta, U_1, U_2\right) &= \left(\mu, \sigma^2\right) \\ &= \left(\log(\alpha\beta/\sqrt{1+1/\alpha}) + U_1, \log(1+1/\alpha)U_2\right), \\ g_{2mm'}\left(\alpha, \beta, U_1, U_2\right) &= \left(U_1, U_2\right). \end{aligned}$$

This move requires generating the stochastic vector $U = (U_1, U_2)$ from $q_{mm'}$.

In the opposite move, from (m', z') to (m, z), solving (11.92) for α, β, U_1, and U_2 yields

$$\begin{aligned} \alpha &= 1/(\exp(\sigma^2/U_2') - 1, \\ \beta &= \exp(\mu - U_1' + \sigma^2/(2U_2'))(\exp(\sigma^2/U_2') - 1), \\ U_1 &= U_1', \\ U_2 &= U_2'. \end{aligned}$$

That is,

$$\begin{aligned}(z,U) &= (\alpha,\beta,U_1,U_2)\\ &= g_{m'm}(\mu,\sigma^2,U_1',U_2')\\ &= g_{1m'm}\left(\mu,\sigma^2,U_1',U_2'\right),\ g_{2m'm}\left(\mu,\sigma^2,U_1',U_2'\right)\\ &= \left(1/(\exp(\sigma^2/U_2')-1),\ \exp(\mu-U_1'+\sigma^2/(2U_2'))(\exp(\sigma^2/U_2')-1)\right),\\ &\quad (U_1',U_2'),\end{aligned}$$

where U_1' and U_2' are generated from $q_{m'm}$ and

$$\begin{aligned}&g_{1m'm}\left(\mu,\sigma^2,U_1',U_2'\right)\\ &= \left(1/(\exp(\sigma^2/U_2')-1),\ \exp(\mu-U_1'+\sigma^2/(2U_2'))(\exp(\sigma^2/U_2')-1)\right),\\ &g_{2m'm}\left(\mu,\sigma^2,U_1',U_2'\right) = (U_1',U_2').\end{aligned}$$

The moves in both directions use stochastic proposals $q_{mm'}$ and $q_{m'm}$. The absolute value of the Jacobian of the transformation in the move from (m,z) to (m',z'), is

$$\begin{aligned}\left|\frac{\partial g_{mm'}(z,u)}{\partial(z,u)}\right| &= \left|\frac{\partial\left(\log(\alpha\beta/\sqrt{1+1/\alpha})+u_1,\log(1+1/\alpha)u_2,u_1,u_2\right)}{\partial(\alpha,\beta,u_1,u_2)}\right|\\ &= u_2\left[\alpha\beta(\alpha+1)\right]^{-1}.\end{aligned}$$

Finally, the acceptance probability for the move from (m,z) to (m',z') is

$$a(z,z') = \min\left(1,\frac{f_2(\mu,\sigma^2,M=2|y)p_{m'm}q_{m'm}}{f_1(\alpha,\beta,M=1|y)p_{mm'}q_{mm'}}u_2\left[\alpha\beta(\alpha+1)\right]^{-1}\right),$$

where the posterior distributions f_1 and f_2 are given by (11.90) and (11.91), respectively.

The move from the gamma to the lognormal model and the reverse move, were chosen here to be stochastic. In this example where the number of parameters is the same in both models, one could have chosen deterministic moves in both directions. In this case, the acceptance probability of moving from z to z' would be given by

$$a(z,z') = \min\left(1,\frac{p_{m'm}p_{m'}f_{m'}(z')}{p_{mm'}p_mf_m(z)}\left|\det\left(g'_{mm'}(z)\right)\right|\right). \tag{11.93}$$

In the example above, the Jacobian is equal to

$$\begin{aligned}\left|\frac{\partial g_{mm'}(z)}{\partial z}\right| &= \left|\frac{\partial\left(\log(\alpha\beta/\sqrt{1+1/\alpha}),\log(1+1/\alpha)\right)}{\partial(\alpha,\beta)}\right|\\ &= \left[\alpha\beta(\alpha+1)\right]^{-1},\end{aligned}$$

and the acceptance probability of moving from the gamma to the lognormal model is

$$a(z, z') = \min\left(1, \frac{f_2(\mu, \sigma^2, M = 2|y)p_{m'm}}{f_1(\alpha, \beta, M = 1|y)p_{mm'}} \left[\alpha\beta(\alpha + 1)\right]^{-1}\right).$$

In terms of the rate of convergence and degree of autocorrelation of the Markov chain, it is difficult a priori to determine which of the two approaches is to be preferred.
This example illustrates that the Jacobian is not an inherent component of dimension-changing MCMC. The Jacobian arises due to the deterministic transformation used in the proposal mechanism, and the change of variable used when equating (11.69) and (11.70). ■

As a concluding remark on the topic, we wish to draw attention to potential difficulties in successful implementation of the algorithm in highly dimensional problems. The competing models under consideration may have different sets of parameters and the reversible jump machinery simply provides no guidance to generate effective jump proposals. A satisfactory rate of transdimensional jumping may require very delicate tuning. A discussion on this topic can be found in Brooks et al. (2001).

11.8 Data Augmentation

We conclude this chapter with a topic that is particularly relevant in the implementation of MCMC. Imagine that there is interest in obtaining the posterior distribution of a parameter $\boldsymbol{\theta}$. Due to analytical intractability, one chooses to approximate $p(\boldsymbol{\theta}|\mathbf{y})$ using MCMC. Often, the fully conditional posterior distributions $p(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$ do not have a standard form and the MCMC algorithm can be difficult to implement. The idea of data augmentation is to augment with the so-called latent data or missing data $\boldsymbol{\varphi}$, in order to exploit the simplicity of the resulting conditional posterior distributions $p(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{\varphi}, \mathbf{y})$. This is in the same spirit as in the EM algorithm: by increasing the dimensionality of the problem, possibly at the expense of extra computing time, although this is not always the case (Swendsen and Wang, 1987), the problem is simplified algorithmically. Notice that the focus of inference is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{\varphi}, \mathbf{y})\, p(\boldsymbol{\varphi}|\mathbf{y})\, d\boldsymbol{\varphi},$$

and this marginalization is carried out via MCMC. A key paper is Tanner and Wong (1987).

Example 11.6 *Inference from truncated data*
Consider a data set (the observed data) consisting of N_o independent obser-

vations from a truncated normal distribution, where the truncation point T is known. Out of a total of N_{To} original observations, each one of the N_o is kept because its value is larger than T. It is also known that there are $N_m = N_{To} - N_o$ missing observations and the only information available on these is that they are i.i.d., that sampling was at random, and that each one is smaller than or equal to T.

The N_{To} original observations are assumed to be independently and normally distributed, with mean μ and variance σ^2. The observed data are denoted by the vector $\mathbf{y}$ of length N_o; the missing data by the vector $\mathbf{z}$ of length N_m. The objective of inference is to characterize μ and σ^2 with the data available.

The p.d.f. of $\mathbf{y}$ is given by the product of truncated normal distributions. The contribution to the likelihood from each element of $\mathbf{y}$ is

$$\begin{aligned} L\left(\mu, \sigma^2 | y_i > T\right) &\propto \frac{p\left(y_i|\mu, \sigma^2\right)}{\int_T^\infty p\left(y_i|\mu, \sigma^2\right) dy_i} \\ &= \frac{p\left(y_i|\mu, \sigma^2\right)}{\left[1 - \Phi\left(\frac{T-\mu}{\sigma}\right)\right]}, \quad i = 1, \ldots, N_o, \end{aligned} \tag{11.94}$$

where $p\left(\cdot|\mu, \sigma^2\right)$ is the p.d.f. of the normal distribution and $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution.

The contribution to the likelihood from each element of $\mathbf{z}$, z_j $(j = 1, \ldots, N_m)$, is

$$\begin{aligned} L\left(\mu, \sigma^2 | z_j \leq T\right) &\propto P\left(z_j \leq T|\mu, \sigma^2\right) \\ &= \int_{-\infty}^{T} p\left(z_j|\mu, \sigma^2\right) dz_j \\ &= \Phi\left(\frac{T-\mu}{\sigma}\right). \end{aligned} \tag{11.95}$$

By virtue of independence, the likelihood is

$$L\left(\mu, \sigma^2|\mathbf{y}\right) \propto \frac{\prod_{i=1}^{N_o} p\left(y_i|\mu, \sigma^2\right)}{\left[1 - \Phi\left(\frac{T-\mu}{\sigma}\right)\right]^{N_o}} \left[\Phi\left(\frac{T-\mu}{\sigma}\right)\right]^{N_m}. \tag{11.96}$$

Assuming independent uniform prior distributions for μ and σ^2, the joint posterior distribution $p\left(\mu, \sigma^2|\mathbf{y}\right)$ is proportional to (11.96). Implementation of the Gibbs sampler requires drawing samples from $p\left(\mu|\sigma^2, \mathbf{y}\right)$ and from $p\left(\sigma^2|\mu, \mathbf{y}\right)$. It is clear from (11.96) that these fully conditional posterior distributions do not reduce to standard form. To facilitate the problem algorithmically one can augment with the missing data $\mathbf{z}$. The complete

data (observed data + missing data) are denoted by $\mathbf{x}' = (\mathbf{z}', \mathbf{y}')$. Thus, the complete data vector $\mathbf{x}$ has i.i.d. elements each with distributional form

$$x_i \sim N\left(\mu, \sigma^2\right).$$

The observed data can be envisaged as being generated in the following manner:

$$y_i|\mu, \sigma^2, y_i > T \sim N\left(\mu, \sigma^2\right) I\left(x_i > T\right) \tag{11.97}$$

and the missing data

$$z_i|\mu, \sigma^2, z_i \leq T \sim N\left(\mu, \sigma^2\right) I\left(x_i \leq T\right), \tag{11.98}$$

where $I\left(x \in A\right)$ is the indicator function, which takes the value 1 if x is contained in the set A, and zero otherwise.
The density of the complete data is

$$p\left(\mathbf{x}|\mu, \sigma^2\right) \propto \prod_{i=1}^{N_{To}} \left[p\left(x_i|\mu, \sigma^2\right) I\left(x_i \leq T\right) + p\left(x_i|\mu, \sigma^2\right) I\left(x_i > T\right)\right]. \tag{11.99}$$

The augmented posterior of the parameters takes the form

$$\begin{aligned} p\left(\mu, \sigma^2, \mathbf{z}|\mathbf{y}\right) &\propto p\left(\mu, \sigma^2, \mathbf{z}\right) p\left(\mathbf{y}|\mu, \sigma^2, \mathbf{z}\right) \\ &= p\left(\mu, \sigma^2\right) p\left(\mathbf{y}, \mathbf{z}|\mu, \sigma^2\right), \end{aligned}$$

which, assuming independent uniform prior distributions for μ and σ^2, is proportional to (11.99). The Gibbs sampler run under the augmentation scheme involves drawing from $p\left(\mathbf{z}|\mu, \sigma^2, \mathbf{y}\right)$, from $p\left(\sigma^2|z, \mu, \mathbf{y}\right)$, and from $p\left(\mu|\mathbf{z}, \sigma^2, \mathbf{y}\right)$.
To derive $p\left(\mathbf{z}|\mu, \sigma^2, \mathbf{y}\right)$, one retains in (11.99) those terms that include $\mathbf{z}$. Therefore, from (11.98) and (11.99),

$$p\left(\mathbf{z}|\mu, \sigma^2, \mathbf{y}\right) \propto \prod_{j=1}^{N_m} p\left(x_j|\mu, \sigma^2\right) I\left(x_j \leq T\right). \tag{11.100}$$

This has the form of a left-truncated normal distribution, with mean μ and variance σ^2, where the truncation point is T.
The fully conditional posterior distribution of σ^2 is

$$\begin{aligned} p\left(\sigma^2|z, \mu, \mathbf{y}\right) &\propto \prod_{i=1}^{N_{To}} p\left(x_i|\mu, \sigma^2\right) \\ &\propto \left(\sigma^2\right)^{-\frac{N_{To}}{2}} \exp\left[-\frac{\sum_{i=1}^{N_{To}} \left(x_i - \mu\right)^2}{2\sigma^2}\right]. \end{aligned}$$

This is the kernel of a scaled inverted chi-square distribution, with scale parameter $\sum_{i=1}^{N_{To}} (x_i - \mu)^2$ and $N_{To} - 2$ degrees of freedom. Therefore,

$$\sigma^2 | z, \mu, \mathbf{y} \sim \left[\sum_{i=1}^{N_{To}} (x_i - \mu)^2 \right] \chi^{-2}_{N_{To}-2}. \tag{11.101}$$

Finally, the fully conditional posterior distribution of μ is

$$p\left(\mu | \sigma^2 \mathbf{z}, \mathbf{y}\right) \propto \prod_{i=1}^{N_{To}} p\left(x_i | \mu, \sigma^2\right).$$

This is proportional to $\exp\left[-\frac{\sum_{i=1}^{N_{To}} (x_i - \mu)^2}{2\sigma^2}\right]$. Adding and subtracting $\overline{x} = \sum_{i=1}^{N_{To}} x_i / N_{To}$ in the squared term yields

$$\exp\left[\frac{\sum_{i=1}^{N_{To}} \left[(x_i - \overline{x}) + (\overline{x} - \mu)\right]^2}{2\sigma^2}\right].$$

Retaining only terms in μ one obtains

$$\mu | \sigma^2 \mathbf{z}, \mathbf{y} \sim N\left(\overline{x}, \frac{\sigma^2}{N_{To}}\right). \tag{11.102}$$

The Gibbs sampling algorithm consists of drawing repeatedly from (11.100), from (11.101), and finally from (11.102). ■

Example 11.7 *ABO blood groups*

Gene frequencies of the ABO blood group data were inferred using maximum likelihood implemented via Newton-Raphson in Example 4.7 of Chapter 4 and implemented via the EM algorithm in Example 9.2 of Chapter 9. Here a Bayesian MCMC approach is implemented via data augmentation, using the data as in Example 4.7.

Assume a Dirichlet prior with parameters $(\alpha_A, \alpha_B, \alpha_O)$ for the gene frequencies. On the basis of the data in Table 4.1 the joint posterior distribution $f\left(p_A, p_B, p_O | \mathbf{n}\right)$ is proportional to

$$\left[p_A^2 + 2p_A p_O\right]^{n_A} \left[2p_A p_B\right]^{n_{AB}} \left[p_B^2 + 2p_B p_O\right]^{n_B} \left[p_O^2\right]^{n_O}$$
$$\left[p_A\right]^{\alpha_A - 1} \left[p_B\right]^{\alpha_B - 1} \left[p_O\right]^{\alpha_O - 1},$$

where $\mathbf{n} = (n_A, n_{AB}, n_B, n_O)'$ is the observed data. It is not possible to extract standard fully conditional posterior distributions for p_A, p_B, and p_O from this joint posterior. Augmenting with the missing counts

$$\mathbf{n}_m = (n_{AO}, n_{AA}, n_{BB}, n_{BO})$$

yields the following augmented posterior distribution

$$f(\mathbf{n}_m, p_A, p_B, p_O|\mathbf{n}) \propto f(\mathbf{n}_m, p_A, p_B, p_O) f(\mathbf{n}|\mathbf{n}_m, p_A, p_B, p_O) = f(\mathbf{n}, \mathbf{n}_m|p_A, p_B, p_O) f(p_A, p_B, p_O). \quad (11.103)$$

The augmented posterior (11.103) has the form

$$[p_A]^{2n_{AA}} [2p_A p_O]^{n_{AO}} [2p_A p_B]^{n_{AB}} [p_B]^{2n_{BB}} [2p_B p_O]^{n_{BO}} [p_O]^{2n_O} [p_A]^{\alpha_A - 1} [p_B]^{\alpha_B - 1} [p_O]^{\alpha_O - 1}, \quad (11.104)$$

which is proportional to

$$\begin{aligned} &[p_A]^{2n_{AA}} [p_A]^{n_{AO}} [p_A]^{n_{AB}} [p_A]^{\alpha_A - 1} \\ &[p_B]^{2n_{BB}} [p_B]^{n_{AB}} [p_B]^{n_{BO}} [p_B]^{\alpha_B - 1} \\ &[p_O]^{2n_O} [p_O]^{n_{AO}} [p_O]^{n_{BO}} [p_O]^{\alpha_O - 1} \\ = &[p_A]^{2n_{AA} + n_{AO} + n_{AB} + \alpha_A - 1} [p_B]^{2n_{BB} + n_{AB} + n_{BO} + \alpha_B - 1} \\ &[p_O]^{2n_O + n_{AO} + n_{BO} + \alpha_O - 1}. \end{aligned}$$

From this expression, the joint conditional posterior distribution

$$[p_A, p_B, p_O|\mathbf{n}_m, \mathbf{n}]$$

is immediately recognized as Dirichlet, with parameters $a = 2n_{AA} + n_{AO} + n_{AB} + \alpha_A$, $b = 2n_{BB} + n_{AB} + n_{BO} + \alpha_B$, and $c = 2n_O + n_{AO} + n_{BO} + \alpha_O$; that is,

$$p_A, p_B, p_O|\mathbf{n}_m, \mathbf{n} \sim Di(a, b, c). \quad (11.105)$$

To derive the fully conditional posterior distribution of n_{AA}, first write $n_{AO} = n_A - n_{AA}$ and extract the terms in n_{AA} from (11.104). This yields

$$n_{AA}|p_A, p_O, n_A \sim Bi\left(\frac{p_A^2}{p_A^2 + 2p_A p_O}, n_A\right), \quad (11.106)$$

with $n_{AO} = n_A - n_{AA}$. Similarly,

$$n_{BB}|p_B, p_O, n_B \sim Bi\left(\frac{p_B^2}{p_B^2 + 2p_B p_O}, n_B\right), \quad (11.107)$$

with $n_{BO} = n_B - n_{BB}$.

The Gibbs sampling algorithm defined by (11.105), (11.106), and (11.107), was run using a chain length equal to 3500. After discarding the first 500 samples, the mean and standard deviation of the marginal posterior distributions were estimated from the remaining 3000 draws. Choosing $\alpha_A = \alpha_B = \alpha_0 = 2$, the Monte Carlo estimate of the posterior means are for $p_A = 0.20925$ and for $p_B = 0.08102$. The corresponding posterior standard deviations are 0.0066312 and 0.0043504 (for p_0 the posterior

standard deviation is 0.0073635). Choosing $\alpha_A = \alpha_B = \alpha_0 = 1$ (leading to a uniform prior for the gene frequencies) yields estimates of posterior means for $p_A = 0.20915$ and for $p_B = 0.08084$; the corresponding posterior standard deviations are 0.0066315 and 0.0043533 (the posterior standard deviation of p_0 is 0.0073300). ■

This page intentionally left blank

12

Implementation and Analysis of MCMC Samples

12.1 Introduction

The typical output of a Bayesian MCMC analysis consists of correlated samples from the joint posterior distribution of all parameters of a single model or of a number of models. Using these samples, the analyst may be interested in estimating various features of the posterior distribution. These could include quantiles or moments of marginal posterior distributions of the parameters or of functions thereof.

Three partly related questions that could be posed in the analysis of MCMC output are:

- Can the simulated or sampled values be considered to be draws from the posterior distribution of interest?
- Are estimates of features of the posterior distribution precise enough?
- Have the empirical averages computed from the Monte Carlo output converged to their expectation under the equilibrium distribution?

Other pertinent problems involve the possible impropriety of posterior distributions (a potential danger when improper priors are employed), and issues related to the sensitivity of an analysis using the MCMC samples. These points are discussed in this chapter and partly in Subsection 16.2.3 of Chapter 16. First, we discuss the relative advantages and disadvantages of conducting the MCMC analysis using one or several independent chains. Subsequently, the effects of inter-correlation between parameters on the

behavior of the MCMC are illustrated, and some techniques for diagnosing convergence are presented. Another section gives an overview of estimators of features of the posterior distribution and of their Monte Carlo precision. The chapter concludes with a discussion of sensitivity assessment.

12.2 A Single Long Chain or Several Short Chains?

In the early 1990s when MCMC methods entered into the statistical arena, considerable discussion centered on the best ways of running the algorithms. Different implementation strategies affect the serial correlations between successive samples of the same or different parameters within a chain. These correlations, as discussed below, influence the rate of convergence to the stationary distribution and the sampling error of estimates of features of this distribution. In short, the specific implementation adopted can have a profound impact on the efficiency of the computations.

A strategy that was advocated in the early literature (Gelfand and Smith, 1990) but that has fallen in disuse thereafter, consists of running several short independent chains, k say, and on saving the last sample (the mth) from each of the chains. This is known as the multiple-chain or short-chain method. Here, mk samples are generated but only k are kept for the post-MCMC analysis. The method is extremely inefficient because $(100\,(m-1)\,/m)\,\%$ of the samples are discarded and the value of m is at least in the dozens. In addition, the length of the chains was often judged to be insufficient to guarantee convergence of each of the runs to the target distribution.

Current recommendations about implementation strategies in the literature range from running either several long chains (Gelman and Rubin, 1992) or a single, very long one (Geyer, 1992). Supporters of the first approach argue that a comparison of results from several seemingly converged chains might reveal genuine differences, if the chains have not yet approached stationarity. Those in favor of the single, long-chain implementation, believe that this method has better chances of producing samples which properly represent the complete support of the target distribution. In practice, one almost always uses more than a single long run, and convergence is assessed graphically or via more formal tests, as discussed later. To avoid possible influences of the starting values, the initial samples are often discarded. This is usually referred to as the burn-in period. After burn-in, unless correlations between adjacent samples are extremely high, or if storage is a problem, all samples are kept for later processing. A very useful reference where many implementation problems are discussed is the tutorial of Kass et al. (1998).

12.3 Convergence Issues

It was seen in Chapters 10 and 11 that an ergodic Markov chain generated via iterative Monte Carlo converges to its stationary distribution asymptotically. This means that the number of iterations of the chain must approach infinity! In practice, however, one runs a chain which is long enough in some sense, so that the values obtained can be regarded as approximate draws from the posterior distribution of interest. The difficulty resides in determining how long the chain must be. Unfortunately, there is no simple answer to this question. Clearly, if the iterations have not proceeded long enough, the draws may be unrepresentative of the whole support of the target distribution and this will probably result in poor inferences.

12.3.1 Effect of Posterior Correlation on Convergence

A strong intercorrelation between parameters in the posterior distribution hampers the behavior of the MCMC scheme. We start this section with a couple of stylized examples. The first one illustrates how a high posterior correlation between parameters can slow down the motion of the chain towards its equilibrium distribution. The second presents a model in which the two parameters of a model are not identifiable from the likelihood function. If improper priors are used, this leads to a situation where even though the conditional distributions are well defined, the joint posterior does not exist. In this case, a Gibbs sampler will lead to absurd results, even though the "numbers" may appear sensible. When proper priors are adopted, it is shown that the analysis produces Bayesian learning, but the influence of the prior does not dissipate asymptotically. Further, the sampler will move very slowly in the parameter space, because of the high intercorrelation between the poorly identified parameters.

Example 12.1 *A* 2×2 *table*
Consider Example 10.4 from Chapter 10, and following O'Hagan (1994) let

$$p_1 = p_4 = \frac{p}{2}.$$
$$p_3 = p_2 = \frac{1-p}{2}.$$

It can be verified that

$$\begin{aligned} Cov\left(X,Y\right) &= E\left(XY\right) - E\left(X\right)E\left(Y\right) \\ &= \frac{1}{4}\left(2p-1\right), \end{aligned}$$

and that

$$Var\left(X\right) = Var\left(Y\right) = \frac{1}{4}.$$

Therefore, the correlation between X and Y is

$$\rho = 2p - 1.$$

The transition probability matrix $\mathbf{P}_{x|x}$ is

$$\mathbf{P}_{x|x} = \begin{bmatrix} 1-2p(1-p) & 2p(1-p) \\ 2p(1-p) & 1-2p(1-p) \end{bmatrix}.$$

The matrix $\mathbf{P}_{x|x}$ is ergodic provided that $p > 0$ and the unique solution to equation (10.8)

$$\boldsymbol{\pi}'\mathbf{P}_{x|x} = \boldsymbol{\pi}',$$

is

$$\boldsymbol{\pi}' = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Therefore $\boldsymbol{\pi}$ is the unique stationary distribution of the Markov chain. The matrix $\mathbf{P}_{x|x}$ has two eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = \rho^2$. The corresponding eigenvectors are

$$\mathbf{c}_1' = \begin{bmatrix} 1 & 1 \end{bmatrix},$$
$$\mathbf{c}_2' = \begin{bmatrix} -1 & 1 \end{bmatrix}.$$

Then the rate of convergence of the Markov chain can be studied using (10.26),

$$\mathbf{P}_{x|x}^n = \lambda_1^n \mathbf{Q}_1 + \lambda_2^n \mathbf{Q}_2, \quad n = 1, 2, \ldots,$$

which in the present example is equal to

$$\begin{aligned} \mathbf{P}_{x|x}^n &= 1^n \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + \left(\rho^2\right)^n \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}\left(1+\rho^{2n}\right) & \frac{1}{2}\left(1-\rho^{2n}\right) \\ \frac{1}{2}\left(1-\rho^{2n}\right) & \frac{1}{2}\left(1+\rho^{2n}\right) \end{bmatrix}. \end{aligned} \tag{12.1}$$

Recall expression (10.7) from Chapter 10,

$$\boldsymbol{\pi}'^{(n)} = \boldsymbol{\pi}'^{(0)}\mathbf{P}_{x|x}^n,$$

where $\boldsymbol{\pi}^{(0)} = (p^{(0)}, 1-p^{(0)})'$ represents the distribution of the initial state of the Markov chain. It is easy to verify from (12.1) that after n transitions

$$p^{(n)} = \frac{1}{2}\left[1 - \rho^{2n}\left(1 - 2p^{(0)}\right)\right].$$

Thus, for high values of ρ, convergence toward the equilibrium distribution is very slow. To illustrate with an extreme case, setting $\rho = 0.9998$ and

$p^{(0)} = 0.1$ leads, after one transition, to $p^{(1)} = 0.10016$ and, after $n = 1000$ transitions, to $p^{(1000)} = 0.23188$, still a long way from the equilibrium value of 1/2. Another way of illustrating the same phenomenon is to set $n = 1000$ in (12.1), which gives

$$\mathbf{P}_{x|x}^{1000} = \begin{bmatrix} 0.835 & 0.165 \\ 0.165 & 0.835 \end{bmatrix}.$$

This is far from the equilibrium value

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

■

Although correlations as large as 0.9998 are not that common, moderate to high correlations between parameters of the model can be encountered frequently. This induces slow mixing of the chain: successive transitions are strongly correlated and convergence can be very slow. Even if the chain converges to the equilibrium distribution, with poor mixing, using the time-average over the chain under the equilibrium distribution, i.e., estimator (11.6), will result in poor inferences.

In highly parameterized models, relatively small correlations can result in a similar behavior of the chain, because of intercorrelations between parameters. From a practical point of view, autocorrelation between successive samples produces long sequences where little change is detected, misleadingly suggesting that the chain has converged. Another consequence of within-sequence correlation is that it leads to less precise inferences than those obtained from the same number of independent samples. Two strategies are often suggested for ameliorating slow mixing in MCMC implementations: reparameterization of the model, and sampling parameters in blocks, rather than sampling each parameter individually. However, the two strategies often lead to added computational complexity.

Example 12.2 *Identifiability and impropriety of posterior distributions*
This example is based on an exercise in Chapter 5 of Carlin and Louis (1996). Suppose that data y_i, $(i = 1, 2, \ldots, n)$, are independent realizations from a normal distribution with known variance

$$y_i|\theta_1, \theta_2 \sim N(\theta_1 + \theta_2, 1). \tag{12.2}$$

Since the variance is known, observations have been rescaled to have a standard deviation equal to 1. The likelihood of $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ can be written

as

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left[-\frac{n}{2}(\theta_1+\theta_2-\overline{y})^2\right]$$
$$\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\mathbf{N}(\boldsymbol{\theta}-\boldsymbol{\mu})\right], \tag{12.3}$$

where $\overline{y} = \frac{1}{n}\sum y_i$, $\boldsymbol{\mu}' = \left[\frac{\overline{y}}{2}, \frac{\overline{y}}{2}\right]$ and

$$\mathbf{N} = \begin{bmatrix} n & n \\ n & n \end{bmatrix}.$$

Because the matrix $\mathbf{N}$ has rank equal to 1, (12.3) is the kernel of a singular bivariate normal distribution, if viewed as a function of $\boldsymbol{\theta}$. It is clearly impossible to obtain ML estimates of θ_1 and θ_2 from (12.3); on the other hand, the ML estimator of $\theta_1+\theta_2$ is $\overline{y}$. If independent, improper uniform prior distributions are adopted for each of the parameters, the density of the joint posterior distribution of θ_1, θ_2 is

$$p(\theta_1, \theta_2|\mathbf{y}) \propto \exp\left[-\frac{n}{2}(\theta_1+\theta_2-\overline{y})^2\right]. \tag{12.4}$$

It can be verified readily that the densities of the conditional posterior distributions are

$$\theta_1|\theta_2, \mathbf{y} \sim N\left(\overline{y}-\theta_2, \frac{1}{n}\right) \tag{12.5}$$

and

$$\theta_2|\theta_1, \mathbf{y} \sim N\left(\overline{y}-\theta_1, \frac{1}{n}\right). \tag{12.6}$$

The marginal posterior density of θ_1, say, is

$$p(\theta_1|\mathbf{y}) = \frac{p(\theta_1, \theta_2|\mathbf{y})}{p(\theta_2|\theta_1, \mathbf{y})}$$
$$= \frac{c_{12}\exp\left[-\frac{n}{2}(\theta_1+\theta_2-\overline{y})^2\right]}{c_{2|1}\exp\left[-\frac{n}{2}(\theta_1+\theta_2-\overline{y})^2\right]}$$
$$= \frac{c_{12}}{c_{2|1}}, \tag{12.7}$$

where c_{12} and $c_{2|1}$ are the constants of integration associated with (12.4) and (12.6), respectively, assuming c_{12} is finite. Hence, this marginal posterior distribution does not depend on θ_1. Clearly, the integral of (12.7) over the real line (the sample space of θ_1) does not converge, indicating that the marginal posterior distribution is improper.

In this setting, the parameters θ_1 and θ_2 in (12.4) are unidentifiable from

each other: the posterior distribution carries information on their sum, $\theta_1 + \theta_2$, but not on each one of them separately. Since the fully conditional posterior distributions (12.5) and (12.6) are proper, a Gibbs sampling implementation of the model is possible, yielding valid inferences about features of the distribution $[\theta_1 + \theta_2|\mathbf{y}]$. However, one could naively use the samples from the chains generated from (12.5) and (12.6), to "infer" features of the marginal posterior distribution of θ_1 or of θ_2. This would lead to meaningless results, since these marginal distributions do not exist. This illustrates the pitfall of using improper priors in hierarchical models. Except in highly stylized models (such as in this example) it is very difficult to assess impropriety analytically (e.g., Hobert and Casella, 1996). All conditional posterior distributions may exist even when the joint distribution is not defined. Yet, the output analysis may produce seemingly "reasonable" results!

Now consider assigning independent normal distributions, a priori, to each of θ_1 and θ_2. Take

$$\theta_i \sim N\left(a_i, b_i^2\right), \quad i = 1, 2,$$

such that the joint prior density is

$$p\left(\theta_1, \theta_2\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \mathbf{a}\right)' \mathbf{B}\left(\boldsymbol{\theta} - \mathbf{a}\right)\right], \tag{12.8}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, $\mathbf{a} = (a_1, a_2)'$, and

$$\mathbf{B} = \begin{bmatrix} \frac{1}{b_1^2} & 0 \\ 0 & \frac{1}{b_2^2} \end{bmatrix}.$$

With proper prior distributions, the problem of the lack of identifiability of the parameters in the posterior distribution disappears. Now using (12.8) and (12.3) the joint posterior density is

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\theta} - \mathbf{a}\right)' \mathbf{B}\left(\boldsymbol{\theta} - \mathbf{a}\right) + \left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)' \mathbf{N}\left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)\right]\right\}. \tag{12.9}$$

Combining these two quadratic forms using results in Box and Tiao (1973), page 418, and keeping only the terms which are functions of $\boldsymbol{\theta}$, leads to

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right)' \left(\mathbf{B} + \mathbf{N}\right)\left(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right)\right]. \tag{12.10}$$

This is the kernel of the density of a bivariate normal distribution with mean vector $\overline{\boldsymbol{\theta}}$ and variance–covariance matrix $(\mathbf{B}+\mathbf{N})^{-1}$, where

$$\begin{aligned}\overline{\boldsymbol{\theta}} &= (\mathbf{B}+\mathbf{N})^{-1}(\mathbf{B}\mathbf{a}+\mathbf{N}\boldsymbol{\mu}) \\ &= \begin{bmatrix} n+\dfrac{1}{b_1^2} & n \\ n & n+\dfrac{1}{b_2^2} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{a_1}{b_1^2}+n\overline{y} \\ \dfrac{a_2}{b_2^2}+n\overline{y} \end{bmatrix}. \end{aligned} \tag{12.11}$$

After some algebra, the posterior mean vector (12.11) can be written as:

$$\begin{bmatrix} \overline{\theta}_1 \\ \overline{\theta}_2 \end{bmatrix} = \begin{bmatrix} a_1+k_1(\overline{y}-a_1-a_2) \\ a_2+k_2(\overline{y}-a_1-a_2) \end{bmatrix}, \tag{12.12}$$

where

$$k_1 = \frac{b_1^2}{b_1^2+b_2^2+\frac{1}{n}},$$

and

$$k_2 = \frac{b_2^2}{b_1^2+b_2^2+\frac{1}{n}}.$$

The posterior variance–covariance matrix is

$$\begin{aligned} &Var(\boldsymbol{\theta}|\mathbf{y}) \\ &= \begin{bmatrix} n+\dfrac{1}{b_1^2} & n \\ n & n+\dfrac{1}{b_2^2} \end{bmatrix}^{-1} = \begin{bmatrix} b_1^2(1-k_1) & -b_1^2k_2 \\ -b_1^2k_2 & b_2^2(1-k_2) \end{bmatrix}. \end{aligned} \tag{12.13}$$

There are a number of important conclusions that can be drawn from this exercise. First, note that the marginal posterior distribution of θ_i is now proper, and that it differs from the prior distribution. In this sense, there is Bayesian learning via the data. From (12.12), it is apparent that the influence of the data on the posterior mean depends on the values of the prior variances, via the "regression" $k_i = b_i^2/\left(b_1^2+b_2^2+\frac{1}{n}\right)$. Second, (12.13) indicates that the variance of the marginal posterior distribution is smaller than the prior variance. Third, (12.12) and (12.13) illustrate that as the number of observations $n \to \infty$, the influence of the prior does not vanish asymptotically. The posterior variance does not tend to zero, and inferences always depend on the relative values of the prior variances b_1^2/b_2^2. For example (for large n), if $b_1^2 >> b_2^2$, the posterior mean of θ_1 will tend to $\overline{y}-a_2$, whereas the posterior mean of θ_2 will be close to a_2, its prior mean. Here, Bayesian learning occurs for θ_1, but not for θ_2. Finally, from (12.13), the posterior correlation between θ_1 and θ_2 is given by:

$$Corr(\theta_1,\theta_2|\mathbf{y}) = -\frac{b_1b_2}{\sqrt{\left(\frac{1}{n}+b_1^2\right)\left(\frac{1}{n}+b_2^2\right)}}, \tag{12.14}$$

which is very close to -1 for moderately large n. From the point of view of implementing a Gibbs sampler, this has important implications. While the proper prior distributions of θ_1 and θ_2 lead to proper marginal posterior distributions, the very high posterior correlation between θ_1 and θ_2 will generate a strong serial correlation between samples of a Gibbs chain. This will have sizable effects on convergence, and will retard the movement of the Gibbs sampler over the support of the posterior distribution. In such a situation, the quality of posterior inferences would be impaired seriously. A broad discussion on strategies for improving MCMC can be found in Gilks and Roberts (1996). ■

MCMC opens the opportunity for fitting complex hierarchical models to data, and these models perhaps describe better the biological system under study. However, the richness and flexibility are accompanied by caveats. For example, it may be dangerous to entertain models that are not well understood analytically. In highly complex models, there is always the pitfall that parameters may be unidentified or very weakly identified. In contrast with the preceding example, it may not always be possible to detect lack of identifiability. Therefore, it is important to learn as much as possible about the model, and to experiment with it step by step before launching a full MCMC-based analysis.

12.3.2 Monitoring Convergence

A large literature on convergence diagnostics has developed during the last decade. Useful reviews can be found in Cowles and Carlin (1996), Brooks and Roberts (1998), Robert (1998), Robert and Casella (1999), Mengersen et al. (1999), and references herein. In this section some of the commonly used convergence diagnostics are described, and the reader is referred to the above reviews for a description of other methods.

Graphical Procedures

Gelfand et al. (1990) suggested informal convergence checks based on graphical techniques. They run a number of independent chains with variable starting values and perform the analysis for each of the parameters of interest. Plots of histograms are overlaid for increasing chain lengths, until the graphs become visually indistinguishable among chains. Stability of the results is assessed by increasing chain length. Another simple graphical tool for studying convergence and mixing behavior of the chain is what is called the "trace plot". Samples of the parameter of interest, from replicated chains started from overdispersed values, are plotted as a function of iterate number. Wavy patterns typically indicate strong autocorrelations within chains, while a zigzag suggest that the parameter moves more freely. In highly dimensional problems, however, it is not feasible to examine the

time series plots of all the parameters. In this case, a selective choice is often made, such that either the focus of inference or a high-level parameter in a hierarchy (e.g., variance components in a generalized linear model) are followed, since these tend to mix more slowly. Here, it is important to be aware that parameters converge at different rates. Further, a slow convergence of nuisance parameters may affect convergence of the parameters of interest adversely. Also, an assessment of convergence of the marginal distributions does not provide an exhaustive diagnostic of convergence of the joint process.

Auto-Correlograms

Examination of lag-t autocorrelations is an easy way of monitoring the mixing behavior of the Markov chain. For instance, one can plot the autocorrelations as a function of the lag (this is called a correlogram), and detect the lag at which the correlation "dies out". As indicated below, these autocorrelations are also useful to estimate the effective chain length (Sorensen et al., 1995).

Between and within-chain Variability of Sample Values

A popular quantitative convergence diagnostic was suggested by Gelman and Rubin (1992). The method involves running m independent chains, each of length $2n$. Each of the chains starts from a different starting point from a distribution that is overdispersed with respect to the target distribution. This is an important design feature, because it can make lack of convergence apparent. As discussed in Gelman and Rubin (1992), satisfying this requirement may involve considerable initial effort toward eliciting a rough estimate of the variance of the marginal posterior distribution of the scalar of interest. This knowledge is important for generating an overdispersed starting sequence. Starting from points far apart will also ensure that the complete support of the target distribution will be visited. The first n iterations of each chain are discarded and the last n are retained. Each feature of interest from the target distribution is monitored separately. The following step consists of carrying out an analysis of variance: approximate convergence is diagnosed when the variability between chains is not larger than that within chains. The rationale of the method is that before convergence is reached, due to the initial over-dispersion, the variance between chains should be relatively large. On the other hand, variation within chains would be relatively small because in the intermediate stages of the iteration the support of the target distribution would be incompletely represented in the simulations.

Suppose that for chain i $(i = 1, 2, \ldots, m)$, simulated values

$$\theta_{ij}, \quad j = 1, 2, \ldots, n$$

are available from the scalar posterior distribution $[\theta|\mathbf{y}]$. Hence, there are m chains each of length n. The simulated values can be organized as a one-way layout, with m classes and n observations per class. The between-chain mean square B and the within-chain mean square W are

$$B = \frac{n \sum_{i=1}^{m} \left(\overline{\theta}_{i.} - \overline{\theta}_{..}\right)^2}{m-1},$$

and

$$W = \frac{\sum_{i=1}^{m} S_i^2}{m}.$$

Here

$$\overline{\theta}_{i.} = \frac{\sum_{j=1}^{n} \theta_{ij}}{n}, \qquad \overline{\theta}_{..} = \frac{\sum_{i=1}^{m} \overline{\theta}_{i.}}{m}$$

are the within-chain sample average and the mean of the chain averages, respectively, and

$$S_i^2 = \frac{\sum_{j=1}^{n} \left(\theta_{ij} - \overline{\theta}_{i.}\right)^2}{n-1}$$

is the estimated variance of sampled values in chain i. Let

$$\mu = \int \theta p\left(\theta|\mathbf{y}\right) d\theta$$

and

$$\sigma^2 = \int \left(\theta - \mu\right)^2 p\left(\theta|\mathbf{y}\right) d\theta,$$

be the mean and variance of the target posterior distribution, respectively. If the simulated values are drawn from $[\theta|\mathbf{y}]$, it can be verified readily that the following expectations hold:

$$E\left(B\right) = \sigma^2, \tag{12.15}$$

and

$$E\left(S_i^2\right) = E\left(W\right) = \sigma^2. \tag{12.16}$$

Gelman and Rubin (1992) suggest the estimator of the posterior variance

$$\widehat{\sigma^2} = \frac{n-1}{n} W + \frac{1}{n} B, \tag{12.17}$$

which is clearly unbiased for σ^2, provided that all draws are from the target distribution. On the other hand, if there is at least some initial overdispersion, the estimator will have an upward bias because the averages $\overline{\theta}_{i.}$ would

vary more than if drawn from the same distribution. This suggests that if the initial draws are not overdispersed enough, $\widehat{\sigma^2}$ can be too low, falsely diagnosing convergence. Gelman and Rubin (1992) suggest that convergence of any scalar quantity of interest can be monitored by the quantity

$$\widehat{R} = \sqrt{\frac{\widehat{\sigma^2}}{W}} = \sqrt{1 + \frac{1}{n}\left(\frac{B}{W} - 1\right)} \tag{12.18}$$

which is expected to be larger than 1, and declines to 1 as $n \to \infty$. Therefore, convergence can be evaluated by examining the proximity of $\widehat{R}$ to 1. Gelman and Rubin (1992) mention that values of $\widehat{R}$ around 1.2 may be satisfactory for most problems. In some cases, however, a higher level of precision may be required. Although normality is not required for unbiasedness of $\widehat{\sigma^2}$, Gelman and Rubin (1992) state that the method works better if the posterior distribution is nearly normal.

A number of shortcomings of the method have been raised. First, constructing a distribution for sampling starting values may be difficult, especially in models in which multimodality is expected. Second, discarding the first n samples is computationally wasteful. Third, the method relies to some extent on approximate normality of the target distribution, and this may not always hold, specially in finite sample situations. Fourth, the procedure will not work if the chains get "trapped" within the same subregion of the parameter space. Finally, the convergence criterion is unidimensional; hence, it gives an inadequate evaluation of convergence to the joint distribution. Brooks and Gelman (1998) have developed a multiparameter version of this approach. These criticisms must be seen in the light of the fact that none of the many methods proposed can be relied upon unilaterally. In a review of convergence diagnosis methods, Cowles and Carlin (1996) concluded that all procedures can fail to detect the sort of convergence failure that they were designed to identify. Therefore, the general recommendation is to use a combination of approaches as diagnostic tools (including graphical methods) and to learn as much as possible from the target distribution before embarking in a MCMC algorithm. As a minimum, ensuring propriety of the posterior distribution is essential.

12.4 Inferences from the MCMC Output

12.4.1 Estimators of Posterior Quantities

The MCMC output is typically used to estimate features of the posterior distribution, such as posterior means and medians, or the posterior variance. While the issue of convergence of the Markov chain to the target distribution is of fundamental importance, many authors place emphasis

on the properties of estimators of features of the posterior distribution. For example, it is of interest to establish whether or not the difference between estimates of a posterior mean obtained from two or more independent chains, can be explained by Monte Carlo sampling error. In this section, two commonly used estimators of features of posterior distributions will be presented. Other estimators are described in Robert and Casella (1999) and in Chen et al. (2000), where a more formal treatment of the subject can be found.

Ergodic Averages

Consider a single chain consisting of correlated samples $\theta^{(i)}$ $(i = 1, 2, \ldots, n)$ from the target distribution $[\theta|\mathbf{y}]$. As presented in the previous chapter (expression (11.5)), for some function $h(\theta)$, the ergodic theorem states that the ergodic average of the function $h(\theta)$, given by

$$\frac{1}{n}\sum_{i=1}^{n} h\left(\theta^{(i)}\right) \tag{12.19}$$

is a consistent estimator of $\int h(\theta)\, p(\theta|\mathbf{y})\, d\theta$, provided that this integral converges. That is,

$$\frac{1}{n}\sum_{i=1}^{n} h\left(\theta^{(i)}\right) \to \int h(\theta)\, p(\theta|\mathbf{y})\, d\theta \tag{12.20}$$

as $n \to \infty$, with probability 1, if $\int h(\theta)\, p(\theta|\mathbf{y})\, d\theta < \infty$. As mentioned above, the rate of convergence may be seriously affected by slow mixing of the chain. At any rate, (12.19) is a consistent estimator of the expected value of $h(\theta)$ with respect to the invariant distribution $[\theta|\mathbf{y}]$, despite any existing autocorrelation between the simulated values $\theta^{(i)}$ of the Markov chain. For example, expression (12.19) is an estimator of:

- the posterior mean if $h(\theta) = \theta$;
- the posterior variance if $h(\theta) = [\theta - E(\theta|\mathbf{y})]^2$. The estimator of the posterior variance is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\theta^{(i)2} - \left[\widehat{E}(\theta|\mathbf{y})\right]^2\right],$$

  where $\widehat{E}(\theta|\mathbf{y}) = \frac{1}{n}\sum_i \theta^{(i)}$;
- the posterior probability that $\theta \in A$, if $h(\theta) = I(\theta \in A)$, such that

$$\Pr(\theta \in A|\mathbf{y}) = \int I(\theta \in A)\, p(\theta|\mathbf{y})\, d\theta.$$

Here, the estimator is $\sum_{i=1}^{n} I\left(\theta^{(i)} \in A\right)/n$. As a special case, the cumulative distribution function is estimated as

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(\theta^{(i)} < t\right).$$

Therefore, the estimator of the posterior probability that $t_1 < \theta < t_2$ is

$$\widehat{\Pr}(t_1 < \theta < t_2|\mathbf{y}) = \frac{1}{n}\left[\sum_{i=1}^{n} I\left(t_1 < \theta^{(i)} < t_2\right)\right];$$

- the posterior predictive density:

$$p(z|\mathbf{y}) = \int p(z|\theta, \mathbf{y})\, p(\theta|\mathbf{y})\, d\theta,$$

where, usually, the form of the problem is such that $p(z|\theta, \mathbf{y}) = p(z|\theta)$. In this setting, $h(\theta) = p(z|\theta)$, and the estimator of the predictive density is $\sum_{i=1}^{n} p\left(z|\theta^{(i)}\right)/n$.

Rao–Blackwell Estimator

Another estimator that has been proposed in the literature is known as the Rao-Blackwell estimator (Gelfand et al., 1990; Liu et al., 1994; Casella and Robert, 1996), which derives its name from the Rao–Blackwell theorem. This theorem states that conditioning an unbiased estimator on a sufficient statistic will result in a uniformly better unbiased estimator.

Let the parameter vector of a model consist of two scalars, that is, $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, and suppose that interest focuses on the mean of the marginal posterior distribution of the function $h(\theta_1)$, or $E[h(\theta_1)|\mathbf{y}]$. As mentioned above, the ergodic average estimator is

$$\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_1^{(i)}\right), \tag{12.21}$$

where $\theta_1^{(i)}$ is a sample from $[\theta_1|\mathbf{y}]$. The Rao–Blackwell estimator is obtained using results discussed in Chapter 1, Section 1.6. Recall that

$$\begin{aligned} E[h(\theta_1)|\mathbf{y}] &= \int \left[\int h(\theta_1)\, p(\theta_1|\theta_2, \mathbf{y})\, d\theta_1\right] p(\theta_2|\mathbf{y})\, d\theta_2 \\ &= \int E_{\theta_1|\theta_2,\mathbf{y}}[h(\theta_1)|\theta_2, \mathbf{y}]\, p(\theta_2|\mathbf{y})\, d\theta_2 \\ &= E_{\theta_2|\mathbf{y}}\left\{E_{\theta_1|\theta_2,\mathbf{y}}[h(\theta_1)|\theta_2, \mathbf{y}]\right\}. \end{aligned}$$

The Rao–Blackwell estimator has the form

$$\widehat{E}\left[h\left(\theta_1\right)|\mathbf{y}\right] \approx \frac{1}{n}\sum_{i=1}^{n} E_{\theta_1|\theta_2,\mathbf{y}}\left[h\left(\theta_1\right)|\theta_2^{(i)},\mathbf{y}\right], \tag{12.22}$$

where $\theta_2^{(i)}$is a draw from the marginal posterior distribution $[\theta_2|\mathbf{y}]$. Hence, the estimator is an ergodic average of conditional means, and one must be able to write these in closed form, to be able to form (12.22). Recall from (1.129) that the variance of $h\left(\theta_1|\mathbf{y}\right)$ in (12.21) can be written as

$$\begin{aligned} &Var\left[h\left(\theta_1|\mathbf{y}\right)\right] \\ &= Var\left\{E\left[h\left(\theta_1\right)|\theta_2,\mathbf{y}\right]\right\} + E\left\{Var\left[h\left(\theta_1\right)|\theta_2,\mathbf{y}\right]\right\}. \end{aligned}$$

Therefore, $Var\left\{E\left[h\left(\theta_1\right)|\theta_2,\mathbf{y}\right]\right\} \leq Var\left[h\left(\theta_1|\mathbf{y}\right)\right]$ indicating that (12.22) can improve upon (12.21) in terms of variance. While making optimal use of the available data is a praiseworthy endeavor, the improvement of (12.22) over (12.21) is often limited in chains that have been run long enough. This improvement comes at the cost of having to know the closed form of the expected value of the conditional posterior distribution $[\theta_1|\theta_2,\mathbf{y}]$, and at a loss of the simplicity with which (12.21) is calculated.

Density Estimation

Another way of estimating features of the posterior distribution of a parameter of interest is to obtain a smooth estimate of the posterior density using the chain output, and then computing moments from this density using numerical integration. This approach seems to be in disuse in output analysis, in favor of the simple practice of approximating the density by histograms, and of estimating features from posterior distributions using (12.19) and (12.22), for example. The reader is referred to classical texts on density estimation by Silverman (1992) and by Scott (1992) for a detailed description of this approach.

12.4.2 Monte Carlo Variance

Definition

Here it is assumed that draws from the stationary distribution $[\theta|\mathbf{y}]$ are available. Because only a finite number of these draws can be obtained, there is always sampling uncertainty associated with an estimator of features of the target distribution, such as (12.19). This sampling variance is known as the Monte Carlo (MC) variance of estimators of posterior quantities. In principle, it can be made as small as desired, by taking a sufficiently large number of samples. This MC variance can be estimated by running several independent chains, and then calculating the empirical, between-chain variance of the estimates obtained for each chain. Since this

is often computationally expensive, one resorts to theoretical estimators of MC variance. These estimators account for the autocorrelation among the samples taken from the target distribution. Useful references are Ripley (1987), Geyer (1992), and Chen et al. (2000). Here, two commonly used estimators are described.

Consider estimating the cumulative distribution function

$$F(t|\mathbf{y}) = \Pr(\theta < t|\mathbf{y})$$

from the MCMC output $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$. The estimator is

$$\widehat{F}(t|\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} I\left(\theta^{(i)} < t\right).$$

Now $I\left(\theta^{(i)} < t\right)$ has a Bernoulli distribution with success probability $F(t|\mathbf{y})$. Thus, if the draws $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ were independent,

$$\sum_{i=1}^{n} I\left(\theta^{(i)} < t\right)$$

would have a binomial distribution with parameters $(F(t|\mathbf{y}), n)$. It follows that with independent draws, the estimator of the Monte Carlo variance of $\widehat{F}(t|\mathbf{y})$ would be equal to

$$\widehat{Var}\left[\widehat{F}(t|\mathbf{y})\right] = \frac{1}{n}\widehat{F}(t|\mathbf{y})\left[1 - \widehat{F}(t|\mathbf{y})\right]. \tag{12.23}$$

However, (12.23) may give a very distorted picture of the true Monte Carlo variance, depending on the pattern of the autocorrelation among the samples from the target distribution. As shown by Liu et al. (1994), this pattern can be complex since, in a reversible chain, even-lag autocorrelations are non-negative, while odd-lag auto-covariances need not be positive.

Geyer's Estimator

Geyer (1992) proposed an estimator of the Monte Carlo variance of the estimator of the mean of $h(\theta)$ based on time-series theory. The estimates produced are larger than or equal to the true Monte Carlo variance. First, from (12.19), define the estimator of the mean of $h(\theta)$ as

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta^{(i)}\right). \tag{12.24}$$

Let the lag-t autocovariance of the stationary Markov chain $h\left(\theta^{(i)}\right)$ be

$$\gamma(t) = Cov\left[h\left(\theta^{(i)}\right), h\left(\theta^{(i+t)}\right)\right], \quad i = 1, 2, \ldots, n.$$

An estimator of $\gamma(t)$ is

$$\widehat{\gamma}(t) = \frac{1}{n}\sum_{i=1}^{n-t}\left\{\left[h\left(\theta^{(i)}\right) - \widehat{\mu}\right]\left[h\left(\theta^{(i+t)}\right) - \widehat{\mu}\right]\right\}. \tag{12.25}$$

Priestley (1981) mentions that it has been asserted that in general, this biased estimator with divisor n has smaller mean square error than the unbiased estimator with divisor $n-t$. One of the estimators of the Monte Carlo variance of $\widehat{\mu}$ proposed by Geyer (1992), which he calls the initial positive sequence estimator, uses (12.25) as input and is equal to

$$\widehat{Var}(\widehat{\mu}) = \frac{1}{n}\left[\widehat{\gamma}(0) + 2\sum_{i=1}^{i=2\delta+1}\widehat{\gamma}(i)\right], \tag{12.26}$$

where δ is chosen such that it is the largest integer satisfying

$$\widehat{\gamma}\left(2\delta'\right) + \widehat{\gamma}\left(2\delta'+1\right) > 0, \quad \delta' = 0, 1, \ldots, \delta.$$

If the samples are independent,

$$\widehat{Var}(\widehat{\mu}) = \frac{1}{n}\widehat{\gamma}(0).$$

An idea of the effect of the autocorrelation on the amount of information contained in the chain for inferring features of the target distribution, can be obtained by computing an "effective chain size". Denote this as Ψ (Sorensen et al., 1995), where

$$\Psi = \frac{\widehat{\gamma}(0)}{\widehat{Var}(\widehat{\mu})}.$$

When the chain consists of independent draws from the target distribution, $\Psi = n$, and the effective and nominal sizes of the chain are equal.

Batching

A popular method of estimating Monte Carlo variances that is easy to implement, is known as "batching" (Hastings, 1970). It is based on the idea that if individual draws are correlated, grouping successive draws into b batches or groups of size m each, and computing the raw averages, will lead to b batch means that are less strongly inter-correlated than the original draws. This can be so provided that m is chosen appropriately. Further, the larger the autocorrelation among samples, the larger m must be. Suppose that a chain of total length n is divided into b batches each of size m. Let the average of the ith batch be

$$\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} h\left(\theta^{(j)}\right), \quad i = 1, 2, \ldots, b.$$

Here, $h\left(\theta^{(j)}\right)$ is some feature of the posterior distribution evaluated at the sampled value $\theta^{(j)}$. The batch estimator of the variance of (12.24), assuming that m is large enough so that the $\overline{x}_i'$s are uncorrelated, is equal to

$$\widehat{Var}_b\left(\widehat{\mu}\right) = \frac{\sum_{i=1}^{b}\left(\overline{x}_i - \widehat{\mu}\right)^2}{b\left(b-1\right)}. \tag{12.27}$$

An estimate of the batch-effective chain size can be obtained as

$$\Psi_b = \frac{\sum_{i=1}^{n}\left[h\left(\theta^{(i)}\right) - \widehat{\mu}\right]^2}{\left(n-1\right)\widehat{Var}_b\left(\widehat{\mu}\right)}.$$

When the samples are independent, $m = 1$, $\overline{x}_i = h\left(\theta^{(i)}\right)$ for all i, and $\Psi_b = n$.

If the autocorrelation among the samples of the chain is very high (> 0.95), estimator (12.26) seems to be preferred over (12.27).

MCMC algorithms converge to the target distribution asymptotically and the samples are typically correlated. A new and exciting approach, termed perfect sampling proposed by Propp and Wilson (1996), avoids problems of convergence and of serial correlations, since it generates independent draws from the target distribution. This is an area where research is just beginning; it is not clear at the moment whether the technique can applied in settings involving high dimensional distributions without nice symmetry properties. A tutorial can be found in Casella et al. (2001).

12.5 Sensitivity Analysis

An important part of a Bayesian analysis is the study of how robust are inferences to modeling assumptions, including prior and likelihood specifications, or presence of outliers. Smith and Gelfand (1992) describe how to address this question using importance sampling, a technique that was described by Hammersley and Handscomb (1964). Geweke (1989) showed how importance sampling can be applied in Bayesian analyses. This technique was already encountered in Chapter 8, in connection with estimation of the marginal likelihood from the Monte Carlo samples, and is discussed again in Chapter 15. Other relevant literature on the subject can be found in Tanner and Wong (1987), Rubin (1987b), and Gelfand and Smith (1990).

The starting point of a Bayesian analysis is the posterior density

$$p_1\left(\theta|\mathbf{y}\right) = c_1 p_1\left(\theta\right) p\left(\mathbf{y}|\theta\right),$$

where c_1 is the typically unknown normalizing constant, $p_1\left(\theta\right)$ is the density of some prior distribution assigned to the parameter, and $p\left(\mathbf{y}|\theta\right)$ is the

likelihood. The expectation of a function $h(\theta)$ with respect to the posterior distribution with density $p_1(\theta|\mathbf{y})$ is

$$E_1[h(\theta)] = \int h(\theta)\, p_1(\theta|\mathbf{y})\, d\theta.$$

Suppose that n draws $\theta^{(i)}$, $(i = 1, 2, \ldots, n)$ are available from this posterior distribution. Based on (12.19), $E_1[h(\theta)]$ can be estimated as

$$\widehat{E}_1[h(\theta)] = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta^{(i)}\right). \tag{12.28}$$

Now, one may be interested in inferences about $h(\theta)$, conditionally on the same data, but using a different set of modeling assumptions. These could involve perturbations either of the likelihood (this could take a new functional form, or perhaps part of the data could be omitted) or of the prior distribution. For example, suppose that one wishes to study the consequences of changing the prior specification, such that the new posterior becomes

$$p_2(\theta|\mathbf{y}) = c_2 p_2(\theta)\, p(\mathbf{y}|\theta). \tag{12.29}$$

Here, $p_2(\theta)$ is the density of the "new" prior distribution, and c_2 is the corresponding integration constant. Using the draws $\theta^{(i)}$ generated under the distribution with density $p_1(\theta|\mathbf{y})$, inferences about $h(\theta)$ under the new posterior with density $p_2(\theta|\mathbf{y})$ can be obtained without having to run the MCMC procedure again. This is done by using $p_1(\theta|\mathbf{y})$ as importance sampling density. Thus, expectations under $p_2(\theta|\mathbf{y})$ can be obtained as follows

$$\begin{aligned} E_2[h(\theta)] &= \frac{\int h(\theta)\, \frac{p_2(\theta|\mathbf{y})}{p_1(\theta|\mathbf{y})} p_1(\theta|\mathbf{y})\, d\theta}{\int \frac{p_2(\theta|\mathbf{y})}{p_1(\theta|\mathbf{y})} p_1(\theta|\mathbf{y})\, d\theta} \\ &= \frac{\int h(\theta)\, \frac{c_2 p_2(\theta) p(\mathbf{y}|\boldsymbol{\theta})}{c_1 p_1(\theta) p(\mathbf{y}|\boldsymbol{\theta})} p_1(\theta|\mathbf{y})\, d\theta}{\int \frac{c_2 p_2(\theta) p(\mathbf{y}|\boldsymbol{\theta})}{c_1 p_1(\theta) p(\mathbf{y}|\boldsymbol{\theta})} p_1(\theta|\mathbf{y})\, d\theta} \\ &= \frac{\int h(\theta)\, w(\theta)\, p_1(\theta|\mathbf{y})\, d\theta}{\int w(\theta)\, p_1(\theta|\mathbf{y})\, d\theta}, \end{aligned} \tag{12.30}$$

where

$$w(\theta) = \frac{p_2(\theta)}{p_1(\theta)}.$$

Note that in the second line of (12.30) the ratio of constants of integration and the likelihood cancel out in the numerator and denominator. A consistent estimator of (12.30) based on (12.19) is

$$\widehat{E}_2[h(\theta)] = \frac{\sum_{i=1}^{n} h\left(\theta^{(i)}\right) w\left(\theta^{(i)}\right)}{\sum_{i=1}^{n} w\left(\theta^{(i)}\right)}, \tag{12.31}$$

where $\theta^{(i)}$, $(i = 1, 2, \ldots, n)$ are the draws from the distribution with density $p_1(\theta|\mathbf{y})$. The weight function w_i is equal to

$$w\left(\theta^{(i)}\right) = \frac{p_2\left[\theta^{(i)}\right]}{p_1\left[\theta^{(i)}\right]}.$$

If $w_i = 1$ for all i, $p_2(\theta|\mathbf{y}) = p_1(\theta|\mathbf{y})$ and (12.30) is equal to (12.28).

Moments and quantiles under the new posterior distribution can be obtained along the same lines, using the draws from the original posterior distribution. For instance

$$\begin{aligned}\widehat{Var}_2[h(\theta)] &= \widehat{E}\left[h^2(\theta)\right] - \left[\widehat{E}[h(\theta)]\right]^2 \\ &= \frac{\sum_{i=1}^{n} h^2\left[\theta^{(i)}\right] w\left(\theta^{(i)}\right)}{\sum_{i=1}^{n} w\left(\theta^{(i)}\right)} - \left[\widehat{E}[h(\theta)]\right]^2 \end{aligned} \tag{12.32}$$

and

$$\widehat{\Pr}_2[h(\theta) < t] = \frac{\sum_{i=1}^{n} I\left[h\left(\theta^{(i)}\right) < t\right] w\left(\theta^{(i)}\right)}{\sum_{i=1}^{n} w\left(\theta^{(i)}\right)}, \tag{12.33}$$

where subscript 2 indicates that inferences are being drawn from the posterior distribution with density $p_2[\theta|\mathbf{y}]$.

Often, it can be computationally advantageous to fit a particular model elicited under a certain prior or likelihood specification. However, the analyst may have in mind an alternative model which is less tractable computationally. The approach described above provides a powerful tool for doing this in a rather straightforward manner. This is illustrated in the following example.

Example 12.3 *Inferences from two beta distributions*
Suppose n independent draws are made from a Bernoulli distribution with unknown probability of success θ. Let x denote the number of successes and y the number of failures. The likelihood is

$$p(x|\theta, n) \propto \theta^x (1-\theta)^y. \tag{12.34}$$

The experimenter wishes to perform the Bayesian analysis under two different sets of prior assumptions. The first model assumes a uniform prior distribution for θ, $Un(0, 1)$:

$$p_1(\theta) = 1, \quad 0 \le \theta \le 1. \tag{12.35}$$

			Mean × 10		Variance × 10^2		Probability	
S	x	y	Exact	IS	Exact	IS	Exact	IS
4	4	1	6.364	6.356	1.9284	1.9757	0.1742	0.1710
100	4	1	6.364	6.361	1.9284	1.9322	0.1742	0.1728
1000	4	1	6.364	6.365	1.9284	1.9286	0.1742	0.1743
4	12	3	7.143	7.134	0.9276	0.9515	0.3155	0.3004
100	12	3	7.143	7.141	0.9276	0.9283	0.3155	0.3124
1000	12	3	7.143	7.143	0.9276	0.9277	0.3155	0.3157

TABLE 12.1. Comparison betwen exact results and estimates based on importance sampling (IS). S: number of samples in thousands; x: number of successes; y: number of failures; Probability: posterior probability that the binomial parameter takes a value between 0.75 and 0.85.

Under this prior, the posterior density is proportional to (12.34)

$$p_1(\theta|x,n) \propto \theta^x (1-\theta)^y, \tag{12.36}$$

which is recognized as the density of a beta distributed random variable with parameters $x+1, y+1$, that is $Be(\theta|x+1, y+1)$. The second model assumes the same likelihood, but the prior distribution for θ is beta, with parameters a and b. The posterior density is now

$$p_2(\theta|x,n) \propto \theta^{a+x-1} (1-\theta)^{b+y-1}, \tag{12.37}$$

which is the density $Be(\theta|a+x, b+y)$. In this example, the form of the posterior distribution is known under either prior, so it is straightforward to draw inferences from (12.36) or from (12.37). To illustrate, independent samples will be drawn from (12.36), and then importance sampling will be used to obtain inferences based on (12.37), using the draws from (12.36). Further, the Monte Carlo-based estimates will then be compared with exact results.

The results for $\theta = 0.8$, obtained with $n = x + y = 5$ or 15, are shown in Table 12.1, for three importance sampling sample sizes. The focus of inference is on the posterior mean and variance, and on the probability that the value of θ lies between 0.75 and 0.85. In the model that provides the basis of inference, $p_2(\theta|x,n)$, the parameters of the Beta prior are $a = b = 3$. The results in the table illustrate that the estimator is consistent: as the number of samples increases from 4000 to 1 million, the estimates based on (12.31), (12.32), and (12.33) converge to the true values.

When the probability to be estimated is small, a larger number of importance samples must be drawn to achieve the same level of precision. For example, the true probability that θ lies between 0.3 and 0.4, based on $p_2(\theta|x,n)$, is 15.68×10^{-4}. Estimates obtained with sample sizes of four thousand, one hundred thousand and one million were 12.10×10^{-4}, 14.66×10^{-4}, and 15.75×10^{-4}, respectively. ■

While in this example the importance sampling approach performs satisfactorily, in higher-dimensional problems the relative weights

$$\frac{w\left(\theta^{(i)}\right)}{\sum_{i=1}^{n} w\left(\theta^{(i)}\right)}$$

may be concentrated on a small number of samples. As a consequence, the Monte Carlo sampling error associated with estimates of posterior features is likely to be large. A larger effective sample size is required in order to mitigate this drawback.

Part IV

Applications in Quantitative Genetics

This page intentionally left blank

13

Gaussian and Thick-Tailed Linear Models

13.1 Introduction

The fourth part of this book illustrates applications of MCMC methods in genetic analyses, in a Bayesian context. The treatment, in parts, is rather schematic, as the objective is to present the mechanics of MCMC sampling in different modeling scenarios. Attention is restricted to models that appear quite often in quantitative genetics, e.g., linear specifications (univariate and multivariate), binary and ordered polychotomous responses, longitudinal trajectories, segregation analysis, and the study of QTL.

We start in this chapter with a class of models that is probably the most common in animal breeding applications, the Gaussian model. Here the data and other random components are assumed to follow a multivariate normal distribution and, further, location parameters and data are linearly related. The model is discussed in several settings, including situations where one (univariate) or several (multivariate) response variables are measured, and where traits may be influenced by maternal effects. Also, procedures for robust (in some sense) analysis of linear models are discussed. The reader should be aware that in several of the applications discussed below, other approaches may be computationally more efficient than the MCMC algorithms presented here. Further, an MCMC algorithm can be tailored in many different ways, and it is not claimed that the implementations discussed are, necessarily, the best ones. The final section gives a brief discussion of the impact of the alternative parameterizations of a linear model on the behavior of Gibbs sampling algorithms.

13.2 The Univariate Linear Additive Genetic Model

This model was introduced in Example 1.18 of Chapter 1. Genetic aspects of the model were described briefly in Subsection 1.4.4 of the same chapter. For analytical details, see Chapter 6.

A phenotypic record for a given trait is modeled as a linear combination of effects of some explanatory variables. It is assumed that the distribution of data $\mathbf{y}$ (vector of order n) for this trait, given some parameters $\boldsymbol{\beta}$, $\mathbf{a}$, and σ_e^2, is the multivariate normal process

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2\right). \tag{13.1}$$

Here $\boldsymbol{\beta}$ is a vector of "fixed" (in a frequentist sense) effects of order p, $\mathbf{a}$ is the vector of additive genetic values of order q, $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices associating $\boldsymbol{\beta}$ and $\mathbf{a}$ with $\mathbf{y}$, $\mathbf{I}$ is an identity matrix of order $n \times n$, and σ_e^2 is the variance of this conditional distribution, often referred to as the residual variance of the model. Genotypic values of individuals in a pedigree result from the sum of a very large number of independent contributions from many independently segregating loci, each with a small effect. The number of individuals in the pedigree (q) is often larger than the number of phenotypic records (n), which implies that $\mathbf{Z}$ has $q - n$ null columns. The genetic model justifies invoking the central limit theorem, which allows us to write

$$\mathbf{a}|\mathbf{A}, \sigma_a^2 \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right). \tag{13.2}$$

Above, $\mathbf{A}$ is the additive genetic relationship matrix (of dimension $q \times q$) and σ_a^2 is the additive genetic variance in some conceptual or "base" population. From a classical point of view, the parameters of the distribution (13.2) result from a hypothetical conceptual repeated sampling process in which vectors of additive genetic values of order $q \times 1$ are drawn at random, while maintaining the pedigree constant (i.e., with $\mathbf{A}$ fixed in every repetition of such sampling). Hence, $\mathbf{a}$ is called a "random" effect. From a Bayesian perspective, on the other hand, (13.2) represents the uncertainty distribution about genetic effects before data are observed or, in other words, it is the prior distribution of such effects. A large value of σ_a^2 implies large uncertainty. The parameters which are the focus of inference are $\boldsymbol{\beta}$, $\mathbf{a}$, σ_a^2, and σ_e^2 and, possibly, functions thereof, such as the coefficient of heritability $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. From a frequentist point of view, $\boldsymbol{\beta}$ must be defined uniquely ($\mathbf{X}$ must have full-column rank); otherwise, there is an identification problem. Hence, inferences would need to center on linearly estimable functions of $\boldsymbol{\beta}$ (Searle, 1971). Hereinafter, only the Bayesian approach is considered, where the identification problem in theory disappears whenever a proper distribution is assigned to $\boldsymbol{\beta}$ (Bernardo and Smith, 1994).

To carry out a Bayesian analysis, prior distributions must be assigned to each of $\boldsymbol{\beta}$, $\mathbf{a}$, σ_a^2 and σ_e^2. A distribution which approximates the notion of vague prior knowledge about $\boldsymbol{\beta}$ is the flat prior

$$p(\boldsymbol{\beta}) \propto \text{constant}. \tag{13.3}$$

This is an improper prior distribution, which can be made proper by assigning upper and lower limits to each of the elements of $\boldsymbol{\beta}$. In this case, the posterior distribution will then be defined within these assigned limits. In a Bayesian model with known variance components, use of priors (13.2) and (13.3) yields normal marginal posterior distributions for both $\boldsymbol{\beta}$ and $\mathbf{a}$, with mean values equal to the ML estimator of $\boldsymbol{\beta}$ and to the BLUP of $\mathbf{a}$, respectively (see Chapter 6).

Two common prior specifications for the variance components are either proper uniform distributions or scaled inverted chi-square distributions. The corresponding densities have the forms

$$p\left(\sigma_i^2\right) = \frac{1}{\sigma_{i\,\max}^2}, \quad 0 < \sigma_i^2 < \sigma_{i\,\max}^2,\ i = a, e, \tag{13.4}$$

and

$$p\left(\sigma_i^2|\nu_i, S_i^2\right) \propto \left(\sigma_i^2\right)^{-\left(\frac{\nu_i}{2}+1\right)} \exp\left(-\frac{\nu_i S_i^2}{2\sigma_i^2}\right), \quad i = a, e. \tag{13.5}$$

In (13.4), $\sigma_{i\,\max}^2$ is the maximum value that σ_i^2 is allowed to take, according to mechanistic considerations or prior knowledge about the trait. In (13.5), ν_i and S_i^2 are parameters of the corresponding scaled inverted chi-square distribution. Here it is assumed that these hyperparameters (like $\sigma_{i\,\max}^2$) are known; otherwise, these can be assigned prior distributions, in a hierarchical manner. The prior specified by (13.5) reduces to an improper uniform distribution by taking $\nu_i = -2$ and $S_i^2 = 0$.

Assuming that $\boldsymbol{\beta}$, $\left(\mathbf{a}, \sigma_a^2\right)$, and σ_e^2 are independent a priori , the joint posterior density of all unknown quantities is proportional to

$$p\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_a^2, \sigma_e^2|\mathbf{y}\right) \propto p(\boldsymbol{\beta})\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right) p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2\right). \tag{13.6}$$

In the notation, conditioning on hyperparameters is omitted. Using (13.1), (13.2), (13.3) and, for instance, (13.5), the joint posterior density (13.6) is given by

$$p\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_a^2, \sigma_e^2|\mathbf{y}\right) \propto \left(\sigma_e^2\right)^{-\left(\frac{n+\nu_e}{2}+1\right)} \left(\sigma_a^2\right)^{-\left(\frac{q+\nu_a}{2}+1\right)}$$
$$\times \exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}) + \nu_e S_e^2}{2\sigma_e^2}\right] \exp\left(-\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + \nu_a S_a^2}{2\sigma_a^2}\right). \tag{13.7}$$

The joint posterior density of a model that assumes instead improper uniform prior distributions for the variance components, is obtained by setting

$\nu_i = -2$ and $S_i^2 = 0$ in (13.7). A word of caution is in order here: when improper priors are assigned to $\left(\boldsymbol{\beta}, \sigma_a^2, \sigma_e^2\right)$, the posterior distribution may not always be proper (Hobert and Casella, 1996).

13.2.1 A Gibbs Sampling Algorithm

The single-site, systematic scan Gibbs sampling algorithm described below is based on the fully conditional posterior distributions of each scalar parameter. As usual, these are deduced from the joint posterior (13.7). However, for the location parameters $\mathbf{a}$ and $\boldsymbol{\beta}$, the derivation is simpler if one appeals to the more general results for mixed linear models given in Chapter 6; see also Example 1.18 in Chapter 1. First, note that the fully conditional posterior density of $\mathbf{a}$ and $\boldsymbol{\beta}$ is proportional to

$$p\left(\boldsymbol{\beta}, \mathbf{a} | \sigma_a^2, \sigma_e^2, \mathbf{y}\right) \propto p\left(\boldsymbol{\beta}\right) p\left(\mathbf{a} | \sigma_a^2\right) p\left(\mathbf{y} | \boldsymbol{\beta}, \mathbf{a}, \sigma_e^2\right). \tag{13.8}$$

Rather than manipulating this expression, we proceed as follows. As in Example 1.18 of Chapter 1, let

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} = \mathbf{W}\boldsymbol{\theta},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{bmatrix},$$

where $\mathbf{W} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$, $\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{a}'\right]'$, and $k = \sigma_e^2/\sigma_a^2$. Then, using results presented in Chapter 6, the conditional posterior distribution of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} | \sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}, \mathbf{C}^{-1}\sigma_e^2\right), \tag{13.9}$$

where $\mathbf{C} = \mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}$, and the posterior mean $\widehat{\boldsymbol{\theta}}$ is the solution to the linear system:

$$\mathbf{C}\widehat{\boldsymbol{\theta}} = \mathbf{W}'\mathbf{y} = \mathbf{r}. \tag{13.10}$$

One way of deriving the fully conditional posterior distribution of the ith element of $\boldsymbol{\theta}$ is as follows. Let $\boldsymbol{\theta}_{-i}$ be $\boldsymbol{\theta}$, except for the ith element $(\boldsymbol{\theta}_i)$, which is removed from the entire location vector. The results that follow hold irrespective of whether or not $\boldsymbol{\theta}_i$ is a scalar or a vector. Based on Example 1.18, one obtains

$$\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left(\widetilde{\boldsymbol{\theta}}_i, \mathbf{C}_{i,i}^{-1}\sigma_e^2\right) \tag{13.11}$$

where $\widetilde{\boldsymbol{\theta}}_i$ satisfies

$$\mathbf{C}_{i,i}\widetilde{\boldsymbol{\theta}}_i = \left(\mathbf{r}_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i}\right). \tag{13.12}$$

For example, when $\boldsymbol{\theta}_i$ is a scalar, $\mathbf{C}_{i,i}$ is the diagonal element of the coefficient matrix of the mixed model equations (13.10), $\mathbf{r}_i$ is the ith element

of the right-hand side vector in (13.10) associated with $\boldsymbol{\theta}_i$, and $\mathbf{C}_{i,-i}$ is a row vector obtained by deleting element i from the ith row of $\mathbf{C}$. Notice that drawing samples from the distribution $\left[\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i},\sigma_a^2,\sigma_e^2,\mathbf{y}\right]$, as required in Gibbs sampling, does not require inversion of matrices if $\boldsymbol{\theta}_i$ is a scalar. Expression (13.11) is quite general and applies, with appropriate minor changes, to all fully conditional densities of location parameters in Gaussian linear models.

In order to derive the fully conditional posterior distribution of the variance components, one retains only those terms in (13.7) that involve the relevant variance component. Note that, given $\boldsymbol{\theta}=\left[\boldsymbol{\beta}',\mathbf{a}'\right]'$, the two variance components are conditionally independent. The fully conditional posterior distribution of σ_a^2 is given by

$$\begin{aligned} p\left(\sigma_a^2|\boldsymbol{\beta},\mathbf{a},\sigma_e^2,\mathbf{y}\right) &\propto \left(\sigma_a^2\right)^{-\left(\frac{q+\nu_a}{2}+1\right)}\exp\left(-\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a S_a^2}{2\sigma_a^2}\right)\\ &= \left(\sigma_a^2\right)^{-\left(\frac{\widetilde{\nu}_a}{2}+1\right)}\exp\left(-\frac{\widetilde{\nu}_a\widetilde{S}_a^2}{2\sigma_a^2}\right), \end{aligned} \tag{13.13}$$

where

$$\widetilde{S}_a^2=\left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a S_a^2\right)/\widetilde{\nu}_a$$

and $\widetilde{\nu}_a=q+\nu_a$. By inspection, it follows that (13.13) is in the form of a scaled inverted chi-square density with parameters $\widetilde{\nu}_a$ and $\widetilde{S}_a^2$. In short, one can then write

$$\sigma_a^2|\boldsymbol{\beta},\mathbf{a},\sigma_e^2,\mathbf{y}\sim\widetilde{\nu}_a\widetilde{S}_a^2\chi_{\widetilde{\nu}_a}^{-2}. \tag{13.14}$$

To sample from (13.14), a draw is made from a chi-square distribution with $\widetilde{\nu}_a=q+\nu_a$ degrees of freedom, and the reciprocal of this number is multiplied by $\widetilde{\nu}_a\widetilde{S}_a^2=\left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a S_a^2\right)$. The resulting quantity is a realization from the scaled inverted chi-square process (13.14).

Similarly, collecting those terms from (13.7) that involve σ_e^2 only yields

$$\begin{aligned} p\left(\sigma_e^2|\boldsymbol{\beta},\mathbf{a},\sigma_a^2,\mathbf{y}\right) &\propto \left(\sigma_e^2\right)^{-\left(\frac{n+\nu_e}{2}+1\right)}\\ \times\exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a})+\nu_e S_e^2}{2\sigma_e^2}\right]&\\ = \left(\sigma_e^2\right)^{-\left(\frac{\widetilde{\nu}_e}{2}+1\right)}\exp\left(-\frac{\widetilde{\nu}_e\widetilde{S}_e^2}{2\sigma_e^2}\right), \end{aligned} \tag{13.15}$$

where $\widetilde{\nu}_e=n+\nu_e$ and

$$\widetilde{S}_e^2=\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a})+\nu_e S_e^2\right]/\widetilde{\nu}_e.$$

By inspection, (13.15) is proportional to the density of the following scaled inverted chi-square distribution with parameters $\widetilde{\nu}_e$ and $\widetilde{S}_e^2$

$$\sigma_e^2|\boldsymbol{\beta},\mathbf{a},\sigma_a^2,\mathbf{y}\sim\widetilde{\nu}_e\widetilde{S}_e^2\chi_{\widetilde{\nu}_e}^{-2}. \tag{13.16}$$

ID	Father	Mother	Sex	Record
1	–	–	1	–
2	–	–	2	–
3	1	–	1	y_3
4	1	2	2	y_4
5	3	4	1	y_5
6	1	4	2	y_6
7	5	6	1	–

TABLE 13.1. A hypothetical example.

The implementation of the Gibbs sampler consists of drawing successively from (13.11) for each location parameter (or block of parameters whenever $\boldsymbol{\theta}_i$ is vector valued), and from the distributions (13.14) and (13.16). The process is repeated as needed to satisfy convergence requirements, and to attain a reasonably small Monte Carlo error.

Example 13.1 *An additive genetic model*
Consider the data in Table 13.1. There are seven individuals, but only those numbered as 3, 4, 5, and 6 have a phenotypic record. Hence, the pedigree information available involves a total of $q = 7$ individuals, and $n = 4$. Note that subjects 4 to 7 have known mothers and fathers, whereas the parents of subjects 1 and 2 are unknown; for individual 3, only the father is known.

Suppose the sex of the individual is the only source of heterogeneity, other than the subjects themselves. Hence, the vector $\boldsymbol{\beta}$ contains the effects of sex $[S_1, S_2]'$ on the trait, and the vector of additive genetic values of the seven individuals is $\mathbf{a} = [a_1, a_2, \ldots, a_7]'$. The variance components will be assigned uniform distributions a priori, as in (13.4), to convey (naively) a state of vague prior knowledge about their values. Arbitrary starting values adopted for the Gibbs sampler are: $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\mathbf{a}^{(0)} = \mathbf{0}$, $\sigma_a^{2(0)} = 5$, $\sigma_e^{2(0)} = 5$, therefore, the implied starting value for the ratio of variance component is $k^{(0)} = 1$. A convenient algorithm for drawing samples from (13.11) is based on the mixed model equations. Note that $\mathbf{C}$ is a 9×9 matrix, the order resulting from $p = 2$ and $q = 7$. The first seven columns of the coefficient matrix of the mixed model equations are

$$\begin{bmatrix}
2.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 1.00 \\
0.00 & 2.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 0.00 & 2.33k & 0.50k & -0.66k & -0.50k & 0.00k \\
0.00 & 0.00 & 0.50k & 1.50k & 0.00k & -1.00k & 0.00k \\
1.00 & 0.00 & -0.66k & 0.00k & 1+2.83k & 0.50k & -1.00k \\
0.00 & 1.00 & -0.50k & -1.00k & 0.50k & 1+3k & -1.00k \\
1.00 & 0.00 & 0.00k & 0.00k & -1.00k & -1.00k & 1+2.62k \\
0.00 & 1.00 & -1.00k & 0.00k & 0.00k & -1.00k & 0.62k \\
0.00 & 0.00 & 0.00k & 0.00k & 0.00k & 0.00k & -1.23k
\end{bmatrix},$$

and the last two columns are

$$\begin{bmatrix} 0.00 & 0.00 \\ 1.00 & 0.00 \\ -1.00k & 0.00k \\ 0.00k & 0.00k \\ 0.00k & 0.00k \\ -1.00k & 0.00k \\ 0.62k & -1.23k \\ 1+2.62k & -1.23k \\ -1.23k & 2.46k \end{bmatrix}.$$

The location parameters are

$$\boldsymbol{\theta} = [S_1, S_2, a_1, \ldots, a_7]'.$$

The single-site, systematic Gibbs sampler draws the parameters at each iteration in the order in which they appear in $\boldsymbol{\theta}$ above. At iteration 1, the first draw from the fully conditional distribution, with density

$$p\left(S_1 | S_2, a_1, \ldots, a_7, \sigma_e^2, \mathbf{y}\right),$$

is obtained as:

$$S_1^{(1)} | S_2^{(0)}, a_1^{(0)}, \ldots, a_7^{(0)}, \sigma_e^{2(0)}, \sigma_a^{2(0)}, \mathbf{y} \sim N\left(\widehat{S}_1^{(1)}, \frac{\sigma_e^{2(0)}}{2}\right),$$

where, in view of the starting values adopted, $\widehat{S}_1^{(1)} = (y_3 + y_5)/2$. Next, draw $S_2^{(1)}$ from

$$S_2^{(1)} | S_1^{(1)}, a_1^{(0)}, \ldots, a_7^{(0)}, \sigma_e^{2(0)}, \sigma_a^{2(0)}, \mathbf{y} \sim N\left(\widehat{S}_2^{(1)}, \frac{\sigma_e^{2(0)}}{2}\right),$$

where $\widehat{S}_2^{(1)} = (y_4 + y_6)/2$. Subsequently, draw $a_1^{(1)}$ from

$$a_1^{(1)} | S_1^{(1)}, S_2^{(1)}, a_2^{(0)}, \ldots, a_7^{(0)}, \sigma_e^{2(0)}, \sigma_a^{2(0)}, \mathbf{y} \sim N\left(\widehat{a}_1^{(1)}, \frac{\sigma_e^{2(0)}}{2.33k^{(0)}}\right),$$

where $\widehat{a}_1^{(1)} = 0/2.33k^{(0)} = 0$. The process is continued systematically for the genetic effects of individuals 2, 3, 4, 5, 6, and 7. For example, additive genetic value 6 in iteration 1, $a_6^{(1)}$, is sampled from

$$a_6^{(1)} | S_1^{(1)}, S_2^{(1)}, a_1^{(1)}, \ldots, a_5^{(1)}, a_7^{(0)}, \sigma_e^{2(0)}, \sigma_a^{2(0)}, \mathbf{y} \sim N\left(\widehat{a}_6^{(1)}, \frac{\sigma_e^{2(0)}}{1+2.62k^{(0)}}\right),$$

where

$$\widehat{a}_6^{(1)} = \frac{y_6 - S_2^{(1)} - k^{(0)}\left(-1.00a_1^{(1)} - 1.00a_4^{(1)} + 0.62a_5^{(1)} - 1.23a_7^{(1)}\right)}{\left(1 + 2.62k^{(0)}\right)}.$$

Finally, for additive genetic value 7, sample $a_7^{(1)}$ from

$$a_7^{(1)}|S_1^{(1)}, S_2^{(1)}, a_1^{(1)}, \ldots, a_6^{(1)}, \sigma_e^{2(0)}, \sigma_a^{2(0)}, \mathbf{y} \sim N\left(\widehat{a}_7^{(1)}, \frac{\sigma_e^{2(0)}}{1 + 2.46k^{(0)}}\right),$$

where

$$\widehat{a}_7^{(1)} = \frac{\left[0 - k^{(0)}\left(-1.23a_5^{(1)} - 1.23a_6^{(1)}\right)\right]}{2.46k^{(0)}}.$$

Having sampled all location parameters, one proceeds to drawing the two variance components, to complete the first iteration of the sampler. First, the following sums of squares are computed:

$$SS_e^{(1)} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(1)} - \mathbf{Z}\mathbf{a}^{(1)}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(1)} - \mathbf{Z}\mathbf{a}^{(1)}\right),$$

and

$$SS_a^{(1)} = \mathbf{a}'^{(1)}\mathbf{A}^{-1}\mathbf{a}^{(1)}.$$

The first-round sample for the additive genetic variance, $\sigma_a^{2(1)}$, is extracted from

$$\sigma_a^2|\boldsymbol{\beta}, \mathbf{a}, \mathbf{y} \sim SS_a^{(1)}\chi_{7-2}^{-2},$$

and the residual variance $\sigma_e^{2(1)}$ from

$$\sigma_e^2|\boldsymbol{\beta}, \mathbf{a}, \mathbf{y} \sim SS_e^{(1)}\chi_{4-2}^{-2}.$$

Algorithmically, the second round of iteration starts by updating the ratio of variance components $k^{(1)} = \sigma_e^{2(1)}/\sigma_a^{2(1)}$, followed by updating the coefficient matrix of the mixed model equations. The iteration then proceeds as sketched above. ■

13.3 Additive Genetic Model with Maternal Effects

This model was proposed by Willham (1963), and has been used widely in animal breeding applications. Here, it is assumed that an offspring's attribute (or phenotypic record) is affected by the "usual" genetic and environmental effects, plus a contribution from its mother's phenotype. The latter may include both genetic and nongenetic components of maternal

origin, but it will be assumed here that the maternal effect is only genetic in nature, although it acts as an environmental influence on the offspring's record. The maternal influence is transmitted in a Mendelian manner, so both males and females carry genes for maternal effects. Naturally, genetic differences for maternal effects can be assessed only when females produce offspring.

A typical example of a trait affected by maternal effects is body weight at weaning in cattle or sheep. At weaning time, variation in body weight can be partly attributed to the calf's genes (direct genetic effects), by the pre- and post-natal environment provided by the dam of the animal (e.g., amount of milk available for suckling), and to environmental effects stemming from sources other than the mother's influence. Again, although the maternal effects are environmental vis-a-vis the measurement made in the offspring, part of the variation in milk yield between dams is assumed to be additive genetic. The calf or lamb receives a sample of 50% of its mother's and father's autosomal genes for milk yield. If the calf is female, and if it becomes a mother, these genes will influence her offspring's weaning weight.

With this model, the dispersion parameters of interest are:
(1) the additive genetic variance of "direct" effects affecting the trait e.g., calf's weaning weight;
(2) the additive genetic variance for maternal effects;
(3) a possible additive genetic covariance between direct and maternal effects; and
(4) the residual variance.

As usual, interest may also focus on the posterior distribution of location parameters or functions thereof, and on genetic parameters such as the heritabilities of direct and maternal effects, and the genetic correlation between direct and maternal effects. For applications to livestock breeding and for variations of the model, see Van Vleck (1993); an extension of Willham's specification can be found in Koerkhuis and Thompson (1997). A Bayesian analysis of a maternal effects model via the Gibbs sampler was described by Jensen et al. (1994).

The model used here to illustrate the algorithm is as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_m\mathbf{m} + \mathbf{Z}_a\mathbf{a} + \mathbf{e}. \tag{13.17}$$

The notation is as for model (13.1), with the only novelty being the introduction of $\mathbf{m}$, a vector of order q of maternal additive genetic values; $\mathbf{a}$ is now the vector of order q of direct additive genetic values, and $\mathbf{Z}_m$ and $\mathbf{Z}_a$ are known incidence matrices associating the data to $\mathbf{m}$ and to $\mathbf{a}$, respectively. The model is written in a manner that allows each individual in a pedigree to have both direct and maternal genetic effects, and this is the reason for which the order of the two vectors is $q \times 1$; this will be illustrated below

ID	Father	Mother	Sex	Record
1^*	–	–	2	–
2	–	–	1	–
3	–	–	2	–
4	2	1^*	1	y_4
5	2	3	2	y_5
6	4	5	1	y_6
7	2	5	2	y_7
8	6	7	1	–

TABLE 13.2. Hypothetical data structure from a maternal effects model.

Example 13.2 *Incidence matrices in a maternal effects model*
To illustrate the structure of matrices $\mathbf{Z}_m$ and $\mathbf{Z}_a$, we revert to the example in Table 13.1. Notice that individual 3 does not have a known mother. It simplifies matters algorithmically, if a "phantom" mother is created. After creating the "phantom" mother (1^*) of individual 3 (which now becomes individual 4, as all individuals are renumbered), the data looks as in Table 13.2.

Based on the data in Table 13.2, the structure of the matrix $\mathbf{Z}_m$, of order 4×8 and which links a data point to its mother, is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

and note that the columns are nonnull only for those individuals that are mothers and that have a measured progeny (i.e., $1^*, 3, 5$). Likewise, the structure of the matrix $\mathbf{Z}_a$, also of order 4×8, is

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Here, the nonnull columns pertain to the individuals with measurements (4 to 7). ■

Proceeding with the definition of the parameters of the model, let $\mathbf{G}_0$ be a 2×2 variance–covariance matrix whose diagonal elements are σ_m^2, the maternal additive genetic variance, and σ_a^2, the direct additive genetic variance; $\sigma_{a,m}$, the additive genetic covariance between direct and maternal effects, is in the off-diagonal position. The genetic correlation between maternal and direct additive genetic effects is, then

$$r_{a,m} = \frac{\sigma_{a,m}}{\sigma_a \sigma_m}.$$

The inverse of the matrix $\mathbf{G}_0$ is

$$\mathbf{G}_0^{-1} = \begin{bmatrix} \sigma_m^2 & \sigma_{a,m} \\ \sigma_{a,m} & \sigma_a^2 \end{bmatrix}^{-1} = \begin{bmatrix} g^{m,m} & g^{m,a} \\ g^{m,a} & g^{a,a} \end{bmatrix}.$$

The covariance structure associated with the entire vector of genetic effects $[\mathbf{m}', \mathbf{a}']'$ is

$$\mathbf{G} = Var \begin{bmatrix} \mathbf{m} \\ \mathbf{a} \end{bmatrix} = \mathbf{G}_0 \otimes \mathbf{A}, \tag{13.18}$$

where the symbol $\otimes$ stands, as usual, for "direct product", and $\mathbf{A}$ is the additive genetic relationship matrix.

According to the model postulated, the variance of $\mathbf{y}$ (in the frequentist sense, where both $\boldsymbol{\beta}$ and $\mathbf{G}_0$ are fixed parameters) is given by

$$\begin{aligned} Var(\mathbf{y}) &= Var(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_m\mathbf{m} + \mathbf{Z}_a\mathbf{a} + \mathbf{e}) \\ &= \begin{bmatrix} \mathbf{Z}_m & \mathbf{Z}_a \end{bmatrix} Var \begin{pmatrix} \mathbf{m} \\ \mathbf{a} \end{pmatrix} \begin{bmatrix} \mathbf{Z}_m' \\ \mathbf{Z}_a' \end{bmatrix} + Var(\mathbf{e}) \\ &= \begin{bmatrix} \mathbf{Z}_m & \mathbf{Z}_a \end{bmatrix} [\mathbf{G}_0 \otimes \mathbf{A}] \begin{bmatrix} \mathbf{Z}_m' \\ \mathbf{Z}_a' \end{bmatrix} + \mathbf{I}\sigma_e^2 \\ &= \mathbf{Z}_m\mathbf{A}\mathbf{Z}_m'\sigma_m^2 + \mathbf{Z}_a\mathbf{A}\mathbf{Z}_a'\sigma_a^2 + \mathbf{Z}_a\mathbf{A}\mathbf{Z}_m'\sigma_{a,m} + \mathbf{Z}_m\mathbf{A}\mathbf{Z}_a'\sigma_{a,m} + \mathbf{I}\sigma_e^2. \end{aligned}$$

The scalar version of this expression is as follows. Let y_a and y_b represent records of individuals a and b, respectively. Let the mothers of a and b be c and d, respectively. Then, if A_{ij} denotes the additive genetic relationship between i and j:

$$Cov(y_a, y_b) = A_{ab}\sigma_a^2 + A_{cd}\sigma_m^2 + (A_{ad} + A_{bc})\sigma_{a,m}$$

with the extra term σ_e^2 when $a = b$.

Return now to the Bayesian implementation via Gibbs sampling. As in (13.1) it is assumed that, conditionally on all location effects $\boldsymbol{\beta}$, $\mathbf{m}$, $\mathbf{a}$, and on the residual variance σ_e^2, the data are a realization from the normal process

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_m\mathbf{m} + \mathbf{Z}_a\mathbf{a}, \mathbf{I}\sigma_e^2\right). \tag{13.19}$$

The joint prior distribution of $\mathbf{m}$ and $\mathbf{a}$, under the assumptions of the infinitesimal model, is

$$\begin{bmatrix} \mathbf{m} \\ \mathbf{a} \end{bmatrix} \Bigg| \mathbf{A}, \mathbf{G}_0 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{G}_0 \otimes \mathbf{A}\right). \tag{13.20}$$

On defining $\mathbf{g} = [\mathbf{m}', \mathbf{a}']'$, one can write the corresponding density as

$$\begin{aligned} p(\mathbf{g}|\mathbf{A}, \mathbf{G}_0) &= |2\pi\mathbf{G}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{g}'\left(\mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right)\mathbf{g}\right] \\ &\propto |\mathbf{G}_0|^{-\frac{q}{2}} \exp\left[-\frac{1}{2}\mathbf{g}'\left(\mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right)\mathbf{g}\right]. \end{aligned} \tag{13.21}$$

This expression is now written in the form of Example 1.17 in Chapter 1, since this makes it easier to derive the fully conditional posterior distribution of $\mathbf{G}_0$. Define

$$\mathbf{S}_g = \begin{bmatrix} \mathbf{m}'\mathbf{A}^{-1}\mathbf{m} & \mathbf{m}'\mathbf{A}^{-1}\mathbf{a} \\ \mathbf{a}'\mathbf{A}^{-1}\mathbf{m} & \mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{g}' \left[\mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right] \mathbf{g} &= \begin{bmatrix} \mathbf{m}' & \mathbf{a}' \end{bmatrix} \begin{bmatrix} g^{m,m}\mathbf{A}^{-1} & g^{m,a}\mathbf{A}^{-1} \\ g^{m,a}\mathbf{A}^{-1} & g^{a,a}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{m} \\ \mathbf{a} \end{bmatrix} \\ &= tr\left(\mathbf{G}_0^{-1}\mathbf{S}_g\right). \end{aligned}$$

Therefore (13.21) can be expressed as

$$p\left(\mathbf{g}|\mathbf{A}, \mathbf{G}_0\right) \propto |\mathbf{G}_0|^{-\frac{g}{2}} \exp\left[-\frac{1}{2}\, tr\left(\mathbf{G}_0^{-1}\mathbf{S}_g\right)\right]. \tag{13.22}$$

The vector $\boldsymbol{\beta}$ is assigned again the uniform prior distribution, with density

$$p\left(\boldsymbol{\beta}\right) \propto constant.$$

The residual variance is assumed to follow, a priori, a scaled inverted chi-square distribution with density

$$p\left(\sigma_e^2|\nu_e, S_e^2\right) \propto \left(\sigma_e^2\right)^{-\left(\frac{\nu_e}{2}+1\right)} \exp\left(-\frac{\nu_e S_e^2}{2\sigma_e^2}\right), \tag{13.23}$$

where ν_e and S_e^2 are hyperparameters, assumed known. Finally, the covariance matrix $\mathbf{G}_0$ is assumed to follow, a priori, a scaled, two-dimensional, inverse Wishart distribution having density function

$$p\left(\mathbf{G}_0|\mathbf{V}, \upsilon\right) \propto |\mathbf{G}_0|^{-\frac{1}{2}(\upsilon+k+1)} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1}\mathbf{V}^{-1}\right)\right], \tag{13.24}$$

where k, the dimension of $\mathbf{G}_0$, is equal to 2 here. The density (13.24) is symbolized as $IW\left(\mathbf{V}, \upsilon\right)$, and it reduces to a two-dimensional improper uniform distribution by setting $\upsilon = -\left(k+1\right)$ and $\mathbf{V} = \mathbf{0}$.

The joint posterior density of all parameters is given by

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \mathbf{G}_0, \sigma_e^2|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \sigma_e^2\right) p\left(\boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \mathbf{G}_0, \sigma_e^2\right) \\ &\propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \sigma_e^2\right) p\left(\mathbf{m}, \mathbf{a}|\mathbf{G}_0\right) p\left(\mathbf{G}_0|\mathbf{V}, \upsilon\right) p\left(\sigma_e^2|\nu_e, S_e^2\right), \end{aligned} \tag{13.25}$$

with the second line of (13.25) arising in view of the prior assumed for $\boldsymbol{\beta}$, and because σ_e^2 and $(\mathbf{m}, \mathbf{a}, \mathbf{G}_0)$ are assumed to be independent, a priori. This joint density (distribution) is the basis for deriving all the fully conditional posterior densities (distributions) needed for implementing the Gibbs sampler.

13.3.1 Fully Conditional Posterior Distributions

As in the previous section, in order to derive the fully conditional posterior distributions of $\boldsymbol{\beta}, \mathbf{m}$, and $\mathbf{a}$, let

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_m\mathbf{m} + \mathbf{Z}_a\mathbf{a} = \mathbf{W}\boldsymbol{\theta}, \tag{13.26}$$

where

$$\mathbf{W} = \left[\begin{array}{ccc} \mathbf{X} & \mathbf{Z}_m & \mathbf{Z}_a \end{array} \right],$$

and put

$$\mathbf{C} = \mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma},$$

with

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k_{m,m} & \mathbf{A}^{-1}k_{m,a} \\ \mathbf{0} & \mathbf{A}^{-1}k_{m,a} & \mathbf{A}^{-1}k_{a,a} \end{array} \right]. \tag{13.27}$$

In (13.27), $k_{mm} = \sigma_e^2 g^{m,m}$, $k_{m,a} = \sigma_e^2 g^{m,a}$, and $k_{a,a} = \sigma_e^2 g^{a,a}$. Then, as in (13.11), the fully conditional posterior distribution of $\boldsymbol{\theta}_i$ is

$$\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \mathbf{G}_0, \sigma_e^2, \mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}_i, \mathbf{C}_{i,i}^{-1}\sigma_e^2\right) \tag{13.28}$$

where $\widehat{\boldsymbol{\theta}}_i$ satisfies

$$\mathbf{C}_{i,i}\widehat{\boldsymbol{\theta}}_i = \mathbf{r}_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i}. \tag{13.29}$$

To derive the fully conditional posterior distribution of the residual variance note, from (13.25), that the only terms that include σ_e^2 are the first and the last ones in the joint density. Therefore,

$$\begin{aligned} p\left(\sigma_e^2 | \boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \mathbf{G}_0, \mathbf{y}\right) &\propto p\left(\mathbf{y} | \boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \sigma_e^2\right) p\left(\sigma_e^2\right) \\ \propto \left(\sigma_e^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\left(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\right)'\left(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\right)}{2\sigma_e^2}\right] \left(\sigma_e^2\right)^{-\left(\frac{\nu_e}{2}+1\right)} \exp\left(-\frac{\nu_e S_e^2}{2\sigma_e^2}\right) \\ \propto \left(\sigma_e^2\right)^{-\left(\frac{\widetilde{\nu}_e}{2}+1\right)} \exp\left(-\frac{\widetilde{\nu}_e \widetilde{S}_e^2}{2\sigma_e^2}\right), \end{aligned} \tag{13.30}$$

where $\widetilde{\nu}_e = \nu_e + n$ and

$$\widetilde{S}_e^2 = \frac{\left(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\right)'\left(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\right) + \nu_e S_e^2}{\widetilde{\nu}_e}.$$

Expression (13.30) is proportional to a scaled inverted chi-square density, with $\widetilde{\nu}_e$ degrees of freedom and scale parameter $\widetilde{S}_e^2$. Thus

$$\sigma_e^2 | \boldsymbol{\beta}, \mathbf{m}, \mathbf{a}, \mathbf{G}_0, \mathbf{y} \sim \widetilde{\nu}_e \widetilde{S}_e^2 \chi_{\widetilde{\nu}_e}^{-2}. \tag{13.31}$$

Finally, the fully conditional posterior distribution of $\mathbf{G}_0$ is obtained by retaining those terms in (13.25) that involve $\mathbf{G}_0$:

$$p\left(\mathbf{G}_0|\boldsymbol{\beta},\mathbf{m},\mathbf{a},\sigma_e^2,\mathbf{y}\right) \propto p\left(\mathbf{m},\mathbf{a}|\mathbf{G}_0\right) p\left(\mathbf{G}_0|\mathbf{V},\upsilon\right)$$
$$\propto |\mathbf{G}_0|^{-\frac{q}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1}\mathbf{S}_g\right)\right] |\mathbf{G}_0|^{-\frac{1}{2}(\upsilon+k+1)} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1}\mathbf{V}^{-1}\right)\right].$$

Collecting terms yields

$$p\left(\mathbf{G}_0|\boldsymbol{\beta},\mathbf{m},\mathbf{a},\sigma_e^2,\mathbf{y}\right) \propto |\mathbf{G}_0|^{-\frac{1}{2}(\upsilon+q+k+1)} \exp\left\{-\frac{1}{2} tr\left[\mathbf{G}_0^{-1}\left(\mathbf{S}_g+\mathbf{V}^{-1}\right)\right]\right\} \tag{13.32}$$

which can be recognized as the kernel of a 2×2 scaled inverted Wishart distribution ($k = 2$), with degrees of freedom equal to $\upsilon+q$ and scale matrix $\mathbf{S}_g + \mathbf{V}^{-1}$. Hence, we write

$$\mathbf{G}_0|\boldsymbol{\beta},\mathbf{m},\mathbf{a},\sigma_e^2,\mathbf{y} \sim IW_2\left[\left(\mathbf{S}_g+\mathbf{V}^{-1}\right)^{-1},\upsilon+q\right]. \tag{13.33}$$

The Gibbs sampler proceeds by sampling from distribution (13.28), either elementwise or blockwise, and from (13.31) and (13.33).

13.4 The Multivariate Linear Additive Genetic Model

This section describes a Gibbs sampler for making Bayesian inferences based on a multiple-trait mixed linear model. This is a model that applies to a situation where several response variables are measured simultaneously on an individual. For example, Smith (1936) considered the problem of selecting among varieties of wheat differing in yield and quality traits; Hazel (1943) applied some of the ideas to pig breeding schemes, where body weights and scores had been collected in each of the animals. The "selection index" procedures suggested by these authors depend on knowledge of genetic and environmental correlations between traits, and these must be inferred using some multivariate model.

The developments presented here are circumscribed to a two-trait case, but extension to more response variables is straightforward. An arbitrary pattern of missing data will be assumed, and use is made of the ideas of data augmentation to "fill-in" the missing observations. This simplifies the Gibbs sampler, because the fully conditional posterior distributions of residual covariance matrices follow standard inverse Wishart distributions, at least for certain forms of the prior distribution. An important assumption is that the missing data are missing at random, in the sense defined by Rubin (1976). This means that the pattern of missing data may depend on the observed data, but not on the missing data. In this case, the missing data

mechanism can be ignored in the formulation of the probabilistic model, simplifying matters considerably. If, on the other hand, the data are not missing at random, it is necessary to model the missing data mechanism, which requires making extra assumptions, and additional parameters enter into the model. See Rubin (1987a), Little and Rubin (1987), Gelman et al. (1995), and Schafer (2000) for a comprehensive discussion of the issues.

Consider two traits denoted by Y and Z, and let $\mathbf{y}_o$ and $\mathbf{z}_o$ denote the vectors of observed data for traits Y and Z, respectively. The individuals are hypothetically from a polytocous (litter bearing) species, so that there may be effects common to individuals raised in the same litter. Ideally, each of the individuals would have measurements for both traits, but this is seldom the case, at least with animal breeding data, where some individuals will have records for both Y and Z, whereas others will have measurements for either Y or Z only. Suppose that the number of individuals having at least one record is n, that is, $n = n_{YZ} + n_Y + n_Z$, where n_{YZ} is the number of individuals having records for each of the two traits, and n_Y (n_Z) is the number of subjects with records only for trait Y (Z). If there were no missing records, the vector of "complete" data would have dimension $2n \times 1$. Let $\mathbf{y}_m$ and $\mathbf{z}_m$ be the vectors of missing data, and let the complete data vectors then be $\mathbf{y}' = [\mathbf{y}'_o, \mathbf{y}'_m]$ and $\mathbf{z}' = [\mathbf{z}'_o, \mathbf{z}'_m]$. The following model will be assumed

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_z \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_y \\ \boldsymbol{\beta}_z \end{bmatrix} + \begin{bmatrix} \mathbf{W}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_z \end{bmatrix} \begin{bmatrix} \mathbf{l}_y \\ \mathbf{l}_z \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_z \end{bmatrix} \begin{bmatrix} \mathbf{a}_y \\ \mathbf{a}_z \end{bmatrix} + \begin{bmatrix} \mathbf{e}_y \\ \mathbf{e}_z \end{bmatrix}, \tag{13.34}$$

where $\boldsymbol{\beta}_y$ ($\boldsymbol{\beta}_z$) is a vector of "fixed effects" affecting trait Y (Z), $\mathbf{l}_y$ ($\mathbf{l}_z$) is a vector of litter effects of order s (s), $\mathbf{a}_y$ ($\mathbf{a}_z$) is the vector of order q (q) of additive genetic values for trait Y (Z) and $\mathbf{e}_y$ ($\mathbf{e}_z$) is a vector of residual effects of order n (n) for trait Y (Z). As stated, the litter effect parameters account for the common environment affecting individuals that are raised together (often, but not always, contemporaneous full-sibs). Alternatively, they could represent, for example, "pen" or "cage" effects in laboratory animals; if such effects do not exist, they can be removed from the model. Matrices $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$, with subscripts y and z, are known incidence arrays relating location effects for each trait to the data. This is a model that could be used, for example, to analyze data on daily gain and feed intake in pigs.

Put now $\boldsymbol{\beta} = \left[\boldsymbol{\beta}'_y, \boldsymbol{\beta}'_z\right]'$, $\mathbf{l} = \left[\mathbf{l}'_y, \mathbf{l}'_z\right]'$, $\mathbf{a} = \left[\mathbf{a}'_y, \mathbf{a}'_z\right]'$, with appropriate partitions for matrices $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$, such that

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_z \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} \mathbf{W}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_z \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_y & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_z \end{bmatrix}.$$

The conditional distribution of the complete data for each individual, given the parameters, is assumed to be bivariate normal. For all individuals the

distribution can be written as

$$\mathbf{v}|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{l} + \mathbf{Z}\mathbf{a}, \mathbf{R}\right), \tag{13.35}$$

where $\mathbf{v}$ contains $\mathbf{y}$ and $\mathbf{z}$. We shall assume that records have been sorted by individual and trait within individual, so that $\mathbf{v}$ is a sequence of (Y, Z) values for each individual. Pairs of records from different individuals are assumed to be conditionally independent, given the parameters, but a correlation between residuals of the same individual is allowed. Hence, the sorting is such that the residual variance–covariance matrix can be written as $\mathbf{R} = \mathbf{I}_n \otimes \mathbf{R}_e$, a block diagonal matrix with n submatrices of residual covariances $\mathbf{R}_e$ (of order 2×2) and, as usual, $\mathbf{I}_n$ is the $n \times n$ identity matrix. The residual dispersion matrix is

$$\mathbf{R}_e = \begin{bmatrix} \sigma^2_{e,y} & \sigma_{e,yz} \\ \sigma_{e,yz} & \sigma^2_{e,z} \end{bmatrix},$$

where $\sigma^2_{e,y}$ is the residual variance for trait Y, $\sigma^2_{e,z}$ is the residual variance for trait Z, and $\sigma_{e,yz}$ is the residual covariance. The residual correlation $\sigma_{e,yz} / \left(\sigma_{e,y}\sigma_{e,z}\right)$ is a measure of the association between traits due to sources other than genetic and litter (or pen) effects.

Prior distributions are specified now, starting with the location effects. For the vector $\boldsymbol{\beta}$, a proper uniform distribution is assigned, with density

$$p\left(\boldsymbol{\beta}\right) \propto \text{constant}, \tag{13.36}$$

with possible boundaries, $\boldsymbol{\beta}_{\min}, \boldsymbol{\beta}_{\max}$, to ensure propriety of the joint posterior distribution. The vector of litter effects is assumed to have a prior uncertainty distribution well-reflected by the multivariate normal process

$$\mathbf{l}|\mathbf{R}_l \sim N\left(\mathbf{0}, \mathbf{R}_l \otimes \mathbf{I}_s\right), \tag{13.37}$$

where $\mathbf{I}_s$ is the $s \times s$ identity matrix, and $\mathbf{R}_l$ is the covariance matrix

$$\mathbf{R}_l = \begin{bmatrix} \sigma^2_{l,y} & \sigma_{l,yz} \\ \sigma_{l,yz} & \sigma^2_{l,z} \end{bmatrix}.$$

Above, $\sigma^2_{l,y}$ $\left(\sigma^2_{l,z}\right)$ is the variance between litter effects for trait Y (Z), and $\sigma_{l,yz}$ is a covariance component. The vector of additive genetic values is assumed to follow, a priori, the multivariate normal distribution

$$\mathbf{a}|\mathbf{G}_0, \mathbf{A} \sim N\left(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A}\right), \tag{13.38}$$

where $\mathbf{A}$ is the additive genetic relationship matrix of order $q \times q$, and

$$\mathbf{G}_0 = \begin{bmatrix} \sigma^2_{a,y} & \sigma_{a,yz} \\ \sigma_{a,yz} & \sigma^2_{a,z} \end{bmatrix}$$

is a 2×2 matrix, whose elements are the additive genetic (co)variance components. The genetic correlation between traits, $\sigma_{a,yz} / (\sigma_{a,y} \sigma_{a,z})$ measures the strength of the linear association between additive effects for traits Y and Z.

Two-dimensional scaled inverted Wishart distributions are assigned as prior processes for each of the $\mathbf{R}_e$, $\mathbf{R}_l$, and $\mathbf{G}_0$ covariance matrices, with the respective densities being

$$p\left(\mathbf{R}_e | \upsilon_e, \mathbf{V}_e\right) \propto \left|\mathbf{R}_e\right|^{-\frac{1}{2}(\upsilon_e + k + 1)} \exp\left[-\frac{1}{2} tr\left(\mathbf{R}_e^{-1} \mathbf{V}_e^{-1}\right)\right], \tag{13.39}$$

$$p\left(\mathbf{R}_l | \upsilon_l, \mathbf{V}_l\right) \propto \left|\mathbf{R}_l\right|^{-\frac{1}{2}(\upsilon_l + k + 1)} \exp\left[-\frac{1}{2} tr\left(\mathbf{R}_l^{-1} \mathbf{V}_l^{-1}\right)\right], \tag{13.40}$$

and

$$p\left(\mathbf{G}_0 | \upsilon_a, \mathbf{V}_a\right) \propto \left|\mathbf{G}_0\right|^{-\frac{1}{2}(\upsilon_a + k + 1)} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1} \mathbf{V}_a^{-1}\right)\right], \tag{13.41}$$

where $k = 2$ in our hypothetical bivariate model. In these expressions, υ_i and $\mathbf{V}_i$ $(i = e, l, a)$ are hyperparameters of the distributions, which are assumed known. As mentioned in connection with (13.24), these inverse Wishart distributions reduce to improper uniform distributions, if $\upsilon_i = -(k+1)$ and $\mathbf{V}_i = \mathbf{0}$.

The joint posterior density of all parameters (after augmentation with the missing records), allowing for dependence of the distribution of the litter and additive effects on the corresponding covariance matrices, but assuming prior independence otherwise, is given by

$$\begin{aligned} & p\left(\mathbf{y}_m, \mathbf{z}_m, \boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e, \mathbf{R}_l, \mathbf{G}_0 | \mathbf{y}_o, \mathbf{z}_o\right) \\ & \propto p\left(\mathbf{y}_m, \mathbf{z}_m, \boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e, \mathbf{R}_l, \mathbf{G}_0\right) p\left(\mathbf{y}_o, \mathbf{z}_o | \mathbf{y}_m, \mathbf{z}_m, \boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e\right) \\ & = p\left(\mathbf{y}_m, \mathbf{y}_o, \mathbf{z}_m, \mathbf{z}_o | \boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e\right) p\left(\mathbf{l} | \mathbf{R}_l\right) p\left(\mathbf{R}_l\right) p\left(\mathbf{a} | \mathbf{G}_0\right) p\left(\mathbf{G}_0\right) p\left(\mathbf{R}_e\right). \end{aligned} \tag{13.42}$$

This density is defined within the bounds specified in connection with the prior (13.36). Note that the first term in (13.42) is the conditional density of the complete data, defined in (13.35). The fully conditional posterior distributions are derived from (13.42) by proceeding in the usual manner, that is, fixing the appropriate conditioning variables in the joint density.

Here scaled inverse Wishart distributions were chosen as prior specifications for the covariance matrices. This makes implementation of the Gibbs sampler straightforward. Clearly, other prior specifications could be chosen, such as assigning independent prior distributions to the 3 elements of the covariance matrices, and using a parameterization in terms of correlations.

13.4.1 Fully Conditional Posterior Distributions

Imputation of Missing Records

Since the missing records are treated as unknowns in the probability model having density (13.42), their values must be imputed via Gibbs sampling by effecting draws from their fully conditional posterior distributions. First, note from the joint posterior density that

$$\begin{aligned} p\left(\mathbf{z}_m|\mathbf{y}_m,\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{R}_l,\mathbf{G}_0,\mathbf{y}_o,\mathbf{z}_o\right) &\propto p\left(\mathbf{y}_m,\mathbf{y}_o,\mathbf{z}_m,\mathbf{z}_o|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e\right) \\ &\propto p\left(\mathbf{z}_m|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{y}_o,\mathbf{z}_o\right) \\ &\propto p\left(\mathbf{z}_m|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{y}_o\right). \end{aligned} \tag{13.43}$$

The preceding follows because:
(1) Missingness is at random, so the distribution of missing records depends on the observed data, and not on the missing observations.
(2) Given $\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e$, the missing observations for trait Z do not depend on the observed records for this trait, as these have been measured on other individuals, and observations from different subjects are conditionally independent.
Furthermore, because of such conditional independence,

$$p\left(\mathbf{z}_m|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{y}_o\right) = \prod_{i\in M_Z} p\left(z_{m,i}|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,y_{o,i}\right), \tag{13.44}$$

where $z_{m,i}$ is the missing record, and $y_{o,i}$ is the observed record for subject i in the set M_Z of individuals with missing data for trait Z, comprising n_Y members. Similarly

$$p\left(\mathbf{y}_m|\mathbf{z}_m,\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{R}_l,\mathbf{G}_0,\mathbf{y}_o,\mathbf{z}_o\right) \propto \prod_{i\in M_Y} p\left(y_{m,i}|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,z_{o,i}\right), \tag{13.45}$$

where now $y_{m,i}$ is the missing record and $z_{o,i}$ is the observed record for subject i in the set M_Y of individuals with missing data for trait Y, comprising n_Z subjects.

Consider an individual i with a record for trait Y and no record for Z. Using (13.35), and the results from multivariate normal theory, the fully conditional posterior distribution of $z_{m,i}$ is

$$z_{m,i}|y_{o,i},\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e \sim N\left(\widehat{z}_{m,i},V_{z_{m,i}}\right), \tag{13.46}$$

where

$$\begin{aligned} \widehat{z}_{m,i} &= E\left(z_{m,i}|y_{o,i},\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e\right) \\ &= \mathbf{x}'_{mz,i}\boldsymbol{\beta}_z + \mathbf{w}'_{mz,i}\mathbf{l}_z + \mathbf{z}'_{mz,i}\mathbf{a}_z + \frac{\sigma_{e,yz}}{\sigma^2_{e,y}}\left(y_{o,i} - \mathbf{x}'_{oy,i}\boldsymbol{\beta}_y - \mathbf{w}'_{oy,i}\mathbf{l}_y - \mathbf{z}'_{oy,i}\mathbf{a}_y\right) \end{aligned} \tag{13.47}$$

and

$$V_{z_{m,i}} = \sigma^2_{e,z}\left(1 - \frac{\sigma^2_{e,yz}}{\sigma^2_{e,y}\sigma^2_{e,z}}\right). \tag{13.48}$$

In these expressions, $\mathbf{x}'_{mz,i}$ is the row of $\mathbf{X}_z$ associating $\boldsymbol{\beta}_z$ to $z_{m,i}$, $\mathbf{w}'_{mz,i}$ is the row of $\mathbf{W}_z$ associating $\mathbf{l}_z$ to $z_{m,i}$, and $\mathbf{z}'_{mz,i}$ is the row of $\mathbf{Z}_z$ associating $\mathbf{a}_z$ to $z_{m,i}$. If, instead, individual i has an observation for trait Z and lacks a record for Y, the fully conditional posterior distribution for $y_{m,i}$ is derived in a similar manner. Thus, the draws for the missing records for Z can be done from (13.46), or from its counterpart when Y is the missing trait.

In this computing strategy, generation of the missing data requires knowledge of elements of the incidence matrices, that is, of the way that location effects enter into the missing record. Often, for some missing records, this information may not be available. An alternative strategy, which is computationally simpler, and which avoids knowledge of incidence matrices altogether, is to generate "missing residuals" instead of missing observations, as suggested by Van Tassell and Van Vleck (1996) and Wang et al. (1997). In this setting, the joint posterior distribution is augmented with the "missing residuals" in lieu of the missing records. The missing residuals are sampled from a normal distribution with mean

$$\widehat{e}_{mz,i} = \frac{\sigma_{e,yz}}{\sigma^2_{e,y}}\left(y_o - \mathbf{x}'_{oy,i}\boldsymbol{\beta}_y - \mathbf{w}'_{oy,i}\mathbf{l}_y - \mathbf{z}'_{oy,i}\mathbf{a}_y\right).$$

The variance of the fully conditional posterior distribution of the missing residual is as in (13.48). Similar expressions apply to the "missing residuals" for trait Y.

Location Effects

The derivation of the fully conditional posterior distributions of the location parameters $\boldsymbol{\beta}$, $\mathbf{l}$ and $\mathbf{a}$ is similar to the one leading to (13.9), with a slight modification needed to accommodate the multiple-trait case. Let

$$\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{l}', \mathbf{a}'\right]',$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{X}_y & \mathbf{0}_y & \mathbf{W}_y & \mathbf{0}_y & \mathbf{Z}_y & \mathbf{0}_y \\ \mathbf{0}_z & \mathbf{X}_z & \mathbf{0}_z & \mathbf{W}_z & \mathbf{0}_z & \mathbf{Z}_z \end{bmatrix},$$

$$\mathbf{R} = \mathbf{R}_e \otimes \mathbf{I}_n,$$

noting that this implies the sorting of individuals within trait,

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_l^{-1} \otimes \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix},$$

and

$$\mathbf{C} = \mathbf{M}'\mathbf{R}^{-1}\mathbf{M} + \boldsymbol{\Omega}.$$

The multiple-trait mixed model equations are

$$\mathbf{C}\widehat{\boldsymbol{\theta}} = \mathbf{t} \tag{13.49}$$

where the right-hand side vector $\mathbf{t}$ is equal to

$$\begin{aligned} \mathbf{t} &= \mathbf{M}'\mathbf{R}^{-1}\mathbf{v} \\ &= \mathbf{M}'\mathbf{R}^{-1}\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}. \end{aligned}$$

Then the fully conditional posterior distribution of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}|\mathbf{R}_e, \mathbf{R}_l, \mathbf{G}_0, \mathbf{y}, \mathbf{z} \sim N\left(\widehat{\boldsymbol{\theta}}, \mathbf{C}^{-1}\right), \tag{13.50}$$

and the multiple-trait version of (13.11) is, for any sub-vector $\boldsymbol{\theta}_i$ of $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathbf{R}_e, \mathbf{R}_l, \mathbf{G}_0, \mathbf{y}, \mathbf{z} \sim N\left(\widetilde{\boldsymbol{\theta}}_i, \mathbf{C}_{i,i}^{-1}\right), \tag{13.51}$$

where $\widetilde{\boldsymbol{\theta}}_i$ satisfies

$$\mathbf{C}_{i,i}\widetilde{\boldsymbol{\theta}}_i = \mathbf{t}_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i}. \tag{13.52}$$

As before, when the Gibbs sampler is implemented in a scalar mode (drawing from the fully conditional posterior distributions one element of $\boldsymbol{\theta}$ at a time), $\boldsymbol{\theta}_i$ and $\mathbf{C}_{i,i}$ above are scalars and $\mathbf{C}_{i,-i}$ is a row vector. On the other hand, when the entire location vector $\boldsymbol{\theta}$ is sampled (as discussed later on), the offset $\mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i}$ is null, and (13.51) reproduces (13.50).

Dispersion Matrices

The fully conditional posterior distributions of the dispersion matrices are derived next. From (13.42):

$$p\left(\mathbf{R}_e|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_l, \mathbf{G}_0, \mathbf{y}, \mathbf{z}\right) \propto p\left(\mathbf{y}, \mathbf{z}|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e\right) p\left(\mathbf{R}_e\right). \tag{13.53}$$

Define

$$\mathbf{S}_e = \begin{bmatrix} \mathbf{e}_y'\mathbf{e}_y & \mathbf{e}_y'\mathbf{e}_z \\ \mathbf{e}_y'\mathbf{e}_z & \mathbf{e}_z'\mathbf{e}_z \end{bmatrix},$$

where

$$\mathbf{e}_y = \mathbf{y} - \mathbf{X}_y\boldsymbol{\beta}_y - \mathbf{W}_y\mathbf{l}_y - \mathbf{Z}_y\mathbf{a}_y,$$

and

$$\mathbf{e}_z = \mathbf{z} - \mathbf{X}_z\boldsymbol{\beta}_z - \mathbf{W}_z\mathbf{l}_z - \mathbf{Z}_z\mathbf{a}_z.$$

The density of the conditional distribution of the complete data, given $\boldsymbol{\beta}$, $\mathbf{l}$, $\mathbf{a}$, $\mathbf{R}_e$, can be expressed as

$$\begin{aligned} p\left(\mathbf{y}, \mathbf{z}|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e\right) &\propto |\mathbf{R}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{R}_e^{-1}\mathbf{S}_e\right)\right] \\ &= |\mathbf{R}_e|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}tr\left(\mathbf{R}_e^{-1}\mathbf{S}_e\right)\right]. \end{aligned}$$

Combining this with prior (13.39) yields

$$p\left(\mathbf{R}_e|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_l,\mathbf{G}_0,\mathbf{y},\mathbf{z}\right) \propto |\mathbf{R}_e|^{-\frac{1}{2}(\upsilon_e+n+k+1)} \times \exp\left\{-\frac{1}{2}\,tr\left[\mathbf{R}_e^{-1}\left(\mathbf{S}_e+\mathbf{V}_e^{-1}\right)\right]\right\}.$$

This is the kernel of a scaled inverted Wishart distribution of order $k=2$ (the number of traits in this case), with (υ_e+n) degrees of freedom and scale matrix $\left(\mathbf{S}_e+\mathbf{V}_e^{-1}\right)^{-1}$. Hence, the Gibbs sampler obtains updates for the residual covariance matrix from

$$\mathbf{R}_e|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_l,\mathbf{G}_0,\mathbf{y},\mathbf{z} \sim IW_2\left(\left(\mathbf{S}_e+\mathbf{V}_e^{-1}\right)^{-1},\upsilon_e+n\right). \tag{13.54}$$

Similarly for $\mathbf{R}_l$, from the joint posterior (13.42) one can deduce that

$$p\left(\mathbf{R}_l|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{G}_0,\mathbf{y},\mathbf{z}\right) \propto p\left(\mathbf{l}|\mathbf{R}_l\right)p\left(\mathbf{R}_l\right).$$

Define

$$\mathbf{S}_l = \begin{bmatrix} \mathbf{l}_y'\mathbf{l}_y & \mathbf{l}_y'\mathbf{l}_z \\ \mathbf{l}_y'\mathbf{l}_z & \mathbf{l}_z'\mathbf{l}_z \end{bmatrix}$$

and express the density $p\left(\mathbf{l}|\mathbf{R}_l\right)$ as

$$p\left(\mathbf{l}|\mathbf{R}_l\right) \propto |\mathbf{R}|^{-\frac{s}{2}} \exp\left[-\frac{1}{2}\,tr\left(\mathbf{R}_l^{-1}\mathbf{S}_l\right)\right].$$

Combining this with prior (13.40) yields

$$p\left(\mathbf{R}_l|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{G}_0,\mathbf{y},\mathbf{z}\right) \propto |\mathbf{R}_l|^{-\frac{1}{2}(\upsilon_l+s+k+1)} \times \exp\left\{-\frac{1}{2}\,tr\left[\mathbf{R}_l^{-1}\left(\mathbf{S}_l+\mathbf{V}_l^{-1}\right)\right]\right\}.$$

Therefore

$$\mathbf{R}_l|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{G}_0,\mathbf{y},\mathbf{z} \sim IW_2\left(\left(\mathbf{S}_l+\mathbf{V}_l^{-1}\right)^{-1},\upsilon_l+s\right),$$

so the samples are also obtained from an inverted Wishart distribution.

Finally, for the genetic covariance matrix, from (13.42) one obtains

$$p\left(\mathbf{G}_0|\boldsymbol{\beta},\mathbf{l},\mathbf{a},\mathbf{R}_e,\mathbf{R}_l,\mathbf{y},\mathbf{z}\right) \propto p\left(\mathbf{a}|\mathbf{G}_0\right)p\left(\mathbf{G}_0\right). \tag{13.55}$$

The term $p\left(\mathbf{a}|\mathbf{G}_0\right)$ is

$$p\left(\mathbf{a}|\mathbf{G}_0\right) \propto |\mathbf{G}_0|^{-\frac{q}{2}} \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{a}_y & \mathbf{a}_z\end{bmatrix}'\left[\mathbf{G}_0\otimes\mathbf{A}\right]^{-1}\begin{bmatrix}\mathbf{a}_y \\ \mathbf{a}_z\end{bmatrix}\right\}.$$

On defining

$$\mathbf{S}_a = \begin{bmatrix} \mathbf{a}_y'\mathbf{A}^{-1}\mathbf{a}_y & \mathbf{a}_y'\mathbf{A}^{-1}\mathbf{a}_z \\ \mathbf{a}_z'\mathbf{A}^{-1}\mathbf{a}_y & \mathbf{a}_z'\mathbf{A}^{-1}\mathbf{a}_z \end{bmatrix},$$

the density $p(\mathbf{a}|\mathbf{G}_0)$ can be expressed as

$$p(\mathbf{a}|\mathbf{G}_0) \propto |\mathbf{G}_0|^{-\frac{q}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1}\mathbf{S}_a\right)\right].$$

Therefore, combining this with prior (13.41) yields

$$p(\mathbf{G}_0|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e, \mathbf{R}_l, \mathbf{y}, \mathbf{z}) \propto |\mathbf{G}_0|^{-\frac{1}{2}(\upsilon_a+q+k+1)} \times \exp\left\{-\frac{1}{2} tr\left[\mathbf{G}_0^{-1}\left(\mathbf{V}_a^{-1}+\mathbf{S}_a\right)\right]\right\},$$

which is recognized as the kernel of the scaled inverse Wishart distribution

$$\mathbf{G}_0|\boldsymbol{\beta}, \mathbf{l}, \mathbf{a}, \mathbf{R}_e, \mathbf{R}_l, \mathbf{y}, \mathbf{z} \sim IW_2\left(\left(\mathbf{V}_a^{-1}+\mathbf{S}_a\right)^{-1}, \upsilon_a+q\right). \tag{13.56}$$

This process completes the specification of all conditional posterior distributions needed for running a Gibbs sampler. The order of visitation of the distributions is dictated primarily by computational convenience. In principle, one can follow an order similar to that described for the univariate additive genetic model in Section 13.2, keeping in mind that imputations for the missing records must be effected before draws are made for the location effects.

13.5 A Blocked Gibbs Sampler for Gaussian Linear Models

In the Gibbs sampler, as stated earlier, the location effects can be drawn either element-by-element or in blockwise manners. In the elementwise or scalar version of the sampler, each location parameter is drawn successively. A consequence of scalar sampling is that convergence may be very slow, especially in models where the parameters in the posterior distribution are highly intercorrelated (Hills and Smith, 1992; Roberts and Sahu, 1997). Also, use of data augmentation in situations where there is a large fraction of missing observations (in the broad sense of including random effects or unobserved latent variables, such as in threshold models), can also slow down convergence, for reasons similar to those hampering the expectation-maximization algorithm (Dempster et al., 1977; Liu et al., 1994).

Liu (1994), Liu et al. (1994), and Roberts and Sahu (1997) compared rates of convergence of various blocking strategies in Gibbs sampling. Liu (1994) and Liu et al. (1994) considered a three-dimensional posterior distribution with parameters a, b, c, say, and contrasted elementwise sampling with a blockwise sampling algorithm, called a group sampler. In the latter, two parameters (a, b) were blocked, so that the draws were made from $[a, b|c, data]$ and from $[c|a, b, data]$. In our context, a and b could correspond

to fixed and random effects, respectively, whereas c could correspond to dispersion parameters. A collapsed sampler was considered as well, where draws were made from $[a, b|data]$ (so integration of parameter c is required here) and from $[c|a, b, data]$. Liu (1994) found that the collapsed sampler works better than the blocked sampler, with the latter performing better than the elementwise implementation. The collapsed sampler has the potential shortcoming that integration may not always be possible in non-stylized models; for further discussion, see Chen et al. (2000). In the study of Roberts and Sahu (1997), where the joint posterior distribution was multivariate normal, it was found that when all partial correlations between parameters were positive, the blocked Gibbs sampler had a faster rate of convergence than the piecewise algorithm. This suggests that grouping positively correlated parameters can be advantageous (e.g., members of a family in a genetic context), although there are instances in which blocking can make things worse.

Hence, the effectiveness of a sampler may be enhanced by drawing parameters in blocks. For example, in the case of the linear additive genetic model discussed in Section 13.2, rather than drawing each element of the vector $\boldsymbol{\theta}$ at a time, one may draw the whole vector $\boldsymbol{\theta}$ in one pass. Clearly, if the dispersion parameters were known, this is identical to sampling directly from the target posterior distribution of the location parameter without creating a Markov chain. However, since the (co)variance parameters are typically unknown, this would be equivalent to the group sampler described above, with c being the location effects, say, and a and b playing the role of the variance components, as stated. Here a strategy presented by García-Cortés and Sorensen (1996) is described, which makes feasible sampling the entire $\boldsymbol{\theta}$ directly for some large Gaussian linear models. For simplicity, the model is restricted to a univariate (single-trait) setting with a sole random effect other than the residual. Extension to models with several variance components and to multivariate (multiple-trait) settings is relatively straightforward.

Consider the same model assumptions as in Section 13.2, and adopt proper prior distributions for $\mathbf{a}$ (13.2), for $\boldsymbol{\beta}$ ((13.3), with upper and lower bounds), and for the two variance components σ_i^2 $(i = a, e)$ as in (13.4). Then the fully conditional posterior distributions, needed for implementing the blocked or grouped Gibbs sampler, are

$$\boldsymbol{\theta}|\sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left(\widehat{\boldsymbol{\theta}}, \mathbf{C}^{-1}\sigma_e^2\right), \tag{13.57}$$

and

$$\sigma_i^2|\boldsymbol{\theta}, \mathbf{y} \sim \widetilde{\nu}_i \widetilde{S}_i \chi_{\widetilde{\nu}_i}^{-2}, \quad i = a, e, \tag{13.58}$$

where $\widetilde{\nu}_i$ and $\widetilde{S}_i$ are defined in connection with (13.14) and (13.16). Recall that the coefficient matrix of the mixed model equation is

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k \end{bmatrix},$$

where k is the ratio between the residual and the genetic components of variance. The vector of the right-hand sides is $\mathbf{W}'\mathbf{y}$, where $\mathbf{W} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$. It is computationally difficult to draw $\boldsymbol{\theta}$ from (13.57) in a single pass when $p+q$ is very large; the usual calculations for extracting a multivariate normal vector involve performing the Cholesky decomposition of the covariance matrix, etc., and these are involved and must be repeated iteration after iteration. The García-Cortés and Sorensen procedure, instead, makes use of the fact that $\widehat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{W}'\mathbf{y}$ can be calculated rapidly using iterative methods for solving linear systems of equations.

Define the following random vector of order $(p+q) \times 1$:

$$\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}} + \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{a} \end{bmatrix} - \mathbf{C}^{-1}\mathbf{W}'\mathbf{z}, \tag{13.59}$$

where:

- The vector $\widehat{\boldsymbol{\theta}}$ is the mean of the conditional posterior distribution

$$\left[\boldsymbol{\theta} | \sigma_a^2, \sigma_e^2, \mathbf{y}\right].$$

- $\boldsymbol{\mu}$ is a $p \times 1$ vector of "pseudo-fixed" effects (we will see later that the value of $\boldsymbol{\mu}$ is immaterial, so each of its elements can be set conveniently equal to 0).

- As before, $\mathbf{a}|\mathbf{A}, \sigma_a^2 \sim N\left(\mathbf{0}, \mathbf{A}\,\sigma_a^2\right)$ is a vector of random genetic effects and $\mathbf{A}$ is the usual, non-stochastic, additive relationship matrix.

- The random vector $\mathbf{z}$ is a vector of pseudo-observations generated according to the process

$$\left[\mathbf{z} | \boldsymbol{\mu}, \mathbf{a}, \sigma_e^2\right] \sim N\left(\mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2\right). \tag{13.60}$$

Since the additive genetic effects are normally distributed, and the conditional distribution (13.60) is normal, it follows that the process

$$[\mathbf{a}, \mathbf{z} | \boldsymbol{\mu}, \mathbf{A}, \sigma_a^2, \sigma_e^2]$$

is jointly normal, with parameters

$$\left.\begin{bmatrix} \mathbf{a} \\ \mathbf{z} \end{bmatrix}\right| \boldsymbol{\mu}, \sigma_a^2, \sigma_e^2 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\sigma_a^2 & \mathbf{A}\mathbf{Z}'\sigma_a^2 \\ \mathbf{Z}\mathbf{A}\sigma_a^2 & \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{I}\sigma_e^2 \end{bmatrix}\right). \tag{13.61}$$

First, write $\boldsymbol{\theta}^*$ in (13.59) as

$$\begin{aligned}\boldsymbol{\theta}^* &= \widehat{\boldsymbol{\theta}} + \mathbf{C}^{-1}\left\{\mathbf{C}\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{a}\end{bmatrix} - \mathbf{W}'\mathbf{z}\right\}\\ &= \widehat{\boldsymbol{\theta}} + \mathbf{C}^{-1}\begin{bmatrix}\mathbf{X}'\mathbf{X}\boldsymbol{\mu} + \mathbf{X}'\mathbf{Z}\mathbf{a} - \mathbf{X}'(\mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{a} + \mathbf{e})\\ \mathbf{Z}'\mathbf{X}\boldsymbol{\mu} + \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k\right)\mathbf{a} - \mathbf{Z}'(\mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{a} + \mathbf{e})\end{bmatrix}\\ &= \widehat{\boldsymbol{\theta}} + \mathbf{C}^{-1}\begin{bmatrix}-\mathbf{X}'\mathbf{e}\\ \mathbf{A}^{-1}k\mathbf{a} - \mathbf{Z}'\mathbf{e}\end{bmatrix},\end{aligned} \tag{13.62}$$

and observe that this is a linear combination of normal vectors, so its distribution (given the observed data) must be normal as well. Second, note that this vector does not depend at all on $\boldsymbol{\mu}$ so there is no loss of generality in assuming that $\boldsymbol{\mu}= \mathbf{0}$. Taking the expectation of $\boldsymbol{\theta}^*$ over $[\mathbf{a}, \mathbf{z}|\boldsymbol{\mu}, \mathbf{A}, \sigma_a^2, \sigma_e^2]$ one gets at once that

$$E(\boldsymbol{\theta}^*) = \widehat{\boldsymbol{\theta}},$$

so the mean of the distribution of $\boldsymbol{\theta}^*$ is identical to the mean of the conditional posterior distribution of $\boldsymbol{\theta}$. Third, taking variances and covariances of representation (13.62) over $\left[\mathbf{a}, \mathbf{z}|\boldsymbol{\mu}, \mathbf{A}, \sigma_a^2, \sigma_e^2\right]$ yields, since both $\widehat{\boldsymbol{\theta}}$ and $\mathbf{C}^{-1}$ are fixed,

$$\begin{aligned}Var(\boldsymbol{\theta}^*) &= \mathbf{C}^{-1}Var\begin{bmatrix}-\mathbf{X}'\mathbf{e}\\ \mathbf{A}^{-1}k\mathbf{a} - \mathbf{Z}'\mathbf{e}\end{bmatrix}\mathbf{C}^{-1}\\ &= \mathbf{C}^{-1}\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k\end{bmatrix}\mathbf{C}^{-1}\sigma_e^2 = \mathbf{C}^{-1}\sigma_e^2,\end{aligned}$$

which is identical to the variance–covariance matrix of the target conditional posterior distribution. Hence, the distribution of $\boldsymbol{\theta}^*$ is precisely that of the posterior process (13.57). The three results together imply that samples from the conditional posterior distribution of the location effects can be obtained by generating $\boldsymbol{\theta}^*$ draws using (13.59). It is convenient to rearrange $\boldsymbol{\theta}^*$ (after setting $\boldsymbol{\mu} = \mathbf{0}$, in view of considerations above) as

$$\begin{aligned}\boldsymbol{\theta}^* &= \begin{bmatrix}\mathbf{0}\\ \mathbf{a}\end{bmatrix} + \widehat{\boldsymbol{\theta}} - \mathbf{C}^{-1}\mathbf{W}'\mathbf{z}\\ &= \begin{bmatrix}\mathbf{0}\\ \mathbf{a}\end{bmatrix} + \mathbf{C}^{-1}\mathbf{W}'(\mathbf{y} - \mathbf{z})\\ &= \begin{bmatrix}\mathbf{0}\\ \mathbf{a}\end{bmatrix} + \begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}'(\mathbf{y} - \mathbf{z})\\ \mathbf{Z}'(\mathbf{y} - \mathbf{z})\end{bmatrix}.\end{aligned} \tag{13.63}$$

To summarize, the algorithm for carrying out the fully blocked implementation of the Gibbs sampler is

1. Provide starting values for σ_a^2 and σ_e^2.

2. Generate $\mathbf{a}^*$ from $N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right)$.

3. Generate $\mathbf{z}^*$ from $N\left(\mathbf{Z}\mathbf{a}^*, \mathbf{I}\sigma_e^2\right)$.

4. Calculate $\mathbf{y} - \mathbf{z}^*$.

5. Compute

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{a}^* \end{bmatrix} + \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\left(\mathbf{y} - \mathbf{z}^*\right) \\ \mathbf{Z}'\left(\mathbf{y} - \mathbf{z}^*\right) \end{bmatrix},$$

   where $\left(\boldsymbol{\beta}', \mathbf{a}'\right)'$ is a draw from $\left[\boldsymbol{\theta}|\sigma_a^2, \sigma_e^2, \mathbf{y}\right]$.

6. Compute $\widetilde{S}_i$, $(i = a, e)$.

7. Sample variance components from (13.58), and update the coefficient matrix of the mixed model equations.

8. Return to step 2 and continue with this loop, until the end of the chain.

Many different iterative algorithms that do not require inversion of $\mathbf{C}$ are available for solving the linear system

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{X}'\left(\mathbf{y} - \mathbf{z}^*\right) \\ \mathbf{Z}'\left(\mathbf{y} - \mathbf{z}^*\right) \end{bmatrix}.$$

The choice of method to apply depends on the dimension of $\mathbf{C}$ and on whether or not one wishes to pay attention to memory requirements or to computing time.

Another useful approach for sampling parameters in blocks, which is not restricted to Gaussian models, is based on the Langevin–Hastings algorithm; this was briefly described in Section 11.6 of Chapter 11. In principle, the Langevin–Hastings algorithm allows joint updates of all the parameters of the model; in the present Gaussian situation, one would update location and dispersion parameters in a single pass.

13.6 Linear Models with Thick-Tailed Distributions

13.6.1 Motivation

It is generally accepted that the normal distribution is sensitive to departures from the assumptions, because of its "thin" tails. Outlier observations can have a marked impact on inferences, so many alternative, "robust",

methods have been developed. For instance, see Hampel et al. (1986). Reviews and applications of robust procedures for parametric linear models are, for example, in Rogers and Tukey (1972), Zellner (1976), and Lange and Sinsheimer (1993). One of the possibilities that have been suggested consists of replacing the normal process with a thicker-tailed distribution, such as the Student-t, either in its univariate or multivariate forms. The use of mixed models with t distributions in quantitative genetics was pioneered by Strandén (1996) and Strandén and Gianola (1999), and some of the ideas were extended by Rosa (1998) to a wider family of distributions. In this section, we motivate the problem and, subsequently, present Gibbs sampling implementations for some variants of the theme. The problem was already encountered in Chapter 6 in the context of a linear regression model with t distributed errors, and will be dealt with again in the chapter on analysis of longitudinal trajectories.

Consider a linear regression model under the "usual" assumptions, assign a flat prior to the regression vector $\boldsymbol{\beta}$, and assume that the residual variance $\left(\sigma^2\right)$ is known. Under these conditions, the joint posterior density of the regression coefficients is

$$p\left(\boldsymbol{\beta}|\mathbf{y},\sigma^2\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right],$$

and the modal vector (equal to the mean vector in this setting) is identical to the maximum likelihood estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1}\frac{\mathbf{X}'\mathbf{y}}{\sigma^2} \\ &= \left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma^2}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i y_i}{\sigma^2}\right), \end{aligned} \tag{13.64}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. Although this point estimator does not depend on σ^2 (because the dispersion parameter cancels out), it is instructive to note that, implicitly, each observation is weighted equally by the reciprocal of the variance (or, equivalently, all observations receive a weight equal to 1).

Abandon the normality assumption for the residuals and suppose that these are independently and identically distributed as $t\left(0,\sigma^2,\nu\right)$, where σ^2 is now the scale parameter and ν is the degrees of freedom parameter, with both assumed known. Then it follows that the observations are independently distributed as

$$y_i|\boldsymbol{\beta},\sigma^2,\nu \sim t_1\left(\mathbf{x}_i'\boldsymbol{\beta},\sigma^2,\nu\right), \tag{13.65}$$

defining a univariate-t distribution. If a flat prior is adopted for the regression vector, the posterior density is then (see Chapter 1 for the form of the

t-density)

$$p\left(\boldsymbol{\beta}|\mathbf{y},\sigma^2,\nu\right) \propto \prod_{i=1}^{n}\left[1+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\sigma^2\nu}\right]^{-\frac{1+\nu}{2}}.$$

Here the modal vector is also identical to the ML estimator, because of the flat prior. To find the mode, we proceed to take derivatives of the log-posterior density (which is equal to the likelihood function, apart from an additive constant K), yielding

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}\log\left[p\left(\boldsymbol{\beta}|\mathbf{y},\sigma^2,\nu\right)\right] &= -\left(\frac{1+\nu}{2}\right)\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\log\left[1+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\sigma^2\nu}\right]+K \\ &= \left(\frac{1+\nu}{\nu}\right)\sum_{i=1}^{n}\frac{\mathbf{x}_i\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}.\end{aligned}$$

Setting this gradient to $\mathbf{0}$, it follows that the posterior mode must satisfy the system

$$\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}\boldsymbol{\beta}=\sum_{i=1}^{n}\frac{\mathbf{x}_i y_i}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}, \tag{13.66}$$

which is not explicit in $\boldsymbol{\beta}$. However, one can construct a functional iteration by assigning a starting value to the regression coefficients, and updating iterates (t denotes round number) as

$$\boldsymbol{\beta}^{[t+1]}=\left[\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}^{[t]}\right)^2}{\nu}}\right]^{-1}\left[\sum_{i=1}^{n}\frac{\mathbf{x}_i y_i}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}^{[t]}\right)^2}{\nu}}\right]. \tag{13.67}$$

This can be written in matrix form by putting

$$\mathbf{D}_{\beta}^{-1}=Diag\left[\frac{1}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}\right],$$

and noting that

$$\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}=\mathbf{X}'\mathbf{D}_{\beta}^{-1}\mathbf{X}$$

and

$$\sum_{i=1}^{n}\frac{\mathbf{x}_i y_i}{\sigma^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}=\mathbf{X}'\mathbf{D}_{\beta}^{-1}\mathbf{y}.$$

Hence (13.67) becomes

$$\boldsymbol{\beta}^{[t+1]}=\left[\mathbf{X}'\left(\mathbf{D}_{\beta}^{-1}\right)^{[t]}\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{D}_{\beta}^{-1}\right)^{[t]}\mathbf{y} \tag{13.68}$$

which defines an iteratively reweighted least-squares algorithm, showing a clear analogy between the linear regression models with normal and t-distributed residuals. Note from (13.67) that the implicit weight received by datum i is

$$\frac{1}{\sigma^2 + \frac{\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2}{\nu}}. \tag{13.69}$$

If the error distribution approaches normality ($\nu \to \infty$), the weight is $1/\sigma^2$, as in the standard regression model. If, on the other hand, the distribution has increasingly thicker tails ($\nu \to 0$), the observation is downweighted further and further. At a fixed value of the degrees of freedom parameter, the weight is inversely proportional to $\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2$, that is, the more an observation deviates from its expected value (given $\boldsymbol{\beta}$), the smaller the weight it receives in the analysis, and the less it perturbs inference. This is the reason why the t-distribution is considered as being more robust than the normal: observations that are in discrepancy with the predictive structure are attenuated, thus reducing the impact on inferences. This phenomenon, clearly, does not take place in the regression model with normally distributed errors. A comprehensive discussion of the effect of outliers is in Barnett and Lewis (1995).

An alternative to the univariate process (13.65) is to adopt a multivariate-t error distribution of order n for the residuals, that is, assume

$$t_n\left(\mathbf{0}, \mathbf{I}\sigma^2, \nu\right),$$

where $\mathbf{I}\sigma^2$ is the scale matrix. Here the residuals are uncorrelated although not independent (recall that in a multivariate-t distribution with a diagonal scale matrix, the joint density cannot be obtained by multiplying the corresponding marginal densities, all of which are univariate-t). The regression model is then based on the data generating scheme

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \nu \sim t_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2, \nu\right).$$

Note that the data constitute a sample of size 1 from a multivariate-t distribution of order n. Assuming that σ^2 and ν are both known, and that $\boldsymbol{\beta}$ has been assigned a flat prior, the posterior density takes the form

$$p\left(\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \nu\right) \propto \left[1 + \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{\sigma^2\nu}\right]^{-\frac{n+\nu}{2}}.$$

A search for the mode gives

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y},$$

as meeting the first-order condition, which is the posterior mode as if the errors had been assigned a multivariate normal distribution. The reason for

this is that all observations here receive the same implicit weight

$$\frac{1}{\sigma^2 + \frac{1}{\nu}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2},$$

so that no "robustness" in inference is gained by adopting this multivariate error distribution, other than an inflation of the posterior standard deviations of individual regression coefficients. Further, Zellner (1976) states that when $\boldsymbol{\beta}$, σ^2, and ν are all unknown, the joint posterior (using flat priors for $\boldsymbol{\beta}$, σ^2, and ν) does not have a maximum, indicating that the joint posterior is improper. On the other hand, Liu and Rubin (1995) describe modifications of the EM algorithm for the situation where the degrees of freedom are unknown (these can be estimated when there is replication of samples from the same multivariate-t process) and obtain reasonable results in their examples. At any rate, posterior modes (or ML estimates in this case) of $\boldsymbol{\beta}$ and of σ^2 exist for any fixed value of the degrees of freedom parameter, even when a single sample is drawn from the multivariate-t distribution (Zellner, 1976; McLachlan and Krishnan, 1997). As implied by Zellner (1976), the scale and degrees of freedom parameters are confounded when a single sample is drawn, even if the data vector contains n observations. If, in a Bayesian context, a proper prior is assigned to the degrees of freedom parameter, the net effect is to spread the uncertainty across all parameters of the model, but it is difficult to assess how each of the marginal distributions would be affected, since the analytical approach is intractable.

Now consider the linear model of (13.1), but under the assumptions

$$y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, \nu_e \sim t_1\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}, \sigma_e^2, \nu_e\right), \tag{13.70}$$

$$\mathbf{a}|\mathbf{A}, \sigma_a^2, \nu_a \sim t_q\left(\mathbf{0}, \mathbf{A}\,\sigma_a^2, \nu_a\right), \tag{13.71}$$

where $\mathbf{z}_i'$ is the ith row of $\mathbf{Z}$, $\mathbf{A}\,\sigma_a^2$ is the scale matrix of the q variate t distribution given above, and ν_a is the corresponding degrees of freedom parameter. This assumption preserves the property of the multivariate normal distribution usually employed for additive genetic effects: all marginal and conditional distributions are univariate- or multivariate-t, and any linear combination of breeding values is t as well. For example, if the joint distribution of the additive genetic effects of the mother, father, and of the segregation residual is trivariate-t, then the breeding value of the progeny is also t. On the other hand, if the breeding values of the father and mother are independently distributed as univariate-t and the segregation residual is also an independent univariate-t variable, the additive genetic value of the offspring is not t. Now suppose that the scales and degrees of freedom are known and, as before, that a flat prior is assigned to the location vector

$\boldsymbol{\beta}$. The logarithm of the joint posterior density is

$$p\left(\boldsymbol{\beta},\mathbf{a}|\sigma_a^2,\nu_a,\sigma_e^2,\nu_e,\mathbf{y}\right)=K-\left(\frac{1+\nu_e}{2}\right)\sum_{i=1}^{n}\log\left[1+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2}{\sigma_e^2\nu_e}\right]$$
$$-\left(\frac{q+\nu_a}{2}\right)\log\left(1+\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}{\sigma_a^2\nu_a}\right). \tag{13.72}$$

The gradients of this density are

$$\frac{\partial p\left(\boldsymbol{\beta},\mathbf{a}|\sigma_a^2,\nu_a,\sigma_e^2,\nu_e,\mathbf{y}\right)}{\partial\boldsymbol{\beta}}=\left(\frac{1+\nu_e}{\nu_e}\right)\sum_{i=1}^{n}\frac{\mathbf{x}_i\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)}{\sigma_e^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2}{\nu_e}},$$

$$\frac{\partial p\left(\boldsymbol{\beta},\mathbf{a}|\sigma_a^2,\nu_a,\sigma_e^2,\nu_e,\mathbf{y}\right)}{\partial\mathbf{a}}$$
$$=\left(\frac{1+\nu_e}{\nu_e}\right)\sum_{i=1}^{n}\frac{\mathbf{z}_i\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)}{\sigma_e^2+\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2}{\nu_e}}-\left(\frac{q+\nu_a}{\nu_a}\right)\frac{\mathbf{A}^{-1}\mathbf{a}}{\left(\sigma_a^2+\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}{\nu_a}\right)}.$$

Let

$$\sigma_{ei}^2=\frac{\sigma_e^2\nu_e+\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2}{\nu_e+1},\quad i=1,2,\ldots,n,$$

and note that this "pseudo-variance" can be viewed as a weighted average of σ_e^2 (known parameter) and of $\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2$; if $\nu_e\to\infty$, then $\sigma_{ei}^2\to\sigma_e^2$ (normality). Similarly, let

$$\overline{\sigma}_a^2=\frac{\nu_a\sigma_a^2+q\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}{q}}{\nu_a+q},$$

which is a weighted average between σ_a^2 and $[\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}]/q$, and observe that if $\nu_a\to\infty$, $\overline{\sigma}_a^2\to\sigma_a^2$. Setting all derivatives to 0 simultaneously and rearranging leads to the iterative system

$$\begin{bmatrix}\mathbf{X}'\mathbf{D}_{\beta u}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{D}_{\beta u}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{D}_{\beta u}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{D}_{\beta u}^{-1}\mathbf{Z}+\frac{\mathbf{A}^{-1}}{\overline{\sigma}_a^2}\end{bmatrix}^{[t+1]}\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{a}\end{bmatrix}^{[t]}=\begin{bmatrix}\mathbf{X}'\mathbf{D}_{\beta u}^{-1}\mathbf{y}\\ \mathbf{Z}'\mathbf{D}_{\beta u}^{-1}\mathbf{y}\end{bmatrix}^{[t]}, \tag{13.73}$$

where

$$\mathbf{D}_{\beta u}^{-1}=Diag\left(\frac{1}{\sigma_{ei}^2}\right)$$

and $\overline{\sigma}_a^2$ change iteratively. This is a set of iteratively reweighted mixed model equations, where observations that are far away from their conditional expectations $(\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{z}_i'\mathbf{a})$ are downweighted, and the more so when the degrees of freedom of the residual distribution are small. Similarly, the

pseudo-variance $\overline{\sigma}_a^2$ is modified iteratively, as the values of $\mathbf{a}$ change from round to round. In summary, this illustrates, at least when the degrees of freedom are fixed, that the univariate-t distribution for the residuals has the effect of attenuating observations that are away, in some sense, from the predictive structure of the model, whereas the multivariate-t assumption for the additive genetic effects has the effect of modifying the impact of the scale parameter σ_a^2 in the light of what the data have to say about breeding values.

A word of caution about modal estimates is in order here. At least for the regression model and the multivariate-t residual distribution, Liu and Rubin (1995) point out that the likelihood function (or posterior distribution under flat priors) can be multimodal when the degrees of freedom parameter is small or unknown. In this case, the point estimates may be of little interest by themselves, even though they may be global or local maxima. McLachlan and Krishnan (1997) gave an example where the data vector was $\mathbf{y}' = [-20, 1, 2, 3]$, and fitted a univariate-t distribution with scale parameter equal to 1 and the degrees of freedom set to 0.05. For this situation, they found that the likelihood of the unknown mean μ had four local maxima:

$$\mu_1 = -19.993,\ \mu_2 = 1.086,\ \mu_3 = 1.997,\ \mu_4 = 2.906,$$

and three local minima:

$$\mu_5 = -14.516,\ \mu_6 = 1.373,\ \mu_7 = 2.647.$$

The plot of the log-likelihood (or posterior density under a flat prior, apart from an additive constant) revealed that the likelihood fell abruptly in the neighborhood of the global maximum $\mu_3 = 1.997$, so there would be little posterior probability mass in the neighboring region. This reinforces the point that often there is no substitute for the entire posterior distribution.

Strandén and Gianola (1998), in a simulation study, evaluated the univariate t residual distribution for coping with the effects of an unknown preferential treatment of some animals in livestock breeding. Alternatively, a multivariate-t residual distribution was used, where residuals were clustered by herd; here the assumption was that the residuals were uncorrelated but not independent. They used a Bayesian model (with Gibbs sampling), where the residual distribution was univariate-t, and treated the degrees of freedom as an unknown, discrete parameter; the usual multivariate normal distribution was assigned to the breeding values. In the simulation, they compared the mean squared error of predicted breeding values (using posterior means), in situations with or without preferential treatment. In the absence of preferential treatment, the t-models were as good as the Gaussian ones. When such treatment was present, the univariate-t model was clearly the best, and the posterior distribution of the degrees of freedom pointed away from the Gaussian assumption. The authors pointed out that

it was encouraging that a symmetric error distribution improved upon the Gaussian one, even under a single-tailed form of preferential treatment. A robust asymmetric distribution, such as in Fernandez and Steel (1998), may do even better, but perhaps at the expense of conceptual and computational simplicity. Their Bayesian implementation, with some slight modifications, is discussed subsequently.

13.6.2 A Student-t Mixed Effects Model

Return to a model with the assumptions as in (13.70) and (13.71) but assume now that the degrees of freedom and the scale parameters are unknown. As seen in Chapter 1 and Chapter 6, the t distributions can be generated by mixing a normal distribution over gamma processes with appropriate parameters. The two assumptions can be replaced by the "augmented" hierarchy

$$y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, w_i \sim N\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}, \frac{\sigma_e^2}{w_i}\right), \quad i = 1, 2, ..., n, \tag{13.74}$$

$$w_i|\nu_e \sim Ga\left(\frac{\nu_e}{2}, \frac{\nu_e}{2}\right), \quad i = 1, 2, ..., n, \tag{13.75}$$

$$\mathbf{a}|\mathbf{A}, \sigma_a^2, w_a \sim N\left(\mathbf{0}, \mathbf{A}\frac{\sigma_a^2}{w_a}\right), \tag{13.76}$$

$$w_a|\nu_a \sim Ga\left(\frac{\nu_a}{2}, \frac{\nu_a}{2}\right). \tag{13.77}$$

Assume now that $\boldsymbol{\beta}$ and the two scale parameters are assigned independent, proper, uniform distributions and that the prior densities of the degrees of freedom are $p(\nu_e)$ and $p(\nu_a)$; also let the two degrees of freedom have independent prior distributions. Then the joint posterior density for the augmented hierarchy has the form

$$p\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, \mathbf{w}, \sigma_a^2, w_a, \nu_e, \nu_a|\mathbf{y}\right) \propto \prod_{i=1}^{n}\left[p\left(y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, w_i\right)p\left(w_i|\nu_e\right)\right]$$
$$\times p\left(\mathbf{a}|\mathbf{A}, \sigma_a^2, w_a\right)p\left(w_a|\nu_a\right)p\left(\nu_e\right)p\left(\nu_a\right), \tag{13.78}$$

where $\mathbf{w} = \{w_i\}$ is the vector of residual weights. The conditional posterior distributions needed for running a MCMC scheme are presented next, making use of results derived in Chapter 6, in connection with the linear regression model with residuals distributed as t. In what follows the usual notation "$ELSE$" is employed to denote the data vector $\mathbf{y}$ and all parameters that are treated as known in the appropriate conditional posterior distribution.

Residual and Genetic "Weights"

Note in (13.78) that the residual weights w_i are conditionally independent of each other, with the individual densities being

$$\begin{aligned} p\left(w_i|ELSE\right) &\propto p\left(y_i|\boldsymbol{\beta},\mathbf{a},\sigma_e^2,w_i\right)p\left(w_i|\nu_e\right) \\ &\propto \left(\frac{\sigma_e^2}{w_i}\right)^{-\frac{1}{2}} w_i^{\frac{\nu_e}{2}-1}\exp\left\{-\frac{w_i}{2}\left[\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2+\nu_e\sigma_e^2}{\sigma_e^2}\right]\right\} \\ &\propto w_i^{\frac{\nu_e+1}{2}-1}\exp\left(-\frac{w_iS_i}{2}\right), \quad i=1,2,...,n. \end{aligned} \tag{13.79}$$

where

$$S_i=\frac{\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{a}\right)^2+\nu_e\sigma_e^2}{\sigma_e^2}.$$

This indicates that the conditional posterior distribution of each w_i is the gamma distribution

$$w_i|ELSE\sim Ga\left(\frac{\nu_e+1}{2},\frac{S_i}{2}\right), \quad i=1,2,...,n. \tag{13.80}$$

Similarly,

$$\begin{aligned} p\left(w_a|ELSE\right) &\propto p\left(\mathbf{a}|\mathbf{A},\sigma_a^2,w_a\right)p\left(w_a|\nu_a\right) \\ &\propto \left(\frac{\sigma_a^2}{w_a}\right)^{-\frac{q}{2}} w_a^{\frac{\nu_a}{2}-1}\exp\left[-\frac{w_a}{2}\left(\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a\sigma_a^2}{\sigma_a^2}\right)\right] \\ &\propto w_a^{\frac{\nu_a+q}{2}-1}\exp\left[-\frac{w_a}{2}\left(\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a\sigma_a^2}{\sigma_a^2}\right)\right], \end{aligned}$$

so its conditional distribution is also gamma

$$w_a|ELSE\sim Ga\left(\frac{\nu_a+q}{2},\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a\sigma_a^2}{2\sigma_a^2}\right). \tag{13.81}$$

Location Effects

From the joint density, it follows that

$$\begin{aligned} p\left(\boldsymbol{\beta},\mathbf{a}|ELSE\right) &\propto \prod_{i=1}^{n} p\left(y_i|\boldsymbol{\beta},\mathbf{a},\sigma_e^2,w_i\right)p\left(\mathbf{a}|\mathbf{A},\sigma_a^2,w_a\right) \\ &\propto \exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)+\frac{w_a\sigma_e^2}{\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right]\right\}, \end{aligned}$$

where $\mathbf{W}=\{w_i\}$ is an $n\times n$ matrix. This form was encountered in Chapter 6 and in Section 13.2 of this chapter. The conditional posterior distribution is multivariate normal with parameters

$$\boldsymbol{\theta}|ELSE\sim N\left(\widehat{\boldsymbol{\theta}},\mathbf{C}^{-1}\sigma_e^2\right), \tag{13.82}$$

where

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z}+\mathbf{A}^{-1}\frac{w_a\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{y} \\ \mathbf{Z}'\mathbf{W}\mathbf{y} \end{bmatrix}, \\ \mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z}+\mathbf{A}^{-1}\frac{w_a\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1}.\end{aligned}$$

Techniques for drawing the location effects either in piecewise or in blockwise manners, or in a single pass, have been discussed earlier in this chapter, with the only novelty being the appearance of the w_i and w_a weights.

Scale Parameters

The conditional posterior density of the scale parameter σ_e^2, making reference to (13.78), is

$$\begin{aligned}p\left(\sigma_e^2|ELSE\right) &\propto \prod_{i=1}^{n}\left[p\left(y_i|\boldsymbol{\beta},\mathbf{a},\sigma_e^2,w_i\right)\right] \\ &\propto \left(\sigma_e^2\right)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma_e^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)\right] \\ &\propto \left(\sigma_e^2\right)^{-\frac{n-2}{2}+1}\exp\left[-\frac{1}{2\sigma_e^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)\right].\end{aligned}$$

It follows that the conditional posterior distribution is the scaled inverted chi-square process

$$\sigma_e^2|ELSE \sim \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)\chi_{n-2}^{-2}. \tag{13.83}$$

Similarly,

$$\begin{aligned}p\left(\sigma_a^2|ELSE\right) &\propto p\left(\mathbf{a}|\mathbf{A},\sigma_a^2,w_a\right) \\ &\propto \left(\sigma_a^2\right)^{-\frac{q-2}{2}+1}\exp\left(-\frac{w_a}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right),\end{aligned}$$

which indicates that

$$\sigma_a^2|ELSE \sim w_a\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\chi_{q-2}^{-2}. \tag{13.84}$$

Degrees of Freedom

The conditional posterior distribution of the degrees of freedom parameters can be deduced from (13.78). One arrives immediately at the result that ν_e and ν_a have conditionally independent posterior distributions, with densities

$$p\left(\nu_e|ELSE\right) \propto \left[\prod_{i=1}^{n}p\left(w_i|\nu_e\right)\right]p\left(\nu_e\right),$$

and

$$p\left(\nu_a|ELSE\right) \propto p\left(w_a|\nu_a\right) p\left(\nu_a\right).$$

It is not possible to go further without being specific about the form of the prior distributions. Here, we will consider two alternative settings.

In the first one, the degrees of freedom (positive parameters) are allowed to take integer values only over a finite set of values having equal prior probability (Albert and Chib, 1993; Geweke, 1993; Strandén, 1996). Let these sets contain values $f_j = 1, 2, \ldots, d_e$ and $q_k = 1, 2, \ldots, d_a$, respectively. The prior distributions are then

$$p\left(\nu_e\right) = \frac{1}{d_e}, \quad \nu_e \in \left\{f_j = 1, 2, ..., d_e\right\},$$

and

$$p\left(\nu_e\right) = \frac{1}{d_a}, \quad \nu_a \in \left\{q_k = 1, 2, ..., d_a\right\}.$$

Recalling that the prior distributions of the weights are in Gamma form, it follows that

$$p\left(\nu_e|ELSE\right) \propto \left[\frac{\left(\frac{\nu_e}{2}\right)^{\left(\frac{\nu_e}{2}\right)}}{\Gamma\left(\frac{\nu_e}{2}\right)}\right]^n \prod_{i=1}^{n}\left[w_i^{\frac{\nu_e}{2}-1} \exp\left(-\frac{\nu_e w_i}{2}\right)\right].$$

This is a discrete distribution, and samples can be drawn by extracting degrees of freedom values with probabilities

$$\Pr\left(\nu_e = f_j|ELSE\right) = C_e\left[\frac{\left(\frac{f_j}{2}\right)^{\left(\frac{f_j}{2}\right)}}{\Gamma\left(\frac{f_j}{2}\right)}\right]^n \prod_{i=1}^{n}\left[w_i^{\frac{f_j}{2}-1} \exp\left(-\frac{f_j w_i}{2}\right)\right], \tag{13.85}$$

where C_e is equal to

$$\left[\sum_{j=1}^{d_e}\left\{\left[\frac{\left(\frac{f_j}{2}\right)^{\left(\frac{f_j}{2}\right)}}{\Gamma\left(\frac{f_j}{2}\right)}\right]^n \prod_{i=1}^{n}\left[w_i^{\frac{f_j}{2}-1} \exp\left(-\frac{f_j w_i}{2}\right)\right]\right\}\right]^{-1}.$$

Likewise, and following a similar type of algebra, the conditional posterior distribution of the degrees of freedom of the multivariate-t distribution of the additive genetic effects can be found to be:

$$\Pr\left(\nu_a = q_k|ELSE\right) = C_a \frac{\left(\frac{q_k}{2}\right)^{\left(\frac{q_k}{2}\right)}}{\Gamma\left(\frac{q_k}{2}\right)} w_a^{\frac{q_k}{2}-1} \exp\left(-\frac{q_k w_a}{2}\right) \tag{13.86}$$

where

$$C_a = \left\{ \sum_{k=1}^{d_a} \left[\frac{\left(\frac{q_k}{2}\right)^{\left(\frac{q_k}{2}\right)}}{\Gamma\left(\frac{q_k}{2}\right)} w_a^{\frac{q_k}{2}-1} \exp\left(-\frac{q_k w_a}{2}\right) \right] \right\}^{-1}.$$

Hence, new states for the degrees of freedom of the genetic distribution can be drawn by sampling with probabilities (13.86). The Gibbs sampling scheme is completed by effecting draws from distributions (13.80)–(13.86) in any suitable order.

If the degrees of freedom are treated as continuous, the conditional posterior densities are not in any recognizable form. Here, one may consider embedding a Metropolis–Hastings step in the MCMC scheme, as in Geweke (1993) and Rodriguez-Zas (1998). For example, Geweke (1993) uses an exponential distribution for these parameters, so that the conditional posterior density of the residual degrees of freedom has the form

$$\begin{aligned} p\left(\nu_e | ELSE\right) &\propto \left[\frac{\left(\frac{\nu_e}{2}\right)^{\left(\frac{\nu_e}{2}\right)}}{\Gamma\left(\frac{\nu_e}{2}\right)} \right]^n \prod_{i=1}^{n} \left[w_i^{\frac{\nu_e}{2}-1} \exp\left(-\frac{\nu_e w_i}{2}\right) \right] \exp\left(-\omega_e \nu_e\right) \\ &\propto \left[\frac{\left(\frac{\nu_e}{2}\right)^{\left(\frac{\nu_e}{2}\right)}}{\Gamma\left(\frac{\nu_e}{2}\right)} \right]^n \exp\left[-\frac{\nu_e \left(2\omega_e + n\overline{w}\right)}{2} \right] \prod_{i=1}^{n} w_i^{\frac{\nu_e}{2}-1}, \end{aligned} \tag{13.87}$$

where ω_e is the parameter of the prior exponential distribution and $\overline{w}$ is the average of the weights w_i at any given iterate. For example, if $\omega_e = 0.15$, this corresponds to a prior mean and variance of 6.6 and 44.4, respectively, for the residual degrees of freedom. The exponential prior can be assessed such that both small and large values of the degrees of freedom receive "high" prior probability, making the specification vague enough. Rodriguez-Zas (1998) used a normal proposal distribution for Metropolis–Hastings with parameters based on the current values of the weights in the course of iteration. Similarly, the conditional posterior density of the degrees of freedom of the genetic distribution would be

$$\begin{aligned} p\left(\nu_a | ELSE\right) &\propto \frac{\left(\frac{\nu_a}{2}\right)^{\left(\frac{\nu_a}{2}\right)}}{\Gamma\left(\frac{\nu_a}{2}\right)} w_a^{\frac{\nu_a}{2}-1} \exp\left(-\frac{\nu_a w_a}{2}\right) \exp\left(-\omega_a \nu_a\right) \\ &\propto \frac{\left(\frac{\nu_a}{2}\right)^{\left(\frac{\nu_a}{2}\right)}}{\Gamma\left(\frac{\nu_a}{2}\right)} \exp\left[-\frac{\nu_a \left(2\omega_a + w_a\right)}{2} \right] w_a^{\frac{\nu_a}{2}-1}, \end{aligned} \tag{13.88}$$

where ω_a is the parameter of the prior exponential distribution.

Making an analogy with the single sample multivariate-t model of Zellner (1976), we conjecture that there is no information in the likelihood available to separate ν_a from σ_a^2. As noted, using proper priors for both parameters solves the identifiability problem and spreads the uncertainty throughout the model, which is realistic. On the other hand, Strandén and Gianola

(1999) suggest fitting a series of models with alternative values for ν_a and then assessing the impact on inferences. The different models can be contrasted using some of the Bayesian model comparison tools discussed earlier in the book. On the other hand, if the t-distribution has replicate samples, such as in a sire model where the transmitting abilities are distributed independently, the two parameters are not confounded. Hence, in practice, one can cluster individuals into families, assume these are independent, and find the most probable value of the degrees of freedom. Then, conditionally on such modal value, one proceeds with the implementation given above, ignoring the uncertainty about its error.

13.6.3 Model with Clustered Random Effects

The setting here is one where observations are clustered in some natural manner, and where cluster effects are independent and identically distributed as univariate-t, with unknown scale and degrees of freedom parameters. For example, individuals may be clustered in nuclear or half-sib families, or in randomly created inbred lines; alternatively, observations could consist of repeated measures on individuals, in which case the subjects constitute the clustering criterion.

The assumptions to be made here are

$$\begin{aligned} y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, w_i &\sim N\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}, \frac{\sigma_e^2}{w_i}\right), \quad i = 1, 2, ..., n, \\ w_i|\nu_e &\sim Ga\left(\frac{\nu_e}{2}, \frac{\nu_e}{2}\right), \quad i = 1, 2, ..., n, \end{aligned}$$

for the observations, thus defining a univariate-t distribution upon integration of the joint distribution over the weights, and

$$\begin{aligned} a_i|\sigma_a^2, w_{ai} &\sim N\left(0, \frac{\sigma_a^2}{w_{ai}}\right), \quad i = 1, 2, ..., q, \\ w_{ai}|\nu_a &\sim Ga\left(\frac{\nu_a}{2}, \frac{\nu_a}{2}\right), \quad i = 1, 2, ..., q. \end{aligned}$$

This is equivalent to stating that the cluster effects are independent and identically distributed as univariate-t, a priori, with null mean, scale parameter σ_a^2, and degrees of freedom ν_a. Here there is "replication" of the second-stage distribution, so both σ_a^2 and ν_a are estimable in the maximum likelihood sense.

The joint posterior density of all unknowns is

$$\begin{aligned} p\left(\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, \mathbf{w}, \sigma_a^2, \mathbf{w}_a, \nu_e, \nu_a|\mathbf{y}\right) \propto \prod_{i=1}^{n} \left[p\left(y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, w_i\right) p\left(w_i|\nu_e\right)\right] \\ \times \prod_{i=1}^{q} \left[p\left(a_i|\sigma_a^2, w_{ai}\right) p\left(w_{ai}|\nu_a\right)\right] p\left(\nu_e\right) p\left(\nu_a\right), \end{aligned}$$

where $\mathbf{w}_a = \{w_{ai}\}$ is a $q \times 1$ vector. The conditional posterior distribution of the residual weights and of σ_e^2 remain as in (13.79) and (13.83), respectively. Further, the conditional posterior distribution of the cluster weights and variance takes the form

$$w_{ai}|ELSE \sim Ga\left(\frac{\nu_a+1}{2}, \frac{a_i^2+\nu_a\sigma_a^2}{2\sigma_a^2}\right), \quad i=1,2,\ldots,q. \tag{13.89}$$

The scale parameter σ_a^2 has as conditional posterior distribution

$$\sigma_a^2|ELSE \sim \left(\sum_{i=1}^{q} w_{ai}a_i^2\right)\chi_{q-2}^{-2}, \tag{13.90}$$

noting that the random effects enter attenuated by the weights w_{ai}, relative to their counterpart in a purely Gaussian model. For example, the scale parameter of this scaled inverted chi-square distribution would be $\sum_{i=1}^{q} a_i^2$ in the latter.

The conditional posterior density of the location effects is

$$p\left(\boldsymbol{\beta}, \mathbf{a}|ELSE\right) \propto \prod_{i=1}^{n} p\left(y_i|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2, w_i\right) \prod_{i=1}^{q} p\left(a_i|\sigma_a^2, w_{ai}\right)$$

$$\propto \exp\left[-\frac{1}{2\sigma_e^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)'\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{a}\right)\right]\exp\left(-\frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{W}_a\mathbf{a}\right),$$

where $\mathbf{W}_a$ is a $q \times q$ diagonal matrix with typical element equal to w_{ai} $(i=1,2,...,q)$. Using well established results, manipulation of the above density leads directly to the normal distribution

$$\boldsymbol{\theta}_w|ELSE \sim N\left(\widehat{\boldsymbol{\theta}}_w, \mathbf{C}_w^{-1}\sigma_e^2\right), \tag{13.91}$$

where

$$\widehat{\boldsymbol{\theta}}_w = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z}+\mathbf{W}_a\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{y} \\ \mathbf{Z}'\mathbf{W}\mathbf{y} \end{bmatrix},$$

and

$$\mathbf{C}_w^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z}+\mathbf{W}_a\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1}.$$

Finally, the conditional posterior densities of the degrees of freedom (continuous case with exponential prior distributions) are as in (13.87) for the residual distribution and

$$p\left(\nu_a|ELSE\right) \propto \left[\frac{\left(\frac{\nu_a}{2}\right)^{\left(\frac{\nu_a}{2}\right)}}{\Gamma\left(\frac{\nu_a}{2}\right)}\right]^{q} \exp\left[-\frac{\nu_a\left(2\omega_a+q\overline{w}_a\right)}{2}\right]\prod_{i=1}^{q} w_{ai}^{\frac{\nu_a}{2}-1}. \tag{13.92}$$

This completes the specification of an MCMC scheme for a linear model where all residuals and cluster effects are distributed as univariate-t with appropriate parameters.

13.7 Parameterizations and the Gibbs Sampler

Roberts and Sahu (1997) noted that high correlations among the elements of the parameter vector in the posterior distribution can lead to poor convergence of the single-site Gibbs sampler. An illustration of the consequences of a large posterior correlation of parameters was discussed in Chapter 12 in Example 12.2. The effect of parameterization on the behavior of the Markov chain is also discussed in Gelfand et al. (1995) and in Gelfand et al. (1996). These authors argued that "hierarchically centered" parameterizations for linear and nonlinear models reduce the extent of posterior intercorrelations and lead to faster mixing and convergence. Here we will give a brief discussion of the problem, by adapting the presentation in Gelfand et al. (1996) to a quantitative genetics setting.

Consider the additive genetic model

$$y_i = \mu + a_i + e_i,$$

where there is a single observation made in each animal, $i = 1, 2, ..., n$; assume that animals are genetically unrelated. As usual, take

$$y_i|\mu, a_i \sim N\left(\mu + a_i, \sigma_e^2\right)$$

as the data generating process and

$$a_i|\sigma_a^2 \sim N\left(0, \sigma_a^2\right)$$

as a prior for the genetic effects. Suppose the two variance components are known, and assign the process $\mu \sim N\left(\mu_0, \sigma_\mu^2\right)$ as a prior distribution for μ, where hyperparameters are taken as known as well. Using standard results employed several times in this book, the posterior variance–covariance matrix of μ and of the additive genetic effects is

$$Var\left(\left.\begin{bmatrix}\mu\\ a_1\\ a_2\\ .\\ .\\ a_n\end{bmatrix}\right| \mu_0, \sigma_\mu^2, \sigma_a^2, \sigma_e^2, \mathbf{y}\right)$$

$$= \begin{bmatrix} n+\frac{\sigma_e^2}{\sigma_\mu^2} & 1 & 1 & . & 1 & 1 \\ 1 & 1+\frac{\sigma_e^2}{\sigma_a^2} & 0 & . & 0 & 0 \\ 1 & 0 & 1+\frac{\sigma_e^2}{\sigma_a^2} & . & 0 & 0 \\ . & . & . & . & . & . \\ 1 & 0 & . & . & 1+\frac{\sigma_e^2}{\sigma_a^2} & 0 \\ 1 & 0 & . & . & 0 & 1+\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \sigma_e^2.$$

The inverse of this matrix can be found using results for partitioned matrices (e.g., Searle, 1982). Gelfand et al. (1996) arrive at the following representations (the conditioning on the known parameters or hyperparameters is suppressed in the notation)

$$\begin{aligned} Corr\left(\mu, a_i|\mathbf{y}\right) &= -\left(1+n\frac{\sigma_e^2}{\sigma_a^2}+\frac{\sigma_e^2}{\sigma_\mu^2}+\frac{\sigma_e^4}{\sigma_\mu^2}\right)^{-\frac{1}{2}}, \\ Corr\left(a_i, a_j|\mathbf{y}\right) &= \left(1+n\frac{\sigma_e^2}{\sigma_a^2}+\frac{\sigma_e^2}{\sigma_\mu^2}+\frac{\sigma_e^4}{\sigma_\mu^2}\right)^{-1}. \end{aligned}$$

As σ_e^2 tends to infinity, the correlations go to 0, but tend to -1 and 1, respectively if either σ_a^2 or σ_μ^2 become larger and larger (vague prior knowledge about the random effects). From an animal breeding point of view, suppose that a flat prior is assigned to μ ($\sigma_\mu^2 \to \infty$) and recall that $\sigma_e^2/\sigma_a^2 = (1-h^2)/h^2$, where h^2 is heritability. Then, when $h^2 \to 0$, so that the variability is dominated by environmental effects,

$$Corr\left(\mu, a_i|\mathbf{y}\right) = -\left(1+n\frac{1-h^2}{h^2}\right)^{-\frac{1}{2}} \to 0$$

and

$$Corr\left(a_i, a_j|\mathbf{y}\right) = \left(1+n\frac{1-h^2}{h^2}\right)^{-1} \to 0.$$

On the other hand, when $h^2 \to 1$, a situation in which one definitely needs the random effects model, the absolute value of the correlations tend to 1. This implies that the Gibbs sampler will mix slower for highly heritable traits, at least with the standard parameterization given above.

Gelfand et al. (1996) also discuss "hierarchical centering". Here, they define $\eta_i = \mu + a_i$, so that the resulting hierarchical model is $y_i|\eta_i \sim N\left(\eta_i, \sigma_e^2\right)$, with $\eta_i|\mu, \sigma_a^2 \sim N\left(\mu, \sigma_a^2\right)$, and $\mu \sim N\left(\mu_0, \sigma_\mu^2\right)$. These two models are probabilistically equivalent, that is, give the same prior predictive distribution or marginal distribution of the observations. The authors encounter

$$\begin{aligned} Corr\left(\mu, \eta_i|\mathbf{y}\right) &= -\left(1+n\frac{\sigma_a^2}{\sigma_e^2}+\frac{\sigma_a^2}{\sigma_\mu^2}+\frac{\sigma_a^4}{\sigma_\mu^2}\right)^{-\frac{1}{2}}, \\ Corr\left(\eta_i, \eta_j|\mathbf{y}\right) &= \left(1+n\frac{\sigma_a^2}{\sigma_e^2}+\frac{\sigma_a^2}{\sigma_\mu^2}+\frac{\sigma_a^4}{\sigma_\mu^2}\right)^{-1}. \end{aligned}$$

Here the correlations do not go to 0 if $\sigma_e^2 \to \infty$. Returning again to the animal breeding setting, with a uniform improper prior for μ, one gets

$$\begin{aligned} Corr\,(\mu, \eta_i|\mathbf{y}) &= -\left(1 + n\frac{h^2}{1-h^2}\right)^{-\frac{1}{2}}, \\ Corr\,(\eta_i, \eta_j|\mathbf{y}) &= \left(1 + n\frac{h^2}{1-h^2}\right)^{-1}. \end{aligned}$$

When heritability tends to 0, the correlations tend to 1, but when the genetic variance becomes relatively more and more important, the correlations go to 0, indicating that the hierarchical parameterization should be preferred. In the standard parameterization, the data inform directly on $\eta_i = \mu + a_i$, so if there is vague prior knowledge about a_i (large value of the additive genetic variance) the data cannot separate μ from a_i; thus, the large negative correlation between these unknowns. Gelfand et al. (1996) note that, in practice, the variance components are unknown, so it is necessary to consider the joint posterior distribution of location effects and dispersion parameters; however, they recommend hierarchical centering as a default procedure. A detailed study of the problem, including longitudinal data settings, is in Gelfand et al. (1995). They concluded that hierarchical centering will usually improve convergence of sampling-based procedures for Bayesian analysis.

Other types of parameterizations and strategies to improve mixing and convergence were studied by Hills and Smith (1992, 1993); Liu (1994) and by Liu et al. (1994).

14

Threshold Models for Categorical Responses

14.1 Introduction

Discrete response variables are ubiquitous in genetics. In particular, categorical responses arise when the outcome is an assignment into one of several mutually exclusive and exhaustive classes. Typical examples include congenital malformations, leg weakness traits in pigs, presence or absence of intra-mammary infection in cows, subjective scores describing difficulties at birth in cattle, X-ray readings of hip-dysplasia in dogs, and litter-size in sheep. When there are two categories of response, the trait is referred to as binary or "all or none" (Dempster and Lerner, 1950). With more than two categories, a distinction must be made as to whether the classes are either unordered or ordered in some manner. Unordered categories can arise when the outcome is a choice; for example, electing an item in a menu or voting for a certain candidate. In this chapter, however, the focus will be on the ordered categories. This is so, because in biological systems, response categories can be ordered almost invariably along some hypothetical gradient. For example, it is possible to think about a fecundity gradient in sheep, from least prolific to most prolific. Here, the litter size observed at birth would be related somehow to this conceptual gradient.

Quantitative geneticists have used the so called threshold model to relate a hypothetical continuous scale to an outward phenotype (the observed category of response). The underlying variate is often called "the latent variable" or, at least in genetics of disease, "liability", after Falconer (1965). The model postulates that the continuous response is rendered discrete via

some fixed thresholds or boundaries delimiting categories (Wright, 1934; Robertson and Lerner, 1949; Dempster and Lerner, 1950; Falconer, 1965, 1967). For example, if there are two categories of response, the observation would be in the second class, e.g., "disease", if the liability exceeds the threshold. The origins of the threshold model can be traced back to Pearson (1900), Wright (1934), Bliss (1935), and Dempster and Lerner (1950). Recent methodological contributions, made primarily from a statistical genetic perspective, include Gianola (1982), Harville and Mee (1984), Gianola and Foulley (1983), Foulley et al. (1987) and Foulley and Manfredi (1991), among others. Curnow (1972) and Curnow and Smith (1975) suggested an alternative concept where, instead of having abrupt thresholds, there is a "risk function". However, at least as these authors formulate the model, their development is mathematically equivalent to one with abrupt thresholds.

The threshold model for analysis of binary traits was encountered in Chapter 4. There, it was pointed out that whenever the model requires specification of random effects, the likelihood function (or marginal posterior distribution) does not have a closed form. In this case, standard likelihood-based analyses have been conducted using Gaussian quadrature approximations in models with independent random effects, as in Anderson and Aitkin (1985) or, when the random effects are correlated, with the Monte Carlo EM algorithm, as in McCulloch (1994). Analyses based on approximations have been developed by Gilmour et al. (1985) and Lee and Nelder (1996) have suggested what are called "hierarchical likelihoods". Basically, these methods can be viewed as approximations to REML and BLUP in the context of generalized linear mixed models. Many of the arguments on which these methods rest are of an asymptotic nature, and the finite sample properties of the procedures are unknown.

In this chapter, an MCMC Bayesian implementation of two models involving ordered categorical traits is described. In all cases, it is assumed that the underlying or latent variable has a (conditional) Gaussian distribution. The object of inference may include, for example, the additive genetic values of individuals (to rank candidates for selection in genetic improvement programs), the probability distribution by category of response for some individuals or experimental conditions or interest, or the genetic variance of the trait in question. The first model discussed extends the analysis of a binary trait presented in Chapter 4 to one with an arbitrary number of ordered response categories. Following Albert and Chib (1993), it is shown that a Gibbs sampler, used with data augmentation, leads to fully posterior distributions that are easy to sample from. The development is based on Sorensen et al. (1995), Heringstad et al. (2001) and Kadarmideen et al. (2002). The second model described can be used when the outcome is an ordered categorical response and an observation on a Gaussian trait, following ideas in Jensen (1994), Sorensen (1996) and in Wang et al. (1997). See Foulley et al. (1983) for an approximate Bayesian analysis of this model.

14.2 Analysis of a Single Polychotomous Trait

14.2.1 Sampling Model

Let all underlying latent variables or liabilities be represented by the vector $\mathbf{l} = \{l_i\}$ $(i = 1, 2, \ldots, n)$, such that for the ith individual or data point it is postulated that

$$l_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a} + e_i. \tag{14.1}$$

Here $\boldsymbol{\beta}$ are some location effects, $\mathbf{a}$ is a $q \times 1$ vector of additive genetic values (perhaps of order larger than n), and $e_i \sim N\left(0, \sigma_e^2\right)$ is a random residual. As usual, $\mathbf{x}_i'$ and $\mathbf{z}_i'$ are incidence row vectors. It will be assumed that, given the location parameters $\boldsymbol{\beta}$ and $\mathbf{a}$, the elements of the vector $\mathbf{l}$ are conditionally independent and distributed as

$$\left(\mathbf{l}|\boldsymbol{\beta}, \mathbf{a}, \sigma_e^2\right) \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2\right). \tag{14.2}$$

Since the liability variate is unobservable, the parameterization $\sigma_e^2 = 1$ will be adopted here (i.e., Cox and Snell, 1989), in order to achieve identifiability in the likelihood. It must be noted that when inferences are based on posterior distributions with proper priors assigned to all parameters, this is not technically required (Bernardo and Smith, 1994). However, this setting is used here nonetheless since it is standard in threshold model analysis. Other parameterizations of the threshold model are discussed in Sorensen et al. (1995).

Let $\mathbf{y} = \{y_i\}$ $(i = 1, 2, \ldots, n)$, denote the vector of observed categorical data. Here, each y_i represents an assignment into one of c mutually exclusive and exhaustive categories of response arrived at more or less arbitrarily. Often, the assignment is subjective, as in analysis of conformation scores. These classes result from the hypothetical existence of $c + 1$ thresholds in the latent scale, such that $t_{\min} < t_1 < t_2 < \cdots < t_{c-1} < t_{\max}$. For example, if the realized value of liability is between t_1 and t_2 , the assignment is in the second category of response. Set the two extreme thresholds to $t_0 = t_{\min}, t_c = t_{\max}$ so that the remaining $c-1$ thresholds can take any value between $t_{\min}$ and $t_{\max}$, subject to the preceding order requirement. However, one of the thresholds must be fixed, so as to center the distribution; a typical assignment is $t_1 = 0$. Then the conditional probability that y_i falls in category j $(j = 1, 2, \ldots, c)$, given $\boldsymbol{\beta}$, $\mathbf{a}$, and $\mathbf{t} = (t_{\min}, t_1, \ldots, t_{c-1}, t_{\max})'$ is given by

$$\begin{aligned}\Pr\left(y_i = j|\boldsymbol{\beta}, \mathbf{a}, \mathbf{t}\right) &= \Pr\left(t_{j-1} < l_i < t_j|\boldsymbol{\beta}, \mathbf{a}, \mathbf{t}\right) \\ &= \Phi\left(t_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right) - \Phi\left(t_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right) \\ &= p\left(y_i|\boldsymbol{\beta}, \mathbf{a}, \mathbf{t}\right).\end{aligned} \tag{14.3}$$

The data are conditionally independent, given $\boldsymbol{\beta}$, $\mathbf{a}$, and $\mathbf{t}$. Therefore the sampling model can be written as

$$p(\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\mathbf{t}) = \prod_{i=1}^{n}\sum_{j=1}^{c} I(y_i=j)\, p(y_i|\boldsymbol{\beta},\mathbf{a},\mathbf{t})$$

$$= \prod_{i=1}^{n}\sum_{j=1}^{c} I(y_i=j)\left[\Phi\left(t_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right) - \Phi\left(t_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right)\right], \quad (14.4)$$

where $I(y_i = j)$ is an indicator function taking the value 1 if the response falls in category j and 0 otherwise.

14.2.2 Prior Distribution and Joint Posterior Density

Adopting the usual hierarchical model building strategy, prior distributions must be assigned to $\boldsymbol{\beta}$, $\mathbf{a}$, and $\mathbf{t}$. It will be assumed, as usual, that the prior distribution of $\mathbf{a}$ depends on an unknown dispersion parameter σ_a^2. The density of the joint prior distribution adopted has the form

$$p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{t},\sigma_a^2\right) = p(\boldsymbol{\beta})\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p(\mathbf{t}).$$

Hence, the joint posterior density is:

$$p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{t},\sigma_a^2|\mathbf{y}\right) \propto p(\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\mathbf{t})\, p(\boldsymbol{\beta})\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p(\mathbf{t})$$

$$= p(\boldsymbol{\beta})\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p(\mathbf{t})$$

$$\times \prod_{i=1}^{n}\sum_{j=1}^{c} I(y_i=j)\left[\Phi\left(t_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right) - \Phi\left(t_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right)\right].$$

The fully conditional posterior distributions of parameters $\boldsymbol{\beta}$, $\mathbf{a}$, $\mathbf{t}$, and σ_a^2 must be derived from the above expression. Irrespective of the form of the joint prior distribution, these conditional processes are not in standard form, because the parameters appear implicitly inside of the normal integrals. Therefore special strategies must be used for implementing a Gibbs sampler; see, for example, Moreno et al. (1997).

An algorithmically simpler approach consists of augmenting the joint posterior distribution with the unobserved liabilities $\mathbf{l}$. If the latent variables are modelled hierarchically as in a Gaussian linear model for observed data, this approach yields fully conditional posterior distributions that have a standard form and which are easy to sample from. Augmenting the joint posterior with $\mathbf{l}$, the resulting density takes the form

$$p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{l},\mathbf{t},\sigma_a^2|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\mathbf{l},\mathbf{t},\sigma_a^2\right) p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{l},\mathbf{t},\sigma_a^2\right)$$

$$= p(\mathbf{y}|\mathbf{l},\mathbf{t})\, p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{l},\mathbf{t},\sigma_a^2\right) = p(\mathbf{y}|\mathbf{l},\mathbf{t})\, p(\mathbf{l}|\boldsymbol{\beta},\mathbf{a})\, p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{t},\sigma_a^2\right)$$

$$= p(\mathbf{y}|\mathbf{l},\mathbf{t})\left[\prod_{i=1}^{n} p(l_i|\boldsymbol{\beta},\mathbf{a})\right] p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{t},\sigma_a^2\right). \quad (14.5)$$

The second and third lines of the expression above follow because: 1) the distribution of the polychotomous observations, given the liabilities, depends only on the thresholds, and 2) given $\boldsymbol{\beta}$ and $\mathbf{a}$, the liabilities are conditionally independent, as indicated in (14.2). All conditional posterior distributions can be derived from (14.5), and this will be discussed later on.

Consider the ith element in the first term on the right hand side of (14.5)

$$\Pr\left(y_i = j | l_i, t_{j-1}, t_j\right) = \begin{cases} 1, & \text{if } t_{j-1} < l_i \leq t_j, \\ 0, & \text{otherwise.} \end{cases} \tag{14.6}$$

That is, the probability that a given data point falls in a given category, given the value of liability and thresholds, is completely specified. This means that $p\left(\mathbf{y}|\mathbf{l}, \mathbf{t}\right)$ is a degenerate distribution. Following the notation in Albert and Chib (1993), $p\left(\mathbf{y}|\mathbf{l}, \mathbf{t}\right)$ can be written as

$$p\left(\mathbf{y}|\mathbf{l}, \mathbf{t}\right) = \prod_{i=1}^{n} \left[\sum_{i=1}^{c} I\left(t_{j-1} < l_i \leq t_j\right) I\left(y_i = j\right) \right]. \tag{14.7}$$

The prior distribution of the vector $\boldsymbol{\beta}$ must be specified judiciously. It is well-established that data structures with a small number of observations (or in extreme cases, when all observations fall into a particular category) per element or level of $\boldsymbol{\beta}$, can lead to poor inferences. Misztal et al. (1989) have referred to this as the extreme category problem (ECP). The problem is related to the fact that when all observations for a location parameter are in one of the extreme categories (e.g., in the binary situation, when all observations are either 0 or 1), the ML estimate of the corresponding parameter is not finite. This is clear in a probit model with a single location parameter: if all observations are $0's$, the ML estimate of the probability of response is 0. Hence, the corresponding ML estimate of the location parameter in the liability scale is $-\infty$. In the context of a hierarchical structure, this problem propagates to other tiers of the model. For example, when there are ECP instances for at least some elements of $\boldsymbol{\beta}$ and when a uniform prior distribution is assigned to this vector, Bayesian MCMC inferences about σ_a^2 can be severely distorted (Hoeschele and Tier, 1995; Moreno et al., 1997). It is not obvious how this problem can be solved satisfactorily, although some ad hoc approaches have been suggested in the literature.

Following Kadarmideen et al. (2002), partition the vector $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \left(\boldsymbol{\beta}_h', \boldsymbol{\beta}_r'\right)',$$

with the corresponding partition for the incidence matrix being

$$\mathbf{X} = \left[\mathbf{X}_h, \mathbf{X}_r\right].$$

Here $\boldsymbol{\beta}_h$ $(H \times 1)$ contains elements of $\boldsymbol{\beta}$ known to have observations with ECPs (such as small herds of cattle in which mastitis is screened), and $\boldsymbol{\beta}_r$ contains elements of $\boldsymbol{\beta}$ with well-structured data. Then, the distorting effects of the ECP on inferences can be tempered somewhat by assuming that the vector $\boldsymbol{\beta}_h$ follows, a priori, a normal distribution with a nonnull mean. A simple possibility is to pose

$$\boldsymbol{\beta}_h|\beta, \sigma^2_{\beta_h} \sim N\left(\mathbf{1}\beta, \mathbf{I}_h\sigma^2_{\beta_h}\right). \tag{14.8}$$

Here $\mathbf{1}$ is a vector of ones, β is a scalar common to all elements of $\boldsymbol{\beta}_h$, and $\sigma^2_{\beta_h}$ is an unknown dispersion parameter. For $\boldsymbol{\beta}_r$ one can assume a vague normal distribution with zero mean and large, known, variance. For instance, assuming that all scalar elements of $\boldsymbol{\beta}_r$ are mutually independent, one can adopt

$$\boldsymbol{\beta}_r \sim N\left(\mathbf{0}, \mathbf{I}_r 10^6\right). \tag{14.9}$$

Assuming prior independence between $\boldsymbol{\beta}_h$ and $\boldsymbol{\beta}_r$, the prior distribution of $\boldsymbol{\beta}$ has density

$$p\left(\boldsymbol{\beta}\right) = p\left(\boldsymbol{\beta}_h|\beta, \sigma^2_{\beta_h}\right) p\left(\boldsymbol{\beta}_r\right). \tag{14.10}$$

The scalar β can be assumed to follow the uniform distribution

$$\beta|\beta_{\min}, \beta_{\max} \sim Un\left(\beta_{\min}, \beta_{\max}\right) \tag{14.11}$$

where $\beta_{\min}$ and $\beta_{\max}$ are chosen appropriately. The prior distribution for $\sigma^2_{\beta_h}$ can be a scaled inverted chi-square process with known parameters ν_β and S_β,with density

$$p\left(\sigma^2_{\beta_h}|\nu_{\beta_h}, S_{\beta_h}\right) \propto \left(\sigma^2_{\beta_h}\right)^{-\left(\frac{\nu_{\beta_h}}{2}+1\right)} \exp\left(-\frac{\nu_{\beta_h} S_{\beta_h}}{2\sigma^2_{\beta_h}}\right). \tag{14.12}$$

The prior for the thresholds arises naturally from the assumptions of the model. This model postulates that the thresholds are ordered, so it is sensible to assume that these are distributed as order statistics from a uniform distribution, in the interval $[t_{\min}, t_{\max}]$. Also, recall that $t_1 = 0$, to give an origin to the underlying distribution; hence, there are $c-2$ unknown thresholds. Therefore, the joint prior density of $\mathbf{t}$ is(Mood et al., 1974):

$$p\left(\mathbf{t}\right) = (c-2)!\left(\frac{1}{t_{\min} - t_{\max}}\right)^{c-2} I\left(\mathbf{t} \in \mathbf{T}\right), \tag{14.13}$$

where $\mathbf{T} = \{(t_1 = 0, t_2, \ldots, t_{c-1})\,|t_{\min} < t_1 < t_2 < \cdots < t_{c-1} < t_{\max}\}$.

If, as stated earlier, the location vector $\mathbf{a}$ consists of additive genetic effects (meaning that genetic variation is due to additive and independent contributions from a large number of loci with small gene substitution effects), a sensible prior is

$$\mathbf{a}|\mathbf{A}, \sigma^2_a \sim N\left(\mathbf{0}, \mathbf{A}\sigma^2_a\right), \tag{14.14}$$

where σ_a^2 is the additive genetic variance and $\mathbf{A}$ is the usual additive relationship matrix. In turn, the additive genetic variance can be conveniently assumed to be distributed, a priori, as a scaled inverted chi-square random variable with density

$$p\left(\sigma_a^2|\nu_a, S_a\right) \propto \left(\sigma_a^2\right)^{-\left(\frac{\nu_a}{2}+1\right)} \exp\left(-\frac{\nu_a S_a}{2\sigma_a^2}\right), \tag{14.15}$$

and ν_a, S_a are known hyperparameters.

The final assumption is that the joint prior density of all unknown parameters, including the liabilities, can be factorized as

$$\begin{aligned} &p\left(\mathbf{l}, \boldsymbol{\beta}, \beta, \mathbf{a}, \mathbf{t}, \sigma_a^2, \sigma_{\beta_h}^2\right) \\ &= p\left(\mathbf{l}|\boldsymbol{\beta}, \mathbf{a}\right) p\left(\boldsymbol{\beta}_h|\beta, \sigma_{\beta_h}^2\right) p\left(\boldsymbol{\beta}_r\right) p\left(\mathbf{a}|\sigma_a^2\right) p\left(\mathbf{t}\right) p\left(\sigma_a^2\right) p\left(\sigma_{\beta_h}^2\right), \end{aligned}$$

where the dependency on the hyperparameters is suppressed in the notation. In view of this and of (14.5), the joint posterior density can be written as

$$\begin{aligned} &p\left(\mathbf{l}, \boldsymbol{\beta}, \beta, \mathbf{a}, \mathbf{t}, \sigma_a^2, \sigma_{\beta_h}^2|\mathbf{y}\right) \propto p\left(\mathbf{y}|\mathbf{l}, \mathbf{t}\right) p\left(\mathbf{t}\right) \\ &\times \left[\prod_{i=1}^{n} p\left(l_i|\boldsymbol{\beta}, \mathbf{a}\right)\right] p\left(\boldsymbol{\beta}_h|\beta, \sigma_{\beta_h}^2\right) p\left(\boldsymbol{\beta}_r\right) p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p\left(\sigma_{\beta_h}^2\right). \end{aligned} \tag{14.16}$$

14.2.3 Fully Conditional Posterior Distributions

Liabilities

Notationally, the fully conditional posterior distribution of parameter x will be represented as $p\left(x|ELSE\right)$, where $ELSE$ refers to data $\mathbf{y}$ and to the values of all the parameters that x depends on. Consider first the fully conditional posterior distribution of liability l_i. In order to obtain this, one must extract the terms involving l_i from the joint posterior (14.16). Since the liabilities are conditionally independent, it follows that, given the parameters and the liabilities, the polychotomous observations are independent as well. This yields

$$p\left(l_i|ELSE\right) \propto p\left(y_i = j|l_i, \mathbf{t}\right) p\left(l_i|\boldsymbol{\beta}, \mathbf{a}\right). \tag{14.17}$$

Given the liabilities and the thresholds, the categorical outcome is not stochastic, since its value is known with certainty. Hence, $p\left(y_i = j|l_i, \mathbf{t}\right)$ is a constant and gets absorbed in the Bayes formula. Therefore,

$$p\left(l_i|ELSE\right) \propto \left[\sum_{i=1}^{c} I\left(t_{j-1} < l_i \leq t_j\right) I\left(y_i = j\right)|\boldsymbol{\beta}, \mathbf{a}\right],$$

where $I\left(t_{j-1} < l_i \leq t_j\right) I\left(y_i = j\right)$ indicates that liability falls in the interval $t_{j-1} < l_i \leq t_j$ if $y_i = j$. Since liability is Gaussian, it follows that this is the density of a truncated normal distribution, with density

$$p\left(l_i|ELSE\right) = \frac{\phi\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{a}, 1\right)}{\Phi\left(t_j - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right) - \Phi\left(t_{j-1} - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{a}\right)}. \tag{14.18}$$

Thresholds

The density of the fully conditional posterior distribution of the ith threshold t_i, is

$$\begin{aligned} p\left(t_i|ELSE\right) &\propto p\left(\mathbf{y}|\mathbf{l},\mathbf{t}\right) p\left(\mathbf{t}\right) \\ &\propto \prod_{j=1}^{N} \left[I\left(t_{i-1} < l_j < t_i\right) I\left(y_j = i\right) + I\left(t_i < l_j < t_{i+1}\right) I\left(y_j = i+1\right)\right]. \end{aligned} \tag{14.19}$$

The preceding is the collection of all terms in the joint posterior density where t_i appears. For example, consider threshold t_2. This will appear either in connection with liabilities corresponding to responses in either the second category (where the threshold is an upper bound) or in the third class (where the threshold is a lower bound). Seen as a function of t_i, (14.19) shows formally that t_i lies in an interval whose limits are as follows: the upper limit must be smaller than or equal to the smallest value of l for which $y_j = i + 1$. The lower limit is given by the maximum value of l for which $y_j = i$. The prior condition ($\mathbf{t} \in \mathbf{T}$) is fulfilled automatically. Within these boundaries, the conditional posterior distribution of threshold t_i is the uniform process

$$p\left(t_i|ELSE\right) = \frac{1}{\min\left(\mathbf{l}|\mathbf{y} = i+1\right) - \max\left(\mathbf{l}|\mathbf{y} = i\right)}, \tag{14.20}$$

where $\min\left(\mathbf{l}|\mathbf{y} = i+1\right)$ indicates the minimum value of the liabilities within observations in category i+1; similarly, $\max\left(\mathbf{l}|\mathbf{y} = i\right)$ denotes the maximum value of liabilities for observations in category i.

A comment about implementation is in order here. The interval

$$\min\left(\mathbf{l}|\mathbf{y} = i+1\right) - \max\left(\mathbf{l}|\mathbf{y} = i\right)$$

is typically very narrow, and varies little between successive iterates of the Gibbs sampler. This leads to strong autocorrelations between samples and slows convergence. Cowles (1996) and Nandram and Chen (1996) propose alternative algorithms for ameliorating this difficulty.

Additive Genetic Variance

The density of the fully conditional posterior distribution of σ_a^2 is

$$p\left(\sigma_a^2|ELSE\right) \propto p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right).$$

This is identical to the expression for the density of the conditional posterior distribution of σ_a^2 in a single-trait additive genetic model presented in Section 13.2. Therefore,

$$\sigma_a^2|ELSE\sim \left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu_a S_a\right)\chi_{\nu_a+q}^{-2}. \tag{14.21}$$

Dispersion Parameter $\sigma_{\beta_h}^2$

Recall that $\sigma_{\beta_h}^2$ is a dispersion parameter describing variability between levels of effects containing (possibly) ECPs. The density of the fully conditional posterior distribution of $\sigma_{\beta_h}^2$ is

$$p\left(\sigma_{\beta_h}^2|ELSE\right)\propto p\left(\boldsymbol{\beta}_h|\beta,\sigma_{\beta_h}^2\right)p\left(\sigma_{\beta_h}^2\right).$$

This is clearly in the same form as the density of the conditional distribution of the additive genetic variance, the difference being that the prior mean of each of the elements of $\boldsymbol{\beta}_h$ is β, instead of 0. After making an offset for the non-null mean, some algebra leads to the scaled inverted chi-square distribution

$$\sigma_{\beta_h}^2|ELSE\sim\left[(\boldsymbol{\beta}_h-\mathbf{1}\beta)'(\boldsymbol{\beta}_h-\mathbf{1}\beta)+\nu_\beta S_\beta\right]\chi_{\nu_{\beta_h}+H}^{-2}, \tag{14.22}$$

where H is the number of elements in $\boldsymbol{\beta}_h$.

Location Effects

We consider first the conditional distribution of $[\boldsymbol{\beta},\mathbf{a}]$ and, subsequently, that of the scalar parameter β. As usual, the fully conditional posterior distribution of $[\boldsymbol{\beta},\mathbf{a}]$ is obtained from (14.16). Extracting the terms containing $\boldsymbol{\beta}$, and $\mathbf{a}$ one obtains

$$\begin{aligned}p\left(\boldsymbol{\beta},\mathbf{a}|\sigma_{\beta_h}^2,\sigma_a^2,\beta,\mathbf{l},\mathbf{y}\right)&\propto\left[\prod_{i=1}^{n}p\left(l_i|\boldsymbol{\beta},\mathbf{a}\right)\right]p\left(\boldsymbol{\beta}_h|\beta,\sigma_{\beta_h}^2\right)p\left(\boldsymbol{\beta}_r\right)p\left(\mathbf{a}|\sigma_a^2\right)\\&\propto p\left(\boldsymbol{\beta},\mathbf{a}|\sigma_{\beta_h}^2,\sigma_a^2,\beta,\mathbf{l}\right).\end{aligned}$$

This follows because, given the liabilities, the categorical responses $\mathbf{y}$ do not bring any additional information about the location effects. The preceding expression has a form similar to (13.8), except that $\mathbf{l}$ here, replaces $\mathbf{y}$. This is precisely the density of the joint posterior distribution of the location effects in a Gaussian hierarchical model, which was discussed extensively in Chapter 6. Using the result in Example 1.18 of Chapter 1, the fully conditional posterior distributions follow rather directly. First

$$\left[\boldsymbol{\beta}_h|\boldsymbol{\beta}_r,\mathbf{a},\sigma_a^2,\sigma_{\beta_h}^2,\beta,\mathbf{l},\mathbf{y}\right]\sim N\left(\widehat{\boldsymbol{\beta}}_h,\left[\mathbf{X}_h'\mathbf{X}_h+\frac{1}{\sigma_{\beta_h}^2}\mathbf{I}_h\right]^{-1}\right), \tag{14.23}$$

where

$$\widehat{\boldsymbol{\beta}}_h = \left[\mathbf{X}_h'\mathbf{X}_h + \frac{1}{\sigma^2_{\beta_h}}\mathbf{I}_h\right]^{-1}\left[\mathbf{X}_h'\left(\mathbf{l} - \mathbf{X}_r\boldsymbol{\beta}_r - \mathbf{Z}\mathbf{a}\right) + \mathbf{1}\frac{\beta}{\sigma^2_{\beta_h}}.\right]$$

Further,

$$\left[\boldsymbol{\beta}_r|\boldsymbol{\beta}_h, \mathbf{a}, \sigma^2_a, \sigma^2_{\beta_h}, \beta, \mathbf{l}, \mathbf{y}\right] \sim N\left(\widehat{\boldsymbol{\beta}}_r, \left[\mathbf{X}_r'\mathbf{X}_r + 10^{-6}\mathbf{I}_r\right]^{-1}\right), \tag{14.24}$$

where

$$\widehat{\boldsymbol{\beta}}_r = \left[\mathbf{X}_r'\mathbf{X}_r + \frac{1}{10^6}\mathbf{I}_r\right]^{-1}\mathbf{X}_r'\left(\mathbf{l} - \mathbf{X}_h\boldsymbol{\beta}_h - \mathbf{Z}\mathbf{a}\right).$$

Likewise, for the additive genetic effects,

$$\left[\mathbf{a}|\boldsymbol{\beta}_h, \boldsymbol{\beta}_r, \sigma^2_a, \sigma^2_{\beta_h}, \beta, \mathbf{l}, \mathbf{y}\right] \sim N\left(\widehat{\mathbf{a}}, \left[\mathbf{Z}'\mathbf{Z} + \frac{1}{\sigma^2_a}\mathbf{A}^{-1}\right]^{-1}\right) \tag{14.25}$$

where

$$\widehat{\mathbf{a}} = \left[\mathbf{Z}'\mathbf{Z} + \frac{1}{\sigma^2_a}\mathbf{A}^{-1}\right]^{-1}\mathbf{Z}'\left(\mathbf{l} - \mathbf{X}_r\boldsymbol{\beta}_r - \mathbf{X}_h\boldsymbol{\beta}_h\right).$$

Using these expressions, a single site updating Gibbs sampler can be developed as for the Gaussian hierarchical model. Alternatively (and perhaps more efficiently from a computational point of view), a joint updating algorithm can be chosen along the lines described in Section 13.5 of the previous chapter.

Finally, the density of the fully conditional posterior distribution of the scalar β is:

$$\begin{aligned} p\left(\beta|\boldsymbol{\beta}, \mathbf{a}, \sigma^2_{\beta_h}, \sigma^2_a, \mathbf{t}, \mathbf{l}, \mathbf{y}\right) &\propto p\left(\boldsymbol{\beta}_h|\beta, \sigma^2_{\beta_h}\right) \\ &\propto \exp\left[-\frac{1}{2\sigma^2_{\beta_h}}\left(\boldsymbol{\beta}_h - \mathbf{1}\beta\right)'\left(\boldsymbol{\beta}_h - \mathbf{1}\beta\right)\right]. \end{aligned}$$

Viewed as a function of β, this is the density of the normal distribution

$$\left[\beta|\boldsymbol{\beta}, \mathbf{a}, \sigma^2_{\beta_h}, \sigma^2_a, \mathbf{t}, \mathbf{l}, \mathbf{y}\right] \sim N\left(\overline{\beta}_h, \frac{\sigma^2_{\beta_h}}{H}\right), \tag{14.26}$$

where

$$\overline{\beta}_h = \frac{1}{H}\sum_{i=1}^{H}\beta_{h_i}$$

and β_{h_i} is the ith element of $\boldsymbol{\beta}_h$, drawn from (14.23).

14.2.4 The Gibbs Sampler

To summarize, the Gibbs sampler consists of iterating through the following loop:

1. Read through the data file and sample the liabilities $\mathbf{l}$ from the truncated normal distribution with density (14.18).

2. Sample the thresholds from the uniform distribution (14.20).

3. Sample the additive genetic variance from the scaled inverted chi-square process (14.21).

4. Sample $\sigma^2_{\beta_h}$ from the scaled inverted chi-square distribution (14.22).

5. Build mixed model equations and their right-hand sides using $\mathbf{l}$ as "data".

6. Sample the location parameters from the normal distributions (14.23), (14.24) and 14.25).

7. Sample β from the normal distribution (14.26)

8. Return to Step 1 or terminate when chain length is adequate to meet convergence diagnostics.

14.3 Joint Analysis of an Ordered Categorical and a Normally Distributed Trait

The results presented in the previous section are extended for a joint analysis of a model for one categorical and one normally distributed trait. Such a model could be relevant to study the genetic associations between growth rate and leg weakness in pigs, for example, since the last trait is scored categorically. The setting is as in Foulley et al. (1983), who introduced the model and suggested an approximate Bayesian analysis. The approach presented here can accommodate a general pattern of missing data. For the purpose of presentation, a simple additive genetic model is postulated for each of the two traits. Bayesian MCMC related work can be found in Jensen (1994), Sorensen (1996), and in Wang et al. (1997). A more general model that encompasses several categorical and Gaussian traits was described by Van Tassell et al. (1998). Korsgaard et al. (2002) proposed a model for the joint analysis of categorical, censored and Gaussian traits using the Gibbs sampler.

14.3.1 Sampling Model

Suppose there are n individuals, each of which is potentially measured for each of the two traits. However, it is typical that there will be at least some individuals on which the measurement is available for one of the traits only. Subscript 1 will refer to the continuous trait, and subscript 2 to the categorical trait. Denote by $\mathbf{y}_{1o}$ and by $\mathbf{y}_{2o}$ the vectors of the observed data for the continuous and categorical trait, respectively, and by $\mathbf{y}_{1m}$ and by $\mathbf{y}_{2m}$, the vectors of the missing data for the continuous and categorical trait, respectively. Let $\mathbf{y}_1' = (\mathbf{y}_{1o}', \mathbf{y}_{1m}')$ and $\mathbf{y}_2' = (\mathbf{y}_{2o}', \mathbf{y}_{2m}')$ be vectors of dimension n each. As in Section 13.4, it is assumed that data are missing at random. Also, as before, let $\mathbf{l}$ (of order $n \times 1$), represent the unobserved liabilities associated with the categorical trait, which can be partitioned in an obvious notation as $\mathbf{l}' = (\mathbf{l}_o', \mathbf{l}_m')$.

The approach followed here is to augment the posterior distribution with $(\mathbf{y}_{1m}', \mathbf{l}')$, that is, with the missing data for the continuous trait and all liabilities. It is assumed that the vector of complete continuous data, which is defined here as $(\mathbf{y}_1', \mathbf{l}') = (\mathbf{y}_{1o}', \mathbf{y}_{1m}', \mathbf{l}')$, is normally distributed, given vectors of location parameters $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$ and $\mathbf{a} = (\mathbf{a}_1', \mathbf{a}_2')'$, with the latter being the additive genetic values for the two traits. The order of $\mathbf{a}_1$ and $\mathbf{a}_2$ is $q \times 1$ each. If the complete continuous data are sorted by trait and by individual within trait, with the resulting vector labeled as $\mathbf{v}$, the conditional distribution of the complete continuous data given the location parameters has the form

$$\mathbf{v}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_e \sim N\left(\left[\begin{array}{c} \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_1\mathbf{a}_1 \\ \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_2\mathbf{a}_2 \end{array}\right], \mathbf{R}\right), \tag{14.27}$$

where the $\mathbf{X}'s$ and $\mathbf{Z}'s$ are incidence matrices of appropriate order. In (14.27), $\mathbf{R} = \mathbf{R}_e \otimes \mathbf{I}_n$, $\mathbf{R}_e$ is the 2×2 variance–covariance matrix

$$\mathbf{R}_e = \left[\begin{array}{cc} \sigma_{e1}^2 & \sigma_{e1,2} \\ \sigma_{e1,2} & \sigma_{e2}^2 \end{array}\right], \tag{14.28}$$

and $\mathbf{I}_n$ is an $n \times n$ identity matrix. As in Section 13.4, the data can be augmented with residuals as in Wang et al. (1997), in which case the appropriate rows of incidence matrices $\mathbf{X}$ and $\mathbf{Z}$ in (14.27) have all elements equal to zero.

Following the results in Example 1.17 of Chapter 1, the density associated with (14.27) can be written as

$$p\left(\mathbf{v}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_e\right) \propto |\mathbf{R}_e|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{R}_e^{-1}\mathbf{S}_e\right)\right]. \tag{14.29}$$

Here

$$\mathbf{S}_e = \left[\begin{array}{cc} \mathbf{e}_1'\mathbf{e}_1 & \mathbf{e}_1'\mathbf{e}_2 \\ \mathbf{e}_2'\mathbf{e}_1 & \mathbf{e}_2'\mathbf{e}_2 \end{array}\right]$$

is a matrix of sums of squares and products involving the residuals

$$\begin{aligned}\mathbf{e}_1 &= \mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{Z}_1\mathbf{a}_1,\\ \mathbf{e}_2 &= \mathbf{l} - \mathbf{X}_2\boldsymbol{\beta}_2 - \mathbf{Z}_2\mathbf{a}_2.\end{aligned}$$

14.3.2 Prior Distribution and Joint Posterior Density

Based on the assumptions of the infinitesimal model , the prior distribution $[\mathbf{a}_1, \mathbf{a}_2|\mathbf{A}, \mathbf{G}_0]$, is taken to be the multivariate normal process

$$\left[\begin{array}{c}\mathbf{a}_1\\ \mathbf{a}_2\end{array}\right]\Bigg|\,\mathbf{A}, \mathbf{G}_0 \sim N\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{0}\end{array}\right], \mathbf{G}_0 \otimes \mathbf{A}\right), \tag{14.30}$$

where

$$\mathbf{G}_0 = \left[\begin{array}{cc}\sigma^2_{a_1} & \sigma_{a_{12}}\\ \sigma_{a_{12}} & \sigma^2_{a_2}\end{array}\right]$$

is the additive genetic (co)variance matrix between the two traits, and $\mathbf{A}$ is the $q \times q$ additive relationship matrix between members of the genealogy (recall that, typically, $q > n$) .

The treatment of the vector $\boldsymbol{\beta}$ requires extending the developments for dealing with potential ECPs for the polychotomous trait to a bivariate situation. We adopt the partition

$$\boldsymbol{\beta} = \left[\begin{array}{c}\boldsymbol{\beta}_{1h}\\ \boldsymbol{\beta}_{1r}\\ \boldsymbol{\beta}_{2h}\\ \boldsymbol{\beta}_{2r}\end{array}\right],$$

where $\boldsymbol{\beta}_{2h}$ are location effects on liabilities whose levels have (possibly) ECPs; the vector $\boldsymbol{\beta}_{1h}$ contains the effects of these levels on the Gaussian trait. Subsequently, it is assumed that the prior distribution of $\boldsymbol{\beta}$ is such that its density factorizes as

$$p(\boldsymbol{\beta}) = p(\boldsymbol{\beta}_{1h}, \boldsymbol{\beta}_{2h})\, p(\boldsymbol{\beta}_{1r})\, p(\boldsymbol{\beta}_{2r}),$$

where

$$p(\boldsymbol{\beta}_{1r}) \propto \text{constant}$$

if $\boldsymbol{\beta}_{1r,\min} < \boldsymbol{\beta}_{1r} < \boldsymbol{\beta}_{1r,\max}$, and

$$\boldsymbol{\beta}_{2r} \sim N\left(\mathbf{0}, \mathbf{I}_{2r}10^6\right).$$

Further,

$$\left[\begin{array}{c}\boldsymbol{\beta}_{1h}\\ \boldsymbol{\beta}_{2h}\end{array}\right]\Bigg|\,\beta, \mathbf{B}_h \sim N\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{1}\beta\end{array}\right], \mathbf{B}_h \otimes \mathbf{I}\right), \tag{14.31}$$

where

$$\mathbf{B}_h = \begin{bmatrix} \sigma^2_{\beta_{1,h}} & \sigma_{\beta_{12,h}} \\ \sigma_{\beta_{12,h}} & \sigma^2_{\beta_{2,h}} \end{bmatrix}.$$

Here $\sigma^2_{\beta_{1,h}}$ and $\sigma^2_{\beta_{2,h}}$ are variance components and $\sigma_{\beta_{12,h}}$ is the covariance between the $\boldsymbol{\beta}_{1h}$ and $\boldsymbol{\beta}_{2h}$ effects on the two traits. As in the previous section, the scalar parameter β is assigned the uniform prior distribution

$$\beta|\beta_{\min}, \beta_{\max} \sim Un\left(\beta|\beta_{\min}, \beta_{\max}\right), \tag{14.32}$$

where $\beta_{\min}$ and $\beta_{\max}$ are appropriately chosen hyperparameters.

It is assumed that the matrices $\mathbf{R}_e$, $\mathbf{G}_0$, and $\mathbf{B}_h$ follow independent scaled-inverted Wishart distributions, a priori. The respective densities are

$$p\left(\mathbf{R}_e|\upsilon_e, \mathbf{V}_e\right) \propto |\mathbf{R}_e|^{-\frac{1}{2}(\upsilon_e+3)} \exp\left[-\frac{1}{2} tr\left(\mathbf{R}_e^{-1}\mathbf{V}_e^{-1}\right)\right], \tag{14.33}$$

$$p\left(\mathbf{G}_0|\upsilon_0, \mathbf{V}_0\right) \propto |\mathbf{G}_0|^{-\frac{1}{2}(\upsilon_0+3)} \exp\left[-\frac{1}{2} tr\left(\mathbf{G}_0^{-1}\mathbf{V}_0^{-1}\right)\right], \tag{14.34}$$

and

$$p\left(\mathbf{B}_h|\upsilon_h, \mathbf{V}_h\right) \propto |\mathbf{B}_h|^{-\frac{1}{2}(\upsilon_h+3)} \exp\left[-\frac{1}{2} tr\left(\mathbf{B}_h^{-1}\mathbf{V}_h^{-1}\right)\right],$$

where υ_i and $\mathbf{V}_i$ $(i = e, 0, h)$, are the usual parameters of the scaled inverted Wishart distributions. An important point: although the liabilities are (conditionally) Gaussian, the fact that the responses are categorical imposes some conditions on the form of the inverse Wishart distribution with density as in (14.33). We shall return to this issue later on.

The unknown thresholds in $\mathbf{t}$ delimiting the c categories of response, are assumed to be distributed a priori as ordered statistics from a uniform distribution in the interval $[t_{\min}, t_{\max}]$, as in (14.13).

The parameter vector is augmented with the missing data for the continuous trait $(\mathbf{y}_{1m})$ and with the unobserved liabilities $\mathbf{l}$. The parameters of the augmented model are represented as $(\boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l})$,where

$$\boldsymbol{\Omega} = (\boldsymbol{\beta}, \mathbf{a}, \mathbf{G}_0, \mathbf{R}_e, \beta, \mathbf{B}_h, \mathbf{t}).$$

Before embarking on the derivation of the fully conditional posterior distributions, we focus on the conditional distribution $[\mathbf{y}_{2o}|\mathbf{l}_o, \boldsymbol{\Omega}]$ of the observed categorical responses, given their liabilities and $\boldsymbol{\Omega}$. Recall that, given the liabilities l_i and the thresholds, the categorical responses $\mathbf{y}_{2o}$ are known with certainty. This means that given $(\boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l})$, that is, the parameters of the augmented model, the observations $\mathbf{y}_{1o}$ $(\mathbf{y}_{1m})$ and $\mathbf{y}_{2o}$ are conditionally independent. Ignoring hyperparameters in the notation, the joint posterior density of all uncertain variables in the augmented model

is

$$\begin{aligned} p\left(\boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l} | \mathbf{y}_{1o}, \mathbf{y}_{2o}\right) &\propto p\left(\mathbf{y}_{1o}, \mathbf{y}_{2o} | \boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l}\right) p\left(\boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l}\right) \\ &\propto p\left(\mathbf{y}_{2o} | \mathbf{y}_{1o}, \boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l}\right) p\left(\mathbf{y}_{1o} | \boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l}\right) p\left(\boldsymbol{\Omega}, \mathbf{y}_{1m}, \mathbf{l}\right) \\ &\propto p\left(\mathbf{y}_{2o} | \mathbf{y}_1, \boldsymbol{\Omega}, \mathbf{l}\right) p\left(\mathbf{y}_1, \mathbf{l} | \boldsymbol{\Omega}\right) p\left(\boldsymbol{\Omega}\right) \\ &\propto p\left(\mathbf{y}_1, \mathbf{l} | \boldsymbol{\Omega}\right) p\left(\mathbf{y}_{2o} | \mathbf{l}, \boldsymbol{\Omega}\right) p\left(\boldsymbol{\Omega}\right). \end{aligned} \tag{14.35}$$

The last line follows from the conditional independence of $\mathbf{y}_1$ and $\mathbf{y}_{2o}$, given $\mathbf{l}$ and $\mathbf{t}$. The first term on the right-hand side term of (14.35) is the density of the sampling model for the complete continuous data, as given in (14.29). The second term is not stochastic (given $\mathbf{l}$ and $\mathbf{t}$ one knows the categories of response with certainty), so it gets absorbed as a constant in Bayes theorem. The third term is the density of the joint prior distribution of the parameters, which is assumed to factorize as

$$p\left(\boldsymbol{\Omega}\right) \propto p\left(\boldsymbol{\beta}_{1h}, \boldsymbol{\beta}_{2h} | \beta, \mathbf{B}_h\right) p\left(\boldsymbol{\beta}_{1r}\right) p\left(\boldsymbol{\beta}_{2r}\right) p\left(\mathbf{a} | \mathbf{G}_0\right) p\left(\mathbf{G}_0\right) p\left(\mathbf{R}_e\right) p\left(\mathbf{B}_h\right) p\left(\mathbf{t}\right). \tag{14.36}$$

Although the conditioning on $\boldsymbol{\Omega}$ will be kept to simplify notation, note that in (14.35)

$$p\left(\mathbf{y}_1, \mathbf{l} | \boldsymbol{\Omega}\right) = p\left(\mathbf{y}_1, \mathbf{l} | \boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_e\right)$$

and

$$p\left(\mathbf{y}_{2o} | \mathbf{l}, \boldsymbol{\Omega}\right) = p\left(\mathbf{y}_{2o} | \mathbf{l}, \mathbf{t}\right).$$

Consider now the prior distribution of the variance–covariance matrix of the sampling model. If $\mathbf{R}_e$ is assigned an inverted Wishart distribution, then σ^2_{e2} (the residual variance of the liability) should be stochastic, so it cannot be fixed at some value. Treating σ^2_{e2} as a random variable requires parameterizing the model such that two thresholds, instead of only one, are given arbitrary values. Again, these must satisfy $t_{\min} < t_1 < t_2 < \cdots < t_{c-1} < t_{\max}$ (Sorensen et al., 1995). Typical choices are $t_1 = 0$ and $t_2 = 1$. However, this requires that the data fall into three or more categories of response. If the data are binary, this parameterization is not possible. When this is the case, a prior density $p\left(\mathbf{R}_e | \sigma^2_{e2} = 1\right)$ in the form of a scaled inverted Wishart can be specified. It turns out that the fully conditional posterior density $p\left(\mathbf{R}_e | ELSE, \sigma^2_{e2} = 1, data\right)$ is also scaled inverted Wishart. A simple way of drawing samples from this distribution is based on the properties of the inverted Wishart distribution described in Subsubsection 1.4.6 of Chapter 1. The algorithm was presented by Korsgaard et al. (1999), and is summarized at the end of this section.

14.3.3 Fully Conditional Posterior Distributions

As before the fully conditional posterior distribution of parameter x say, is written as $p\left(x | ELSE\right)$, where now $\mathbf{y} = \left(\mathbf{y}_{1o}, \mathbf{y}_{2o}\right)$. We start by deriving the fully conditional posterior distribution of the missing data $\left(\mathbf{y}_{1m}, \mathbf{l}\right)$.

Case	Observed data	Missing data	Generate
1	$y_{1o,i}, y_{2o,i}$	*none*	$l_{o,i}$
2	$y_{1o,i}$	*trait 2*	$l_{m,i}$
3	$y_{2o,i}$	*trait 1*	$y_{1m,i}, l_{o,i}$

TABLE 14.1. Possible patterns of observed and missing data.

Missing Continuous Data and Liabilities

From the sampling model in (14.27), it follows that a missing continuous record for individual i, say, is sampled from

$$y_{1m,i}|ELSE \sim N\left(E\left(y_{1m,i}|ELSE\right), Var\left(y_{1m,i}|ELSE\right)\right).$$

Here,

$$E\left(y_{1m,i}|ELSE\right) = \mathbf{x}'_{1,i}\boldsymbol{\beta}_1 + \mathbf{z}'_{1,i}\mathbf{a}_1 + \frac{\sigma_{e1,2}}{\sigma^2_{e2}}\left(l_i - \mathbf{x}'_{2i}\boldsymbol{\beta}_2 - \mathbf{z}'_{2i}\mathbf{a}_2\right) \quad (14.37)$$

and

$$Var\left(y_{1m,i}|ELSE\right) = \sigma^2_{e1}\left(1 - \frac{\left(\sigma_{e1,2}\right)^2}{\sigma^2_{e1}\sigma^2_{e2}}\right), \quad (14.38)$$

where $\sigma^2_{e2} = 1$. Thus, the conditional distribution of a missing Gaussian observation, given the liabilities, the observed data, and all parameters does not depend on the observed data. In these expressions, $\mathbf{x}'_{1i}$ ($\mathbf{x}'_{2i}$) and $\mathbf{z}'_{1i}$ ($\mathbf{z}'_{2i}$) are rows of matrices $\mathbf{X}_1$ ($\mathbf{X}_2$) and $\mathbf{Z}_1$ ($\mathbf{Z}_2$) associated with individual i. In (14.37), if the joint posterior is augmented with the residuals (instead of with the missing observations), elements of $\mathbf{x}'_{1i}$ and of $\mathbf{z}'_{1i}$ are all equal to zero.

The next step involves drawing samples of the underlying vector $\mathbf{l}$, noting that all liabilities are conditionally independent. The possible patterns of missing data for the i^{th} record ($i = 1, 2, \ldots, n$) are shown in Table 14.1. In the table, subscripts $1o, i$ ($2o, i$) associated with y, represent the observed continuous (categorical) record of the i^{th} individual, and subscripts $1m, i$ ($2m, i$) represent the missing continuous (categorical) record of the i^{th} individual.

If the pattern of missing records is as in Case (1), both records on the individual are available, so the fully conditional posterior distribution of $l_{o,i}$ must be derived here. From the joint posterior density presented in (14.35), and exploiting the conditional independence assumptions, the required conditional density is

$$\begin{aligned} p\left(l_{o,i}|ELSE\right) &\propto p\left(y_{1o,i}, l_{o,i}|\boldsymbol{\Omega}\right) p\left(y_{2o,i}|l_{o,i}, \boldsymbol{\Omega}\right) \\ &\propto p\left(l_{o,i}|y_{1o,i}, \boldsymbol{\Omega}\right)\left[\sum_{j=1}^{c} I\left(t_{j-1} < l_{o,i} \leq t_j\right) I\left(y_{2o,i} = j\right)\right]. \end{aligned} \quad (14.39)$$

From (14.27), density (14.39) is recognized as that of a truncated conditional normal distribution, with truncation points at t_{j-1} and t_j. The mean of the untruncated distribution is

$$E\left(l_{o,i}|y_{1o,i},\mathbf{\Omega}\right)=\mathbf{x}'_{2,i}\boldsymbol{\beta}_2+\mathbf{z}'_{2,i}\mathbf{a}_2+\frac{\sigma_{e1,2}}{\sigma^2_{e1}}\left(y_{1o,i}-\mathbf{x}'_{1,i}\boldsymbol{\beta}_1-\mathbf{z}'_{1,i}\mathbf{a}_1\right) \quad (14.40)$$

and the variance is

$$Var\left(l_{o,i}|y_{1o,i},\mathbf{\Omega}\right)=\sigma^2_{e2}\left(1-\frac{\left(\sigma_{e1,2}\right)^2}{\sigma^2_{e1}\sigma^2_{e2}}\right), \quad (14.41)$$

where $\sigma^2_{e2}=1$.

If the pattern of missing data is as in Case (2), the observation for the categorical trait is missing and $l_{m,i}$ must be generated. With $y_{2o,i}$ absent, it follows from (14.35) that the density of the fully conditional posterior distribution of $l_{m,i}$ is proportional to $p\left(\mathbf{y}_1,\mathbf{l}|\mathbf{\Omega}\right)$. Therefore, recalling the conditional independence structure,

$$\begin{aligned} p\left(l_{m,i}|ELSE\right) &\propto p\left(y_{1o,i},l_{m,i}|\mathbf{\Omega}\right) \\ &\propto p\left(l_{m,i}|y_{1o,i},\mathbf{\Omega}\right). \end{aligned} \quad (14.42)$$

From (14.35), this is a conditional normal distribution with mean and variance given by (14.40) and (14.41), respectively. Thus,

$$l_{m,i}|ELSE\sim N\left[E\left(l_{o,i}|y_{1o,i},\mathbf{\Omega}\right),Var\left(l_{o,i}|y_{1o,i},\mathbf{\Omega}\right)\right]. \quad (14.43)$$

Finally, if the pattern of missing data is as in Case (3), both $y_{1m,i}$ and $l_{o,i}$ must be sampled. From (14.35) we can write

$$\begin{aligned} &p\left(y_{1m,i},l_{o,i}|ELSE\right)\propto p\left(y_{1m,i},l_{o,i}|\mathbf{\Omega}\right)p\left(y_{2o,i}|l_{o,i},\mathbf{\Omega}\right) \\ &=p\left(y_{1m,i},l_{o,i}|\mathbf{\Omega}\right)\sum_{j=1}^{c}I\left(t_{j-1}<l_{o,i}\leq t_j\right)I\left(y_{2o,i}=j\right) \\ &=p\left(y_{1m,i}|l_{o,i},\mathbf{\Omega}\right)p\left(l_{o,i}|\mathbf{\Omega}\right)\sum_{j=1}^{c}I\left(t_{j-1}<l_{o,i}\leq t_j\right)I\left(y_{2o,i}=j\right). \end{aligned} \quad (14.44)$$

A simple way of obtaining samples from the distribution with density (14.44) is first to sample a realized value $l^*_{o,i}$ from the normal distribution $\left[l_{o,i}|\mathbf{\Omega}\right]$ truncated at t_{j-1} and at t_j, and second, to sample $y_{1m,i}$ from the conditional normal distribution $\left[y_{1m,i}|l^*_{o,i},\mathbf{\Omega}\right]$. Again, conditional independence holds, so the process can be effected piecewise, observation by observation.

Location Effects

We derive now the fully conditional posterior distribution of $\boldsymbol{\theta}'=\left(\boldsymbol{\beta}',\mathbf{a}'\right)$ and, subsequently, that of the scalar β. From (14.35), and noting that $p\left(\mathbf{y}_{2o}|\mathbf{l},\mathbf{\Omega}\right)$ is not a function of $\boldsymbol{\theta}$, we get

$$\begin{aligned} p(\boldsymbol{\theta}|ELSE) &\propto p(\mathbf{y}_1, \mathbf{l}|\boldsymbol{\Omega})\, p(\boldsymbol{\Omega}) \\ &= p(\mathbf{y}_1, \mathbf{l}, \boldsymbol{\theta}, \beta, \mathbf{G}_0, \mathbf{R}_e, \mathbf{B}_h, \mathbf{t}) \\ &\propto p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{l}, \beta, \mathbf{G}_0, \mathbf{R}_e, \mathbf{B}_h)\,. \end{aligned} \tag{14.45}$$

This is the density of a bivariate Gaussian hierarchical model in which the continuous responses and the liabilities (the complete continuous data) are the observations. Hence, the conditional posterior distribution of the location effects $\boldsymbol{\theta}$ is a multivariate normal process, with its mean vector and variance–covariance matrix calculated as seen in Chapter 13. The general expression for the fully conditionals is very similar to equations (13.50), (13.51), and (13.52), with a slight modification in the interpretation of some terms, to accommodate the different model scenarios.

Consider the fully conditional posterior distributions of $\boldsymbol{\beta}$ and $\mathbf{a}$. Let

$$\mathbf{G}_0^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix},$$

$$\mathbf{R}_e^{-1} = \begin{bmatrix} r^{11} & r^{12} \\ r^{21} & r^{22} \end{bmatrix},$$

and

$$\mathbf{B}_h^{-1} = \begin{bmatrix} b^{11} & b^{12} \\ b^{21} & b^{22} \end{bmatrix}.$$

Then

$$[\boldsymbol{\beta}|ELSE] \sim N\left(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{V}}_\beta\right), \tag{14.46}$$

where the mean vector has the following four partitions:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{1h} \\ \widehat{\boldsymbol{\beta}}_{2h} \\ \widehat{\boldsymbol{\beta}}_{1r} \\ \widehat{\boldsymbol{\beta}}_{2r} \end{bmatrix}.$$

An explicit representation of the mean vector is arrived at by writing the incidence matrix of all elements of $\boldsymbol{\beta}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{1h} & \mathbf{0} & \mathbf{X}_{1r} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{2h} & \mathbf{0} & \mathbf{X}_{2r} \end{bmatrix}.$$

With this notation

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)\mathbf{X} + \mathbf{P}^{-1}\right]^{-1}\left[\mathbf{X}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)(\mathbf{v} - \mathbf{Z}\mathbf{a}) + \mathbf{m}\right],$$

where

$$\mathbf{Z}\mathbf{a} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix}\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix},$$

$$\mathbf{P}^{-1} = \begin{bmatrix} b^{11}\mathbf{I} & b^{12}\mathbf{I} & \mathbf{0} & \mathbf{0} \\ b^{21}\mathbf{I} & b^{22}\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 10^{-6}\mathbf{I} \end{bmatrix},$$

and

$$\mathbf{m} = \mathbf{P}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}\beta \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} b^{12}\beta\mathbf{I} \\ b^{22}\beta\mathbf{I} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

The variance–covariance matrix of the conditional posterior distribution in (14.46) is

$$\widehat{\mathbf{V}}_\beta = \left[\mathbf{X}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)\mathbf{X} + \mathbf{P}^{-1}\right]^{-1}.$$

The fully conditional posterior distribution of $\mathbf{a}$ is

$$[\mathbf{a}|ELSE] \sim N\left(\widehat{\mathbf{a}}, \widehat{\mathbf{V}}_a\right), \tag{14.47}$$

where

$$\widehat{\mathbf{a}} = \left[\mathbf{Z}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)\mathbf{Z} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right]^{-1} \mathbf{Z}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)(\mathbf{v} - \mathbf{X}\boldsymbol{\beta}),$$

and

$$\widehat{\mathbf{V}}_a = \left[\mathbf{Z}'\left(\mathbf{R}_e^{-1} \otimes \mathbf{I}_n\right)\mathbf{Z} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1}\right]^{-1}.$$

The location effect remaining to be sampled is β. It follows from the joint posterior density (14.35) and from the form of the prior in (14.36) that

$$\begin{aligned} p(\beta|ELSE) &\propto p(\boldsymbol{\beta}_{1h}, \boldsymbol{\beta}_{2h}|\beta, \mathbf{B}_h) \\ &\propto p\left(\boldsymbol{\beta}_{1h}|\sigma^2_{\beta_{1,h}}\right) p(\boldsymbol{\beta}_{2h}|\boldsymbol{\beta}_{1h}, \beta, \mathbf{B}_h) \\ &\propto p(\boldsymbol{\beta}_{2h}|\boldsymbol{\beta}_{1h}, \beta, \mathbf{B}_h), \end{aligned}$$

which is multivariate normal, since $[\boldsymbol{\beta}_{1h}, \boldsymbol{\beta}_{2h}|\beta, \mathbf{B}_h]$ is multivariate normal. Further, (14.31) indicates that the pairs $\left(\beta_{1h,k}, \beta_{2h,k}\right)$ and $\left(\beta_{1h,k'}, \beta_{2h,k'}\right)$ are mutually independent, a priori, where $k = 1, 2, ..., H$, denotes a level of the factor possibly associated with ECP problems. Hence

$$\begin{aligned} p(\beta|ELSE) &\propto \prod_{k=1}^{H} p\left(\beta_{2h,k}|\beta_{1h,k}, \beta, \mathbf{B}_h\right) \\ &\propto \prod_{k=1}^{H} \exp\left[-\frac{\left(\beta_{2h,k} - \mu_{\beta_{2.1,k}}\right)^2}{2\tau^2_{\beta_{2.1}}}\right], \end{aligned}$$

where

$$\mu_{\beta_{2.1,k}} = \beta + \frac{\sigma_{\beta_{12,h}}}{\sigma^2_{\beta_{1,h}}}\beta_{1h,k},$$

and

$$\tau^2_{\beta_{2.1}} = \sigma^2_{\beta_{2,h}} \left(1 - \frac{\left(\sigma_{\beta_{12,h}}\right)^2}{\sigma^2_{\beta_{1,h}} \sigma^2_{\beta_{2,h}}} \right).$$

Now, defining the offset $\beta^*_{2h,k} = \beta_{2h,k} - \frac{\sigma_{\beta_{12,h}}}{\sigma^2_{\beta_{1,h}}} \beta_{1h,k}$, the density of the conditional posterior distribution of β can be written as

$$p\left(\beta | ELSE\right) \propto \prod_{k=1}^{H} \exp\left[-\frac{\left(\beta^*_{2h,k} - \beta\right)^2}{2\tau^2_{\beta_{2.1}}} \right].$$

Viewed as a function of β, it follows that this is the density of a normal distribution with mean $\overline{\beta} = \frac{1}{H} \sum_{k=1}^{H} \beta^*_{2h,k}$ and variance $\frac{1}{H}\tau^2_{\beta_{2.1}}$, where H is the number of levels of the factor with ECPs. In short,

$$\beta | ELSE \sim N\left(\frac{\sum_{k=1}^{H} \beta^*_{2h,k}}{H}, \frac{\sigma^2_{\beta_{2,h}}}{H} \left(1 - \frac{\left(\sigma_{\beta_{12,h}}\right)^2}{\sigma^2_{\beta_{1,h}} \sigma^2_{\beta_{2,h}}} \right) \right). \tag{14.48}$$

Dispersion Parameters

The fully conditional posterior distributions of the covariance matrices are derived by extracting the relevant terms from (14.35). In view of the form of the prior in (14.36), all dispersion matrices are conditionally independent, given all other parameters. Thus,

$$\begin{aligned} p\left(\mathbf{G}_0 | ELSE\right) &\propto p\left(\mathbf{\Omega}\right) \\ &\propto p\left(\mathbf{a} | \mathbf{G}_0\right) p\left(\mathbf{G}_0\right) \end{aligned} \tag{14.49}$$

which is identical to (13.55). Therefore, the distribution is as in (13.56)

$$\mathbf{G}_0 | ELSE \sim IW_2\left(\left(\mathbf{V}_a^{-1} + \mathbf{S}_a \right)^{-1}, \upsilon_a + q \right). \tag{14.50}$$

Likewise,

$$p\left(\mathbf{B}_h | ELSE\right) \propto p\left(\boldsymbol{\beta}_{1h}, \boldsymbol{\beta}_{2h} | \beta, \mathbf{B}_h\right) p\left(\mathbf{B}_h\right). \tag{14.51}$$

This is the density of the inverse Wishart process

$$\mathbf{B}_h | ELSE \sim IW_2\left(\left(\mathbf{V}_h^{-1} + \mathbf{S}_h \right)^{-1}, \upsilon_h + H \right), \tag{14.52}$$

where

$$\mathbf{S}_h = \begin{bmatrix} \boldsymbol{\beta}'_{1h}\boldsymbol{\beta}_{1h} & \boldsymbol{\beta}'_{1h}\left(\boldsymbol{\beta}_{2h} - \mathbf{1}\beta\right) \\ \left(\boldsymbol{\beta}_{2h} - \mathbf{1}\beta\right)'\boldsymbol{\beta}_{1h} & \left(\boldsymbol{\beta}_{2h} - \mathbf{1}\beta\right)'\left(\boldsymbol{\beta}_{2h} - \mathbf{1}\beta\right) \end{bmatrix}.$$

Similarly, for the residual covariance matrix,

$$\begin{aligned} p\left(\mathbf{R}_e|ELSE\right) &\propto p\left(\mathbf{y}_1, \mathbf{l}|\boldsymbol{\Omega}\right) p\left(\boldsymbol{\Omega}\right) \\ &\propto p\left(\mathbf{y}_1, \mathbf{l}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{R}_e\right) p\left(\mathbf{R}_e\right). \end{aligned}$$

This is identical to (13.53). The resulting fully conditional distribution is:

$$\mathbf{R}_e|ELSE \sim IW_2\left(\left(\mathbf{S}_e + \mathbf{V}_e^{-1}\right)^{-1}, \upsilon_e + n\right) \tag{14.53}$$

with the term $\mathbf{S}_e$ appropriately defined.

Thresholds

The fully conditional posterior distribution of the j^{th} unknown threshold is obtained as follows. From (14.35)

$$\begin{aligned} &p\left(t_j|ELSE\right) \propto p\left(\mathbf{y}_{2o}|l, \boldsymbol{\Omega}\right) p\left(\Omega\right) \\ &\propto \prod_{i=1}^{n_{2o}} \left[I\left(t_{j-1} < l_{o,i} < t_j\right) I\left(y_{2o,i} = j\right) + I\left(t_j < l_{o,i} < t_{j+1}\right) I\left(y_{2o,i} = j+1\right)\right], \end{aligned} \tag{14.54}$$

where n_{2o} is the number of observed categorical records. This expression is identical to (14.19).

14.3.4 The Gibbs Sampler

To summarize, a Gibbs sampler can be run as follows:

1. Read in the data file and generate missing data (and liabilities) from (14.39), (14.43) or (14.44), depending on the pattern of missing data.
2. Build the mixed model equations.
3. Sample $\boldsymbol{\theta}$ from (14.45), with the distributions given explicitly in (14.46) and (14.47). Recall that the samples can be drawn either blockwise or piecewise, in which case the expression must be modified slightly.
4. Sample the scalar β from (14.48).
5. Sample the covariance matrices from (14.50), (14.52) and (14.53).
6. Sample the thresholds from (14.54).
7. Go to Step 1 or exit if chain is long enough.

Like in the Gaussian multiple-trait case, note that the coefficient matrix of the mixed model equations has to be recreated every iterate with the strategy described above.

14.3.5 Implementation with Binary Traits

As mentioned above, when there are only two categories of response the residual covariance matrix can be specified as

$$\mathbf{R}_e = \begin{bmatrix} \sigma^2_{e1} & \sigma_{e1,2} \\ \sigma_{e1,2} & 1 \end{bmatrix}.$$

From a Bayesian perspective, this is a random matrix with two stochastic elements $\left(\sigma^2_{e1}, \sigma_{e1,2}\right)$ instead of three. This is so because the residual variance in the liability scale is set equal to 1, since the parameter is not identifiable from the likelihood. In this situation, rather than adopting an inverse Wishart prior for $\mathbf{R}_e$, one can assign a conditional inverse Wishart prior (an inverse Wishart distribution, conditional on $\sigma^2_{e2} = 1$). The resulting fully conditional posterior distribution

$$\left[\mathbf{R}_e | ELSE, \sigma^2_{e2} = 1, data\right] \tag{14.55}$$

turns out to have the form of a conditional inverse Wishart distribution also (conditional on $\sigma^2_{e2} = 1$). Korsgaard et al. (1999) have shown how to draw samples from such a distribution. They described a simple algorithm, where a more general formulation than the one described here can be found. Based on the properties of the Wishart distribution presented in Subsection 1.4.6 of Chapter 1, the Gibbs sampler can proceed as follows. Implement the algorithm in the manner described in the previous section, with the exception of the draw involving (14.53), which is replaced by a draw from (14.55). In order to obtain a realized value from the latter,

- Sample X_1 from (1.112).
- Sample X_2 from (1.113).
- Construct T_{11} from (1.109), T_{12} from (1.110) and set $T_{22} = X_3^{-1} = 1$ Then T_{11}, T_{12} and T_{22} are the elements of the draw from

$$\mathbf{R}_e | ELSE, \sigma^2_{e2} = 1, data.$$

A final comment is in order. The threshold model is perhaps appealing in quantitative genetics because all the theory available for additive inheritance carries on to the liability scale. The reader must be aware that there are alternative methods of analysis of categorical responses; see, for example, Fahrmeir and Tutz (2001) or Agresti (1989, 1990, 1996).

15

Bayesian Analysis of Longitudinal Data

15.1 Introduction

Suppose individuals are sampled from a set of populations, with the latter defined in a statistical, rather than demographic, sense. In at least some of the individuals, the trajectory of a trait is measured over a period of time, collecting, thus, a time series of observations. Examples of such longitudinal trajectories are: milk yield of a dairy cow at several points in the course of lactation, body weight or feed intake measured repeatedly during some test period in which animals grow, prolificacy (litter size produced) of a sow in the course of her lifetime, wool growth assessed at different stages of the development of a ewe, presence or absence of clinical mastitis in a dairy cow in each bi-weekly period from calving until the end of lactation, the height of a tree as it grows, etc. The trait monitored may be either "continuous" (e.g., milk yield or body weight), or a count (e.g., lambs per litter over a series of litters), or binary (presence or absence of a disease at a given time). Issues of interest may include inferring the expected trajectory of the trait within an individual or assessing sources of variation, genetic and nongenetic, among the trajectory patterns of groups of individuals. This type of problem is fairly old in the study of animal systems. For example, lactation curves and growth functions have been the subject of research for decades in milk- and meat-producing species, respectively. Animals differ in their rate of growth or adult body weight, and it is known that there is genetic variation for these features, both between and within breeds, that

can be exploited in animal breeding programs. The question posed is: How can this variation be assessed adequately?

In the animal and veterinary sciences, there has been renewed interest in the analysis of longitudinal records of performance. Perhaps this is a consequence of more intensive recording systems (for instance, in dairy cattle production it is possible to monitor instantaneous milk flow) and of better statistical methods for the analysis of longitudinal mixed effects models. In particular, linear random regression models (Laird and Ware, 1982) or similar approaches have been applied in animal breeding, where there is a large body of literature in connection with the analysis of "test-day" yields in dairy cattle (e.g., Kirkpatrick et al., 1994; Jamrozik and Schaeffer, 1997; Wiggans and Goddard, 1997). Similar applications have been made in meat-producing species. For example, an assessment of growth in beef cows from 19 to 119 months of age is in Meyer (1999).

In this chapter, the analysis of longitudinal data will be dealt with in a parametric Bayesian framework, to illustrate the flexibility and attractiveness of the paradigm. First, a description is given of hierarchical or multistage models for describing longitudinal observations, assuming Gaussian processes throughout. Second, methods for an approximate Bayesian analysis of these models are described. These methods can be construed as generalizations of the usual "tandem" employed for analysis of mixed effects linear models in animal breeding, consisting of BLUP (of random effects) plus likelihood-based procedures for parameter inference. However, at least in theory, a Bayesian hierarchical probability model can yield exact finite sample inferences about all unknowns. Hence, a subsequent section discusses an implementation based on MCMC procedures. We also discuss some extensions, including alternative structures for the residual dispersion of the process.

15.2 Hierarchical or Multistage Models

Envisage a setting where, in a randomly drawn sample, each individual is measured longitudinally at several times. For example, suppose that male and female rabbits from several breeds are weighed at several phases of their development, from near birth to the adult stage. An objective might be to study growth patterns of the two sexes in each of the breeds, while taking into account interindividual variability. Typically, there will be variation in the number of measurements per individual, leading at least to longitudinal unbalancedness. In individuals with sparse information, the individual trajectories would probably be estimated imprecisely, unless information from relatives is abundant or prior information about the expected trajectory is sharp.

A hierarchical or multistage model consists of a series of nested functional specifications, together with the associated distributional assumptions. An important paper introducing the basic ideas is that of Lindley and Smith (1972). In the context of longitudinal data, at the first stage of the model, a mathematical function is used to describe the expected trajectory of individuals, and a stochastic residual having some distribution reflects the departure of the observations from such a trajectory. At the second stage, a submodel is used to describe the interindividual variation of parameters of the first-stage specification. A second-stage residual is included to reflect the inability of the submodel to explain completely the variation of the parameters. Additional stages can be imposed in a Bayesian context to describe uncertainty about all other unknown parameters. We shall now proceed to describe each of these stages systematically.

15.2.1 First Stage

The trajectory (body weights of the same individual, for example) will be described with the parametric model

$$\mathbf{y}_i = \mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right) + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \ldots, M, \tag{15.1}$$

where $\mathbf{y}_i = \{y_{ij}\}$ $(i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, n_i)$ is an $n_i \times 1$ vector of records on the trajectory of individual i; $\mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right)$ is its expected trajectory (e.g., expected growth curve) given a vector of animal-specific parameters $\boldsymbol{\theta}_i$ of order $r \times 1$, and $\mathbf{t}_i$ is an $n_i \times 1$ vector of known times of measurement. In (15.1), the $n_i \times 1$ residual vector $\boldsymbol{\varepsilon}_i$ represents the inability of the function $\mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right)$ of reproducing the observed body weights $\mathbf{y}_i$ exactly. An observation on individual i at time j is then

$$y_{ij} = f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right) + \varepsilon_{ij}, \tag{15.2}$$

so the parameters $\boldsymbol{\theta}_i$ dictate the form of the expected trajectory of individual i. For example, when describing animal growth, use is often made of what is called a Gompertz growth function (e.g., Blasco and Varona, 1999). Here $r = 3$; one of the parameters represents adult or asymptotic weight (i.e., body weight as time goes to infinity), the second parameter is related to growth rate, and the third parameter bears an interpretation in terms of the "initial conditions" of growth. Often, animals that are measured at a given time are clustered in different contemporary groups. For example, for some dairy cows the yield may be recorded in February, while for other cows it may be recorded in December. In this situation, the model can be written in a slightly more general form as

$$y_{ijk} = G_k + f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right) + \varepsilon_{ijk},$$

where G_k is an effect peculiar to all measurements taken on individuals belonging to group k $(k = 1, 2, ..., K)$, e.g., month–year at which milk yield

at time t_{ij} is measured on cow i. Since the group effect enters linearly in the model, it can be dealt with in a straightforward manner. For example, in the context of a Gibbs sampler, after G_k is drawn, one would form the offset $y_{ijk} - G_k$, with the conditional model thus being again in the form of specification (15.2). Hence, the simpler model suffices for descriptive purposes without great loss of generality.

The relationship between observed body weights and parameters may be linear or nonlinear, the latter being the case in the Gompertz function. In a linear specification, the derivatives of the model with respect to the parameters do not depend on $\boldsymbol{\theta}_i$. This can be stated as

$$\frac{\partial f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)}{\partial \boldsymbol{\theta}_i} = \mathbf{h}_{ij},$$

where $\mathbf{h}_{ij}$ is an $r \times 1$ vector of constants not involving $\boldsymbol{\theta}_i$. On the other hand, in a nonlinear model,

$$\frac{\partial f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)}{\partial \boldsymbol{\theta}_i} = \mathbf{h}_{ij}\left(\boldsymbol{\theta}_i\right),$$

indicating that the vector $\mathbf{h}_{ij}\left(\boldsymbol{\theta}_i\right)$ involves the parameters, although some may enter linearly into the model.

Example 15.1 *Quadratic trajectory*
Suppose a longitudinal process can be described with the first-stage model

$$y_{ij} = a_i + b_i t_{ij} + c_i t_{ij}^2 + \varepsilon_{ij}, \quad i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, n_i.$$

Here,

$$\boldsymbol{\theta}_i = \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix}, \text{ and } \mathbf{t}_i = \begin{bmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{in_i} \end{bmatrix}.$$

In matrix form, the observations made on individual i can be written as

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{bmatrix} \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix},$$

so

$$\mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right) = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{bmatrix} \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} = \mathbf{H}_i \boldsymbol{\theta}_i,$$

where $\mathbf{H}_i$ is an incidence matrix. Then

$$\frac{\partial \mathbf{f}_i'(\boldsymbol{\theta}_i, \mathbf{t}_i)}{\partial \boldsymbol{\theta}_i} = \mathbf{H}_i' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_{i1} & t_{i2} & \cdots & t_{in_i} \\ t_{i1}^2 & t_{i2}^2 & \cdots & t_{in_i}^2 \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{h}_{i1} & \mathbf{h}_{i2} & \cdots & \mathbf{h}_{in_i} \end{bmatrix}.$$

Since the matrix of derivatives (or incidence matrix) does not involve any of the parameters, the model is linear. ■

Example 15.2 *Growth curve*

Typically, animals grow at an increasing rate from birth to puberty, and at a decreasing rate thereafter, until a mature or asymptotic weight is reached. The resulting average growth curve (assuming a large group of animals treated under similar conditions) is usually S-shaped. Suppose the trajectory of the body weight of an individual from birth onward can be described by the mathematical model

$$y_{ij} = A_i \left[1 - B_i \exp\left(-K_i t_{ij}\right)\right]^{-1} + \varepsilon_{ij},$$

where A_i, B_i, and K_i are parameters indicating some aspect of growth. For example, for positive K_i, when $t_{ij} \to \infty$, $E(y_{ij}) \to A_i$, which is interpretable as adult body weight. Likewise, as $t_{ij} \to 0$, $E(y_{ij}) \to A_i/(1 - B_i)$, interpretable as birth weight. Hence, $1/(1 - B_i)$ is the fraction of adult body weight attained at birth and $-B_i/(1 - B_i)$ represents the fraction yet to be attained during the growth process. Then, by definition, B_i must be a negative parameter and related to degree of maturity at birth (as $B_i \to 0$ the animal is more mature at birth). The parameter K_i is a rate. Here, as in Example 15.1, $r = 3$ and the derivatives of interest are

$$\mathbf{h}_{ij}(\boldsymbol{\theta}_i) = \begin{bmatrix} \dfrac{\partial f_{ij}(\boldsymbol{\theta}_i, t_{ij})}{\partial A_i} \\ \dfrac{\partial f_{ij}(\boldsymbol{\theta}_i, t_{ij})}{\partial B_i} \\ \dfrac{\partial f_{ij}(\boldsymbol{\theta}_i, t_{ij})}{\partial K_i} \end{bmatrix}$$
$$= \begin{bmatrix} \left[1 - B_i \exp\left(-K_i t_{ij}\right)\right]^{-1} \\ A_i \exp\left(-K_i t_{ij}\right) \left[1 - B_i \exp\left(-K_i t_{ij}\right)\right]^{-2} \\ -A_i B_i t_{ij} \exp\left(-K_i t_{ij}\right) \left[1 - B_i \exp\left(-K_i t_{ij}\right)\right]^{-2} \end{bmatrix}.$$

Clearly the model is not linear, although the degree of nonlinearity varies from parameter to parameter. For example, the partial gradient of the model with respect to A_i involves only B_i and K_i, while the other two partial gradients involve all three parameters. ■

Example 15.3 *Inverse third-order polynomial*
A third-order inverse polynomial has the functional form

$$y_{ij} = \frac{1}{\beta_{i0} + \beta_{i1}t_{ij} + \beta_{i2}t_{ij}^2 + \beta_{i3}t_{ij}^3 + \varepsilon_{ij}},$$

where $\boldsymbol{\theta}_i' = [\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i3}]$ are parameters peculiar to individual i and ε_{ij} is a residual having a null expectation. The first derivatives of the model with respect to the parameters can be written as

$$\frac{\partial f_{ij}(\boldsymbol{\theta}_i, t_{ij})}{\partial \beta_k} = \frac{-x_{ijk}}{\left(\beta_{i0} + \beta_{i1}t_{ij} + \beta_{i2}t_{ij}^2 + \beta_{i3}t_{ij}^3 + \varepsilon_{ij}\right)^2},$$
$$x_{ijk} = t_{ij}^k, \quad k = 0, 1, 2, 3.$$

The model, thus, is not linear in the parameters. Note, however, that if a reciprocal transformation of the observations is made, the model becomes linear since

$$z_{ij} = \frac{1}{y_{ij}} = \beta_{i0} + \beta_{i1}t_{ij} + \beta_{i2}t_{ij}^2 + \beta_{i3}t_{ij}^3 + \varepsilon_{ij},$$

and the derivatives now do not involve any of the parameters. It is important to observe, however, that if the original model had the error entering as

$$y_{ij} = \frac{1}{\beta_{i0} + \beta_{i1}t_{ij} + \beta_{i2}t_{ij}^2 + \beta_{i3}t_{ij}^3} + \varepsilon_{ij},$$

the model would be nonlinear, even after making a reciprocal transformation. The two models for y_{ij} are not the same, as distinct assumptions are made about the errors. In the first model, the errors are additive to the expectation of the reciprocal (z_{ij}) of the random variable of interest. In the second model, the errors are additive to the reciprocal of the expectation of the reciprocal of y_{ij}. ■

Returning to the first-stage model, the entire vector of records can be represented as

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}, \mathbf{t}) + \boldsymbol{\varepsilon}, \tag{15.3}$$

where $\boldsymbol{\theta}$ is the $Mr \times 1$ vector of parameters of all individuals, $\mathbf{t}$ contains times of measurement, and $\boldsymbol{\varepsilon}$ is the $\sum_{i=1}^{M} n_i \times 1$ vector of residuals. Commonly, it is assumed that the first-stage residuals are mutually independent between individuals, but some dependence within trajectories may exist. Hereinafter, given the parameters, the observations taken in different individuals will be assumed to be conditionally independent of each other. Possible dependencies, such as those resulting from genetic or environmental relatedness between individuals, can be introduced in the next stage of the model. Assuming normality of the residuals (sometimes, a thick-tailed

distribution, such as Student-t, may be a more sensible specification), the density of the first-stage distribution is expressible as

$$\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\gamma} \sim \mathbf{N}\left[\mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right), \mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right], \quad i = 1, 2, \ldots, M, \tag{15.4}$$

with $\mathbf{y}_i$ being independent of $\mathbf{y}_j$, for all such pairs, conditionally on the parameters and on $\boldsymbol{\gamma}$. In (15.4), $\mathbf{R}_i\left(\boldsymbol{\gamma}\right)$ is an $n_i \times n_i$ first-stage variance–covariance matrix, which depends on $\boldsymbol{\gamma}$, a vector of dispersion parameters. For example, if residuals are independently and identically distributed within individuals, then $\mathbf{R}_i\left(\boldsymbol{\gamma}\right) = \mathbf{I}_{n_i}\gamma$, where γ is the variance about the expected trajectory, so $\boldsymbol{\gamma}$ would be a scalar parameter here.

The form of the matrix $\mathbf{R}_i\left(\boldsymbol{\gamma}\right)$ depends on the dispersion assumptions made. Some possible alternatives to the preceding specification are discussed below.

(1) Residuals may be independently distributed, but heteroscedastic across individuals. Then

$$\mathbf{R}_i\left(\boldsymbol{\gamma}\right) = \mathbf{I}_{n_i}\gamma_i, \quad i = 1, 2, \ldots, M.$$

Here $\boldsymbol{\gamma}' = [\gamma_1, \gamma_2, \ldots, \gamma_M]$ and γ_i is the variance about the trajectory of individual i.

(2) The residuals may be heteroscedastic across times of measurement. In this situation, there would be a first-stage residual variance component for each of the times at which the trajectory is evaluated (if this is done at fixed times, as in experimental settings, for example, every three weeks when measuring body weights in children).

(3) Perhaps the residuals are neither homoscedastic nor independently distributed. For example, there may be a first-order autoregressive process with heterogeneous variance across the times at which measurements are taken, such that

$$Cov\left(\varepsilon_{it}, \varepsilon_{i(t+k)}\right) = \rho^k \gamma_t,$$

where ρ is a first-stage residual correlation and γ_t is the residual variance at time t $(t = 1, 2, \ldots, T)$. If the variance were homogeneous, this would make observations taken adjacently in time to be more correlated than those farther apart. Another specification may be a Markov-type process, where observations are dependent only if the measurement times are contiguous, but not otherwise.

(4) A structural model may be entertained for the residual variance. For example, Blasco and Varona (1999) used a Gompertz function to describe body weight growth in rabbits, and proposed modeling the trajectory of the residual standard deviation over time using the Gompertz function as well. The parameters of this function were assumed to be homogeneous across individuals. An even more ambitious model would consist of building up a hierarchical model for the trajectory of the standard deviation, where parameters vary according to some explanatory variables.

The density of the conditional distribution in (15.4), over all individuals, is then

$$p\left(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\gamma}\right) \propto \prod_{i=1}^{M} \left|\mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon}_i\right\}, \tag{15.5}$$

where $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}_i\right)$, from (15.1). If individual parameters are inferred via maximum likelihood from this first stage model, unstable estimates may be obtained, specially for subjects having few observations. If heterogeneity between individuals is accounted for somehow using a submodel, perhaps the total variation can be described in terms of fewer parameters.

15.2.2 Second Stage

The second stage of the model is a statement of how individual-specific parameters are thought to vary according to explanatory factors, some of these perhaps representing genetic sources of variation. In brief, as stated earlier, the first stage of the model delineates the expected trajectory that longitudinal observations take within a given subject, while the second stage accounts for cross-sectional (between-subject) heterogeneity of parameters of the trajectory.

In order to facilitate implementation, it is convenient to assume that the second stage of the model is linear in the effects of the explanatory variables. However, at least in theory, there is no reason to preclude a nonlinear specification, particularly if this is dictated by mechanistic considerations. Hereinafter, it will be assumed that the trajectory parameters are described suitably by the linear model

$$\boldsymbol{\theta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i, \quad i = 1, 2, \ldots, M. \tag{15.6}$$

Above, the vector $\boldsymbol{\beta}$ represents the effects of p explanatory variables contained in the $r \times p$ matrix $\mathbf{X}_i$ (without loss of generality, this matrix will be assumed to have full-column rank), $\mathbf{u}_i$ are subject-specific effects on each of the r parameters, and $\mathbf{e}_i$ is a vector of second-stage residuals. Similar to the errors in the first stage, these residuals represent discrepancies between the second-stage explanatory structure $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$ and the "true values" $\boldsymbol{\theta}_i$. In animal breeding applications, for example, the vector $\mathbf{u}_i$ may be additive genetic effects on trajectory parameters, and these may or may not be identifiable separately from the residual vector $\mathbf{e}_i$, depending on the genetic relationship structure. For example, suppose the $\boldsymbol{\theta}'$s represent parameters of a three-coefficient lactation curve for each of 100 cows, and that such cows are progeny of a set of 20 sires, each having 5 daughters. In this case, one may wish to employ a double subscript in the notation and let $\boldsymbol{\theta}_{ij}$ be the coefficients for daughter j of sire i. Here $\mathbf{u}_i$ might be a 3×1 vector of "sire effects", common to all progeny of sire i $(i = 1, 2, \ldots, 20)$ and

the second-stage residual vector would be $\mathbf{e}_{ij}$, representing the discrepancy $\boldsymbol{\theta}_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} - \mathbf{u}_i$ peculiar to cow ij. Parameters of the distributions of $\mathbf{u}_i$ and $\mathbf{e}_{ij}$ are identifiable, as is the case in a standard analysis of variance, even if the $\boldsymbol{\theta}_{ij}$ are not observable.

The second-stage distributional assumptions pertain to the uncertainty induced by the presence of $\mathbf{e}_i$ in model (15.6), given $\boldsymbol{\beta}$ and $\mathbf{u}_i$. It is often convenient to postulate that

$$\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e \sim N\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i, \boldsymbol{\Sigma}_e\right), \tag{15.7}$$

implying that

$$\mathbf{e}_i | \boldsymbol{\Sigma}_e \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_e\right),$$

where the second stage variance–covariance matrix has the form

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1r} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_r^2 \end{bmatrix}. \tag{15.8}$$

Here the diagonal elements are the variances of the second-stage residuals and the off-diagonals are corresponding covariances; for example, σ_r^2 is the second-stage variance of parameter r and $\sigma_{r-1,r}$ is the second-stage covariance between parameters $r-1$ and r. In some instances, one may wish to assign a thick-tailed or robust distribution to the residuals, e.g., an r-variate t distribution. In this situation, one would write $\mathbf{e}_i | \nu_e, \boldsymbol{\Sigma}_e \sim t_r\left(\mathbf{0}, \boldsymbol{\Sigma}_e, \nu_e\right)$ to denote a t distribution of dimension r, having a null mean vector, variance–covariance $\boldsymbol{\Sigma}_e$, and degrees of freedom ν_e. It must be noted that in a multivariate-t distribution, $\boldsymbol{\Sigma}_e = \frac{\nu_e}{\nu_e - 2}\mathbf{S}_e$, where $\mathbf{S}_e$ is a scale matrix, so $\nu_e > 2$ is a necessary condition for the existence of the variance–covariance matrix (Zellner, 1971). An implementation based on the t distribution will be discussed later.

Often, it is assumed that second-stage residuals are mutually independent across individuals. Then the joint density of all parameters at the second stage can be expressed as

$$p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M | \boldsymbol{\beta}, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M, \boldsymbol{\Sigma}_e\right) = \prod_{i=1}^{M} p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right). \tag{15.9}$$

Put $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_M'\right]'$ and $\mathbf{u} = [\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_M']'$. Under the normality assumption made in (15.7), the preceding takes the form

$$\begin{aligned} p\left(\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}_e\right) &\propto |\boldsymbol{\Sigma}_e|^{-\frac{M}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{M} \mathbf{e}_i' \boldsymbol{\Sigma}_e^{-1} \mathbf{e}_i\right] \\ &\propto |\boldsymbol{\Sigma}_e|^{-\frac{M}{2}} \exp\left[-\frac{1}{2}\, tr \sum_{i=1}^{M} \mathbf{e}_i' \boldsymbol{\Sigma}_e^{-1} \mathbf{e}_i\right] \\ &\propto |\boldsymbol{\Sigma}_e|^{-\frac{M}{2}} \exp\left[-\frac{1}{2}\, tr\left(\boldsymbol{\Sigma}_e^{-1}\mathbf{B}\right)\right], \end{aligned} \tag{15.10}$$

where $\mathbf{e}_i = \boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{u}_i$, as before, and $\mathbf{B} = \sum_{i=1}^{M} \mathbf{e}_i \mathbf{e}_i'$ is an $r \times r$ matrix. The diagonal elements of this matrix contain the sum of squared deviations of the appropriate parameter from their second-stage conditional expectations; the off-diagonals are sums of products of such parameter deviations. It must be noted that in many models, motivated by mechanistic considerations about growth and lactation, the values of the trajectory parameters must be defined within a restricted range. For example, it is clear that adult body weight or total milk yield produced cannot be negative. Such restrictions are easy to incorporate in the model via an appropriate definition of the parameter space. This issue will be discussed further later.

It is useful to note that if all parameters are concatenated vertically, the second stage structure in (15.6) can be represented as

$$\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_M \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_M \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_M \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_M \end{bmatrix}$$

or, more compactly, as

$$\boldsymbol{\theta}_{Mr\times 1} = \mathbf{X}_{Mr\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{u}_{Mr\times 1} + \mathbf{e}_{Mr\times 1}.$$

This indicates that the second-stage distribution of all parameters of all individuals is

$$\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}_e \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \mathbf{I} \otimes \boldsymbol{\Sigma}_e\right). \tag{15.11}$$

An alternative formulation can be obtained by arranging individuals within parameters; here $\mathbf{X}$ must be redefined accordingly, and the covariance matrix of the process would then be $\boldsymbol{\Sigma}_e \otimes \mathbf{I}$. The choice between the two alternative orders is entirely a matter of computational convenience.

Example 15.4 *Two-stage model for a quadratic trajectory*

Suppose that feed intake measurements are taken serially in each of a number of descendants of a random sample of boars to be evaluated in a progeny

test. Consider the second-order quadratic model of Example 15.1, and assume it provides a reasonable description of the trajectory of feed intake over time. Let y_{ijk} be measurement k taken in offspring j of boar i. The first-stage model can be written as

$$\begin{gathered} y_{ijk} = a_{ij} + b_{ij}t_{ijk} + c_{ij}t_{ijk}^2 + \varepsilon_{ijk}, \\ i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, o_i,\ k = 1, 2, \ldots, n_{ijk}, \\ y_{ijk}|a_{ij}, b_{ij}, c_{ij}, \sigma_\varepsilon^2 \sim NIID(a_{ij} + b_{ij}t_{ijk} + c_{ij}t_{ijk}^2, \sigma_\varepsilon^2), \end{gathered}$$

where a_{ij}, b_{ij}, and c_{ij} are coefficients peculiar to offspring j of boar i. As usual, ε_{ijk} is a discrepancy from the expected trajectory of individual ij when measured at time k. If the sample of boars is homogeneous in every possible respect, it might be sensible to fit the second stage model

$$\begin{bmatrix} a_{ij} \\ b_{ij} \\ c_{ij} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix} + \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} + \begin{bmatrix} e_{a_{ij}} \\ e_{b_{ij}} \\ e_{c_{ij}} \end{bmatrix},$$

where a_0, b_0, and c_0 are regression parameters common to all observations; a_i, b_i, and c_i are deviations common to all progeny of boar i, and the e's are the second-stage residuals, peculiar to offspring j of boar i. Here, $\mathbf{X}_{ij} = \mathbf{I}_3$, $\forall\, i, j$,

$$\mathbf{u}_i = \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_{e_a}^2 & \sigma_{e_{ab}} & \sigma_{e_{ac}} \\ \sigma_{e_{ab}} & \sigma_{e_b}^2 & \sigma_{e_{bc}} \\ \sigma_{e_{ac}} & \sigma_{e_{bc}} & \sigma_{e_c}^2 \end{bmatrix}$$

is the variance–covariance matrix between progeny-specific regression parameters, conditionally on a_i, b_i, and c_i.

It is instructive to write the first-stage model in terms of the second-stage structure, to obtain

$$\begin{aligned} y_{ijk} = &\left(a_0 + b_0 t_{ijk} + c_0 t_{ijk}^2\right) \\ &+ \left[a_i + b_i t_{ijk} + c_i t_{ijk}^2\right] \\ &+ \left\{e_{a_{ij}} + e_{b_{ij}} t_{ijk} + e_{c_{ij}} t_{ijk}^2\right\} + \varepsilon_{ijk}. \end{aligned}$$

The function in parentheses, $(\cdot)$, can be construed as a "population regression", common to all individuals measured; the second function, $[\cdot]$, is a deviation from the population regression shared by all descendants of boar i; and the function $\{\cdot\}$ is a deviation specific to offspring j of boar i. Conditionally on the parameters of the second-stage model, and assuming

normality throughout, it follows that

$$y_{ijk}|a_0, b_0, c_0, a_i, b_i, c_i, t_{ijk}, v\left(t_{ijk}\right) \sim$$
$$N\left[\left(a_0 + a_i\right) + \left(b_0 + b_i\right) t_{ijk} + \left(c_0 + c_i\right) t_{ijk}^2, v\left(t_{ijk}\right)\right],$$

where

$$v\left(t_{ijk}\right) = \left[\begin{array}{ccc} 1 & t_{ijk} & t_{ijk}^2 \end{array}\right] \left[\begin{array}{ccc} \sigma_{e_a}^2 & \sigma_{e_{ab}} & \sigma_{e_{ac}} \\ \sigma_{e_{ab}} & \sigma_{e_b}^2 & \sigma_{e_{bc}} \\ \sigma_{e_{ac}} & \sigma_{e_{bc}} & \sigma_{e_c}^2 \end{array}\right] \left[\begin{array}{c} 1 \\ t_{ijk} \\ t_{ijk}^2 \end{array}\right] + \sigma_\varepsilon^2$$

is a "variance function". Likewise, and conditionally on the second-stage parameters, the covariance between observations taken at times k and k' on individual ij is

$$Cov\left(y_{ijk}, y_{ijk'}|\text{second-stage parameters}\right) = \mathbf{t}_k' \left[\begin{array}{ccc} \sigma_{e_a}^2 & \sigma_{e_{ab}} & \sigma_{e_{ac}} \\ \sigma_{e_{ab}} & \sigma_{e_b}^2 & \sigma_{e_{bc}} \\ \sigma_{e_{ac}} & \sigma_{e_{bc}} & \sigma_{e_c}^2 \end{array}\right] \mathbf{t}_{k'},$$

where $\mathbf{t}_k' = \left[\begin{array}{ccc} 1 & t_{ijk} & t_{ijk}^2 \end{array}\right]$ and $\mathbf{t}_{k'}' = \left[\begin{array}{ccc} 1 & t_{ijk'} & t_{ijk'}^2 \end{array}\right]$. The corresponding "correlation function" is then

$$Corr\left(y_{ijk}, y_{ijk'}|\text{second stage parameters}\right) = \frac{\mathbf{t}_k' \boldsymbol{\Sigma}_e \mathbf{t}_{k'}}{\sqrt{\left(\mathbf{t}_k' \boldsymbol{\Sigma}_e \mathbf{t}_k\right)\left(\mathbf{t}_{k'}' \boldsymbol{\Sigma}_e \mathbf{t}_{k'}\right)}}.$$

Suppose, further, that the boar-specific parameters are assigned the distribution

$$\left[\begin{array}{c} a_i \\ b_i \\ c_i \end{array}\right] \Bigg| \boldsymbol{\Sigma}_s \sim N\left(\left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right], \boldsymbol{\Sigma}_s\right),$$

where

$$\boldsymbol{\Sigma}_s = \left[\begin{array}{ccc} \sigma_{s_a}^2 & \sigma_{s_{ab}} & \sigma_{s_{ac}} \\ \sigma_{s_{ab}} & \sigma_{s_b}^2 & \sigma_{s_{bc}} \\ \sigma_{s_{ac}} & \sigma_{s_{bc}} & \sigma_{s_c}^2 \end{array}\right]$$

is the covariance matrix between boar-specific deviations from the overall population regression. Then, the additional assumption that boar-specific and progeny-specific deviations are independent yields

$$y_{ijk}|a_0, b_0, c_0, t_{ijk}, \boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_s \sim N\left(a_0 + b_0 t_{ijk} + c_0 t_{ijk}^2, \mathbf{t}_k'\left(\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e\right)\mathbf{t}_k\right).$$

The distribution given above has a mean and variance describing how the trait and its variability change in time, averaged over all individuals in the population. ■

A special case is when the first and second stages of the model are both linear in the parameters. If the first stage is linear, (15.1) can be written as

$$\mathbf{y}_i = \mathbf{T}_i \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, ..., M, \tag{15.12}$$

for some known matrix $\mathbf{T}_i$. Employing (15.6) in (15.12) gives the representation

$$\begin{aligned}\mathbf{y}_i &= \mathbf{T}_i\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i\right) + \boldsymbol{\varepsilon}_i \\ &= \left(\mathbf{T}_i\mathbf{X}_i\right)\boldsymbol{\beta} + \mathbf{T}_i\mathbf{u}_i + \mathbf{T}_i\mathbf{e}_i + \boldsymbol{\varepsilon}_i \\ &= \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{T}_i\mathbf{u}_i + \mathbf{T}_i\mathbf{e}_i + \boldsymbol{\varepsilon}_i, \end{aligned} \tag{15.13}$$

where $\mathbf{X}_i^* = \mathbf{T}_i\mathbf{X}_i$. Under the usual normality assumptions, this induces the conditional distributions

$$\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \mathbf{e}_i, \boldsymbol{\gamma} \sim N\left[\mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{T}_i\mathbf{u}_i + \mathbf{T}_i\mathbf{e}_i, \mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right], \tag{15.14}$$

and

$$\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} \sim N\left[\mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{T}_i\mathbf{u}_i, \mathbf{T}_i\boldsymbol{\Sigma}_e\mathbf{T}_i' + \mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right]. \tag{15.15}$$

It follows from (15.13)-(15.15) that a two-stage linear hierarchy for longitudinal data is a special case of the general mixed effects linear model. Further deconditioning can be obtained by introducing additional tiers in the hierarchical structure.

The form of (15.12) implies that in a linear hierarchical model, once a functional form is adopted for the first stage, then all second-stage elements contribute to the overall model in a similar form; for example, the gradient of the observations with respect to either $\mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{u}_i$, or $\mathbf{e}_i$ is always $\mathbf{T}_i'$. It is possible to give a Bayesian implementation in a more general and flexible manner, but this is notationally awkward (specially if a nonlinear specification is chosen for the first stage). Hence, the hierarchical representation is kept, with the understanding that all subsequent developments apply to most models, at least conceptually.

15.2.3 Third Stage

In a Bayesian model, as pointed out in the chapters on Bayesian inference, prior distributions must be assigned to all unknown quantities in the statistical system posited. Thus, priors must be adopted for $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$.

Let the vector $\mathbf{u}$ represent additive genetic effects on the trajectory parameters. In this case, a classical (and convenient) assumption made in quantitative genetics is that

$$\mathbf{u}|\mathbf{G}_0 \sim N\left(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_0\right), \tag{15.16}$$

where it is implied that parameters are ordered within individuals. Above, $\mathbf{A}$ is the additive genetic relationship matrix between the M individuals, and $\mathbf{G}_0$ is an $r \times r$ additive genetic variance–covariance matrix between

parameters, that is

$$\mathbf{G}_0 = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_{12}} & \dots & \sigma_{u_{1r}} \\ \sigma_{u_{21}} & \sigma^2_{u_2} & \dots & \sigma_{u_{2r}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_{r1}} & \sigma_{u_{r2}} & \dots & \sigma^2_{u_r} \end{bmatrix},$$

with the understanding that $\sigma_{u_{ij}} = \sigma_{u_{ji}}$. If $\mathbf{G}_0$ is unknown, a prior distribution must be elicited for this matrix as well. For a linear model as in (15.15) the preceding prior implies that, given $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_e$, $\mathbf{G}_0$, and $\boldsymbol{\gamma}$, the prior predictive distribution is

$$\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \mathbf{G}_0, \boldsymbol{\gamma} \sim N\left[\mathbf{X}^*_i\boldsymbol{\beta}, \mathbf{T}_i\left(\mathbf{G}_0 + \boldsymbol{\Sigma}_e\right)\mathbf{T}'_i + \mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right].$$

It will be assumed further that

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma} \sim N\left(\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \tag{15.17}$$

where $\boldsymbol{\alpha}$, and $\boldsymbol{\Gamma}$ are known hyperparameters. The joint prior density of all unknowns can be taken to be equal to

$$p\left(\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) = p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u}|\mathbf{G}_0\right) p\left(\mathbf{G}_0\right) p\left(\boldsymbol{\Sigma}_e\right) p\left(\boldsymbol{\gamma}\right). \tag{15.18}$$

The preceding implies, a priori, that $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$ are mutually independent of each other and of $\mathbf{u}$ and $\mathbf{G}_0$, with the only dependence assumed a priori being that of the distribution of $\mathbf{u}$ on $\mathbf{G}_0$. As seen before, after data are combined with the prior via formal use of Bayes theorem, parameters become interdependent, even if the data set contains just a few observations. Vague priors can be adopted for the dispersion components, with the corresponding densities being

$$p\left(\mathbf{G}_0\right) \propto \text{constant}, \quad |\mathbf{G}_0| > 0, \tag{15.19}$$

$$p\left(\boldsymbol{\Sigma}_e\right) \propto \text{constant}, \quad |\boldsymbol{\Sigma}_e| > 0, \tag{15.20}$$

and

$$p\left(\boldsymbol{\gamma}\right) \propto \text{constant}, \quad \boldsymbol{\gamma} \in \Re_{\boldsymbol{\gamma}}, \tag{15.21}$$

where $\Re_{\boldsymbol{\gamma}}$ is the allowable parameter space of the dispersion vector $\boldsymbol{\gamma}$. For example, if this vector contains a single residual variance component, its parameter space would be the positive part of the real line, $\mathbb{R}_+$.

The preceding prior distributions may be bounded, based on either prior knowledge of parameter values or on theoretical considerations. It must be emphasized that an advantage of the Bayesian approach resides in the possibility of incorporating external stochastic information into the analysis. If such information exists, and if one wishes to use it, the prior densities should be modified accordingly. For instance, one may adopt informative

priors for the dispersion parameters, for example, inverted Wishart distributions for the covariance matrices and scaled inverted chi-square distributions for variance components. If the priors are conjugate, a Markov chain Monte Carlo implementation does not become much more difficult than one based on bounded uniform priors.

It is often convenient, especially in the case of nonlinear models, to augment the prior distribution with the trajectory parameters $\boldsymbol{\theta}$, leading to the joint prior

$$\begin{aligned} p\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} | \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) &= p\left(\boldsymbol{\theta} | \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \\ &\quad p\left(\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} | \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right). \end{aligned} \tag{15.22}$$

Now, because the parameters of different individuals, conditionally on $\boldsymbol{\beta}$ and $\mathbf{u}$, are independently distributed, with joint distribution as in (15.9), use of this and of the prior densities (15.18)–(15.21) in (15.22) leads to the following form for the augmented joint prior density

$$p\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} | \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \propto \prod_{i=1}^{M} p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right) p\left(\boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u} | \mathbf{G}_0\right). \tag{15.23}$$

The joint posterior distribution has support for any value of $\boldsymbol{\beta}$ in $\mathbb{R}^p$ and for any value of $\mathbf{u}_i$ in $\mathbb{R}^r$. However, as noted earlier, if trajectory parameters take values only within a restricted range, it may not always be reasonable assigning an r-variate normal distribution as prior for the $\boldsymbol{\theta}_i$ coefficients. Nevertheless, the normality assumption may hold well for a transformation of such parameters. For example, if a hypothetical curve for somatic cell count in milk (a measure of udder health in dairy cattle) has four parameters representing, for example, initial conditions, growth rate, change in growth rate, and level near the end of lactation, it may be sensible to adopt a parameterization in a log-scale, with the normal prior assigned to the ensuing parameterization. At any rate, this seldom causes serious problems, provided that the longitudinal series is "long enough" for most individuals (so the prior has a mild effect on inferences), and that the genetic relationship structure is sufficiently dense, so that information between related individuals can be exchanged. Alternative parameterizations do not complicate the Bayesian analysis conceptually, but can make implementation more involved.

15.2.4 Joint Posterior Distribution

From Bayes theorem, the joint posterior density of all unknowns is formed by combining the density of the sampling model in (15.5) with the joint

prior (15.23), yielding

$$\begin{aligned} &p\left(\boldsymbol{\theta}, \beta, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \\ &\propto \left\{\prod_{i=1}^{M}\left|\mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon}_i\right] p\left(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right)\right\} \\ &\qquad \times p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u}|\mathbf{G}_0\right). \end{aligned} \tag{15.24}$$

As a side issue, it is instructive to consider the density of the distribution of the observations unconditionally on $\boldsymbol{\beta}$ and $\mathbf{u}$, and this can be written explicitly when the trajectory is linear in the parameters. From (15.15), it follows that the resulting prior predictive distribution is:

$$\mathbf{y}_i|\boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{G}_0, \boldsymbol{\alpha}, \boldsymbol{\Gamma} \sim N\left[\mathbf{X}_i^*\boldsymbol{\alpha}, \mathbf{X}_i^*\Gamma\mathbf{X}_i^{*\prime} + \mathbf{T}_i\left(\mathbf{G}_0 + \boldsymbol{\Sigma}_e\right)\mathbf{T}_i' + \mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right]. \tag{15.25}$$

This distribution, given $\boldsymbol{\Sigma}_e$, $\boldsymbol{\gamma}$, $\mathbf{G}_0$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Gamma}$ can be interpreted as the probability distribution of the data before observations are collected.

15.3 Two-Step Approximate Bayesian Analysis

Bayesian inference always relies on applying the probability calculus to some target distribution, this being the joint posterior of all unknowns in (15.24). Probability statements are obtained from appropriate sets of marginal, joint, or conditional distributions, depending in the type of inference sought. However, in the case of the probability model with density given by (15.24), it is impossible to arrive at the marginal distributions of interest because the required integrals cannot be evaluated in closed form; furthermore, numerical quadrature seldom works well beyond a few dimensions. An alternative consists in extracting samples from the joint posterior, with the appropriate coordinate of the sample being a draw from the corresponding marginal distribution. From such samples, features of the posterior distribution of interest (e.g., mean, median, variance, quantiles or densities) can be estimated. A MCMC analysis can be used for this purpose, and this will be discussed later. Often, however, an approximate, simpler, analysis can lead to satisfactory inferences, at least for some features of the posterior distribution. For example, in animal breeding it is customary to carry out the following two-step analysis for mixed effects linear models: first, estimate dispersion parameters by some method, such as REML. Second, conditionally on the estimates of dispersion parameters, find point estimates and (sometimes) a measure of uncertainty for $\boldsymbol{\beta}$ and for $\mathbf{u}_i$ $(i = 1, 2, \ldots, M)$. A similar two-step analysis is described for the hierarchical model under discussion.

15.3.1 Estimating $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$ when Variances are Known

Suppose there is no uncertainty about the dispersion parameters, that is, $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$ are known without error. Recall from (15.3) that the first-stage residual vector can be expressed as

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t}) = \mathbf{y} - \mathbf{f}[\mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \mathbf{t}],$$

so that, for individual i,

$$\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t}_i) = \mathbf{y}_i - \mathbf{f}_i(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i, \mathbf{t}_i).$$

The joint posterior density of $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$, given $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$, can be written (without making use of augmentation with the $\boldsymbol{\theta}$ parameters) as

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}|\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) &\propto \prod_{i=1}^{M} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}_i\right] \\ &\quad p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma})\, p(\mathbf{u}|\mathbf{G}_0)\, p(\mathbf{e}|\boldsymbol{\Sigma}_e) \\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{M}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}_i + (\boldsymbol{\beta}-\boldsymbol{\alpha})'\Gamma^{-1}(\boldsymbol{\beta}-\boldsymbol{\alpha})\right]\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\mathbf{u}'\left(\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1}\right)\mathbf{u} + \sum_{i=1}^{M}\mathbf{e}_i'\boldsymbol{\Sigma}_e^{-1}\mathbf{e}_i\right]\right\}. \end{aligned} \tag{15.26}$$

An approximate Bayesian analysis, conditionally on the dispersion components, consists of approximating the joint posterior distribution with density in (15.26) by a Gaussian process having mean vector equal to the joint mode, and a variance–covariance matrix given by the inverse of the corresponding negative Hessian matrix. If the first-stage model is linear, the approximation is exact, as verified subsequently. In general, the modal vector needs to be calculated with an iterative algorithm (converging in one step for a linear trajectory). Considering that second derivatives are needed for completing inferences, the Newton–Raphson or scoring algorithms are natural candidates here.

Let $l(\boldsymbol{\beta}, \mathbf{u})$ be the logarithm of the joint posterior density (15.26). Apart from an additive constant

$$\begin{aligned} l(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}) &= -\frac{1}{2}\left[\sum_{i=1}^{M}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}_i + (\boldsymbol{\beta}-\boldsymbol{\alpha})'\Gamma^{-1}(\boldsymbol{\beta}-\boldsymbol{\alpha})\right] \\ &\quad -\frac{1}{2}\left[\mathbf{u}'\left(\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1}\right)\mathbf{u} + \sum_{i=1}^{M}\mathbf{e}_i'\boldsymbol{\Sigma}_e^{-1}\mathbf{e}_i\right] \\ &= -\frac{1}{2}\left[\boldsymbol{\varepsilon}'\mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon} + (\boldsymbol{\beta}-\boldsymbol{\alpha})'\Gamma^{-1}(\boldsymbol{\beta}-\boldsymbol{\alpha})\right] \\ &\quad -\frac{1}{2}\left[\mathbf{u}'\left(\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1}\right)\mathbf{u} + \mathbf{e}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{e}\right], \end{aligned} \tag{15.27}$$

with $\boldsymbol{\varepsilon}' = [\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_M']$, $\mathbf{R}^{-1}(\boldsymbol{\gamma}) = \mathbf{R}_1^{-1}(\boldsymbol{\gamma}) \oplus \mathbf{R}_2^{-1}(\boldsymbol{\gamma}) \oplus \cdots \oplus \mathbf{R}_M^{-1}(\boldsymbol{\gamma})$, and $\oplus$ is the direct-sum operator, denoting that $\mathbf{R}$ is a block-diagonal matrix with individual blocks being $\mathbf{R}_i(\boldsymbol{\gamma})$.

First Derivatives

Expressions for the first and second derivatives are facilitated by employing the chain rule of calculus. Observing matrix comformability, one can write

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}}\right]_{p\times 1} = \left[\frac{\partial \boldsymbol{\theta}'}{\partial \boldsymbol{\beta}}\right]\left[\frac{\partial \mathbf{f}'(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}\right]. \tag{15.28}$$

Similarly,

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{u}}\right]_{q\times 1} = \left[\frac{\partial \boldsymbol{\theta}'}{\partial \mathbf{u}}\right]\left[\frac{\partial \mathbf{f}'(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}\right], \tag{15.29}$$

and

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{e}}\right]_{rM\times 1} = \left[\frac{\partial \boldsymbol{\theta}'}{\partial \mathbf{e}}\right]\left[\frac{\partial \mathbf{f}'(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}\right]. \tag{15.30}$$

In (15.29) note that if one includes in $\mathbf{u}$ the additive genetic values of related individuals that lack longitudinal information, then $q > rM$. This augmentation is a routine practice in animal breeding, since it permits inferring the genetic worth of candidates for selection indirectly, via the additive genetic relationship matrix $\mathbf{A}$. Put

$$q = rM + r\overline{M} = r\left(M + \overline{M}\right),$$

where $\overline{M}$ is the number of individuals without measurements. It is convenient to write the second-stage model as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where

$$\mathbf{Z}_{rM\times r\left(M+\overline{M}\right)} = \left[\mathbf{I}_{rM\times rM}, \mathbf{0}_{rM\times r\overline{M}}\right].$$

The order of the vectors is dropped in the notation hereinafter. Now

$$\frac{\partial \boldsymbol{\theta}'}{\partial \boldsymbol{\beta}} = \frac{\partial\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}\right)'}{\partial \boldsymbol{\beta}} = \mathbf{X}', \tag{15.31}$$

$$\frac{\partial \boldsymbol{\theta}'}{\partial \mathbf{u}} = \frac{\partial\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}\right)'}{\partial \mathbf{u}} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, \tag{15.32}$$

$$\frac{\partial \boldsymbol{\theta}'}{\partial \mathbf{e}} = \frac{\partial\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}\right)'}{\partial \mathbf{e}} = \mathbf{I}, \tag{15.33}$$

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\boldsymbol{\varepsilon}}{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}\right] = -2\mathbf{R}^{-1}(\boldsymbol{\gamma})\left[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})\right], \tag{15.34}$$

and define

$$\mathbf{H}'(\boldsymbol{\theta}) = \left[\frac{\partial \mathbf{f}'(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\theta}}\right] = \left[\frac{\partial [\mathbf{f}_1'(\boldsymbol{\theta}_1, \mathbf{t}) \;\; \mathbf{f}_2'(\boldsymbol{\theta}_2, \mathbf{t}) \;\; \ldots \;\; \mathbf{f}_M'(\boldsymbol{\theta}_M, \mathbf{t})]}{\partial \boldsymbol{\theta}}\right]$$
$$= \begin{bmatrix} \mathbf{H}_1'(\boldsymbol{\theta}_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2'(\boldsymbol{\theta}_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{H}_M'(\boldsymbol{\theta}_M) \end{bmatrix}, \tag{15.35}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_M]'$. Above

$$\mathbf{H}_i'(\boldsymbol{\theta}_i) = \frac{\partial \mathbf{f}_i'(\boldsymbol{\theta}_i, \mathbf{t})}{\partial \boldsymbol{\theta}_i}$$

is an $r \times n_i$ matrix, being independent of $\boldsymbol{\theta}_i$ only when the first-stage model is linear in the parameters. Hence

$$\mathbf{H}'(\boldsymbol{\theta}) = \mathbf{H}_1'(\boldsymbol{\theta}) \oplus \mathbf{H}_2'(\boldsymbol{\theta}) \oplus \cdots \oplus \mathbf{H}_M'(\boldsymbol{\theta}).$$

Applying (15.31)–(15.35) in (15.28), (15.29) and (15.30):

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}}\right] = -2\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma}) [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})],$$

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon}}{\partial \mathbf{u}}\right] = -2 \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma}) [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})]$$
$$= -2 \begin{bmatrix} \mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma}) [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})] \\ \mathbf{0} \end{bmatrix},$$

and

$$\left[\frac{\partial \boldsymbol{\varepsilon}' \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon}}{\partial \mathbf{e}}\right] = -2\mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma}) [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})].$$

The gradient vector of the logarithm of the conditional posterior density of $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$ in (15.27) is, using the preceding expressions,

$$\frac{\partial L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon} - \Gamma^{-1}(\boldsymbol{\beta} - \boldsymbol{\alpha}), \tag{15.36}$$

$$\frac{\partial L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \mathbf{u}} = \begin{bmatrix} \mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon} \\ \mathbf{0} \end{bmatrix} - \left(\mathbf{A}^{-1} \otimes \mathbf{G}_0^{-1}\right) \mathbf{u}, \tag{15.37}$$

$$\frac{\partial L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \mathbf{e}} = \mathbf{H}'(\boldsymbol{\theta})\, \mathbf{R}^{-1}(\boldsymbol{\gamma})\, \boldsymbol{\varepsilon} - \left(\mathbf{I} \otimes \boldsymbol{\Sigma}_e^{-1}\right) \mathbf{e}, \tag{15.38}$$

with $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})$.

Second Derivatives

The Newton–Raphson or scoring algorithms, and the Gaussian approximation to the posterior distribution, require second derivatives. Hence, an additional differentiation is needed. For simplicity of notation let, hereinafter,

$$\mathbf{R}^{-1}(\boldsymbol{\gamma}) = \mathbf{R}^{-1}.$$

Now, note that

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} &= \frac{\partial \left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\boldsymbol{\varepsilon} - \Gamma^{-1}(\boldsymbol{\beta} - \boldsymbol{\alpha})\right]}{\partial \boldsymbol{\beta}'} \\ &= \frac{\partial \left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\boldsymbol{\varepsilon}\right]}{\partial \boldsymbol{\beta}'} - \Gamma^{-1} \\ &= \mathbf{X}'\left[\frac{\partial \mathbf{H}'(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}'}\right]\mathbf{R}^{-1}\boldsymbol{\varepsilon} - \mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\frac{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\beta}'} - \Gamma^{-1}.\end{aligned}$$

The expression $[\partial \mathbf{H}'(\boldsymbol{\theta})/\partial \boldsymbol{\beta}']$ is an informal representation, as the derivatives of a matrix with respect to a vector require a "three-dimensional matrix" (technically, one needs to arrange elements of $\mathbf{H}'(\boldsymbol{\theta})$ into a vector, and then take derivatives). Ignoring this, note that if expectations of the second derivatives are taken with respect to the first-stage distribution, given in (15.4), one has

$$E_{\mathbf{y}|\boldsymbol{\theta}}\left[\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})\right] = \mathbf{0}.$$

Hence, considerable simplification is obtained, the result being

$$\begin{aligned}E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right] &= -\left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\frac{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\beta}'} + \Gamma^{-1}\right] \\ &= -\left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\frac{\partial \mathbf{f}(\boldsymbol{\theta}, \mathbf{t})}{\partial \boldsymbol{\theta}'}\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'} + \Gamma^{-1}\right] \\ &= -\left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\mathbf{X} + \Gamma^{-1}\right].\end{aligned} \tag{15.39}$$

Likewise,

$$\begin{aligned}E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e})}{\partial \mathbf{u}\, \partial \mathbf{u}'}\right] &= -\left[\mathbf{Z}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\mathbf{Z} + \mathbf{A}^{-1} \otimes \mathbf{G}_0^{-1}\right] \\ = -\left\{\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\left[\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\right]\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix} + \mathbf{A}^{-1} \otimes \mathbf{G}_0^{-1}\right\}.\end{aligned} \tag{15.40}$$

Now the additive genetic relationship matrix $\mathbf{A}$ can be partitioned in a form consistent with that of the vector $\mathbf{u}$, such that

$$\begin{aligned}\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1} &= \begin{bmatrix}\mathbf{A}_{MM} & \mathbf{A}_{M\overline{M}}\\ \mathbf{A}_{\overline{M}M} & \mathbf{A}_{\overline{M}\overline{M}}\end{bmatrix}^{-1}\otimes\mathbf{G}_0^{-1}\\ &= \begin{bmatrix}\mathbf{A}^{MM}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{M\overline{M}}\otimes\mathbf{G}_0^{-1}\\ \mathbf{A}^{\overline{M}M}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{\overline{M}\overline{M}}\otimes\mathbf{G}_0^{-1}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{G}^{MM} & \mathbf{G}^{M\overline{M}}\\ \mathbf{G}^{\overline{M}M} & \mathbf{G}^{\overline{M}\overline{M}}\end{bmatrix}.\end{aligned}$$

Using this in (15.40)

$$E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta},\mathbf{u},\mathbf{e})}{\partial\mathbf{u}\partial\mathbf{u}'}\right] = -\begin{bmatrix}\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})+\mathbf{G}^{MM} & \mathbf{G}^{M\overline{M}}\\ \mathbf{G}^{\overline{M}M} & \mathbf{G}^{\overline{M}\overline{M}}\end{bmatrix}. \tag{15.41}$$

Further,

$$E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta},\mathbf{u},\mathbf{e})}{\partial\mathbf{e}\partial\mathbf{e}'}\right] = -\left[\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})+\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right]. \tag{15.42}$$

Using similar algebra, the second-order mixed partial derivatives are

$$\begin{aligned}E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta},\mathbf{u},\mathbf{e})}{\partial\boldsymbol{\beta}\partial\mathbf{u}'}\right] &= -\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\mathbf{Z}\\ &= -\begin{bmatrix}\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta}) & \mathbf{0}\end{bmatrix},\end{aligned} \tag{15.43}$$

$$E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta},\mathbf{u},\mathbf{e})}{\partial\boldsymbol{\beta}\partial\mathbf{e}'}\right] = -\left[\mathbf{X}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\right], \tag{15.44}$$

and

$$\begin{aligned}E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2 L(\boldsymbol{\beta},\mathbf{u},\mathbf{e})}{\partial\mathbf{u}\partial\mathbf{e}'}\right] &= -\left[\mathbf{Z}'\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\right]\\ &= -\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\\ &= -\begin{bmatrix}\mathbf{H}'(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\theta})\\ \mathbf{0}\end{bmatrix}.\end{aligned} \tag{15.45}$$

Gaussian Approximation to the Conditional Posterior Distribution

Using results given in the chapter on approximate methods for inference, the conditional posterior distribution

$$[\boldsymbol{\beta},\mathbf{u},\mathbf{e}|\mathbf{G}_0,\boldsymbol{\Sigma}_e,\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\Gamma}]$$

can be approximated as

$$\boldsymbol{\beta}, \mathbf{u}, \mathbf{e} | \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma} \sim N(\mathbf{c}, \mathbf{C}_\mathbf{c}^{-1}), \tag{15.46}$$

where

$$\mathbf{c} = Arg \max_{\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}} p\left(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e} | \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \tag{15.47}$$

is the modal vector, and $\mathbf{C}_\mathbf{c}$ is the negative of the matrix of second derivatives of $l\left(\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}\right)$ with respect to $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{e}$, evaluated at the modal value $\mathbf{c}$. Collecting the submatrices of second derivatives in (15.39), and (15.41)–(15.45), the symmetric negative Hessian matrix can be expressed as

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} + \Gamma^{-1} & \mathbf{X}'\mathbf{W} & \mathbf{0} & \mathbf{X}'\mathbf{W} \\ . & \mathbf{W} + \mathbf{G}^{MM} & \mathbf{G}^{M\overline{M}} & \mathbf{W} \\ . & . & \mathbf{G}^{\overline{M}M} & \mathbf{0} \\ . & . & . & \mathbf{W} + \mathbf{I} \otimes \boldsymbol{\Sigma}_e^{-1} \end{bmatrix}, \tag{15.48}$$

where

$$\mathbf{W} = \mathbf{W}\left(\boldsymbol{\theta}, \boldsymbol{\gamma}\right) = \mathbf{H}'\left(\boldsymbol{\theta}\right) \mathbf{R}^{-1} \mathbf{H}\left(\boldsymbol{\theta}\right),$$

noting that this is a symmetric matrix and with order $rM \times rM$.

The Scoring Algorithm

Computation of the modal vector $\mathbf{c}$ using the scoring algorithm proceeds with the iteration

$$\mathbf{c}_{[t+1]} = \mathbf{c}_{[t]} + \mathbf{C}_{[t]}^{-1} \mathbf{g}_{[t]}, \tag{15.49}$$

where $[t]$ denotes the iterate number and $\mathbf{g}_{[t]}$ is the gradient vector at iteration $[t]$. An alternative representation is

$$\mathbf{C}_{[t]} \mathbf{c}_{[t+1]} = \mathbf{C}_{[t]} \mathbf{c}_{[t]} + \mathbf{g}_{[t]}. \tag{15.50}$$

Making use of (15.36)–(15.38), and putting $\boldsymbol{\varepsilon}^{[t]} = \mathbf{y} - \mathbf{f}\left(\boldsymbol{\theta}^{[t]}, \mathbf{t}\right)$,

$$\mathbf{g}_{[t]} = \begin{bmatrix} \mathbf{X}'\mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right) \mathbf{R}^{-1} \boldsymbol{\varepsilon}^{[t]} - \Gamma^{-1}\left(\boldsymbol{\beta}^{[t]} - \boldsymbol{\alpha}\right) \\ \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right) \mathbf{R}^{-1} \boldsymbol{\varepsilon}^{[t]} - \mathbf{G}^{MM} \mathbf{u}_M^{[t]} - \mathbf{G}^{M\overline{M}} \mathbf{u}_{\overline{M}}^{[t]} \\ -\mathbf{G}^{\overline{M}M} \mathbf{u}_M^{[t]} - \mathbf{G}^{\overline{M}\overline{M}} \mathbf{u}_{\overline{M}}^{[t]} \\ \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right) \mathbf{R}^{-1} \boldsymbol{\varepsilon}^{[t]} - \left(\mathbf{I} \otimes \boldsymbol{\Sigma}_e^{-1}\right) \mathbf{e}^{[t]} \end{bmatrix}.$$

Note that

$$\mathbf{C}_{[t]} \mathbf{c}_{[t]} = \begin{bmatrix} \mathbf{X}'\mathbf{W}^{[t]} \left[\mathbf{X}\boldsymbol{\beta}^{[t]} + \mathbf{u}_M^{[t]} + \mathbf{e}^{[t]}\right] + \Gamma^{-1}\boldsymbol{\beta}^{[t]} \\ \mathbf{W}^{[t]} \left[\mathbf{X}\boldsymbol{\beta}^{[t]} + \mathbf{u}_M^{[t]} + \mathbf{e}^{[t]}\right] + \mathbf{G}^{MM} \mathbf{u}_M^{[t]} + \mathbf{G}^{M\overline{M}} \mathbf{u}_{\overline{M}}^{[t]} \\ \mathbf{G}^{\overline{M}M} \mathbf{u}_M^{[t]} + \mathbf{G}^{M\overline{M}} \mathbf{u}_{\overline{M}}^{[t]} \\ \mathbf{W}^{[t]} \left[\mathbf{X}\boldsymbol{\beta}^{[t]} + \mathbf{u}_M^{[t]} + \mathbf{e}^{[t]}\right] + \left(\mathbf{I} \otimes \boldsymbol{\Sigma}_e^{-1}\right) \mathbf{e}^{[t]} \end{bmatrix}.$$

Since

$$\boldsymbol{\theta}^{[t]} = \mathbf{X}\boldsymbol{\beta}^{[t]} + \mathbf{u}_M^{[t]} + \mathbf{e}^{[t]},$$

adding the two preceding vectors gives

$$\mathbf{C}_{[t]}\mathbf{c}_{[t]} + \mathbf{g}_{[t]} = \begin{bmatrix} \mathbf{X}'\mathbf{W}^{[t]}\boldsymbol{\theta}^{[t]} + \mathbf{X}'\mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\boldsymbol{\varepsilon}^{[t]} + \Gamma^{-1}\boldsymbol{\alpha} \\ \mathbf{W}^{[t]}\boldsymbol{\theta}^{[t]} + \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\boldsymbol{\varepsilon}^{[t]} \\ \mathbf{0} \\ \mathbf{W}^{[t]}\boldsymbol{\theta}^{[t]} + \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\boldsymbol{\varepsilon}^{[t]} \end{bmatrix}. \quad (15.51)$$

In (15.51), observe that

$$\begin{aligned} &\mathbf{W}^{[\mathbf{t}]}\boldsymbol{\theta}^{[t]} + \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\boldsymbol{\varepsilon}^{[t]} \\ &= \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\mathbf{H}\left(\boldsymbol{\theta}^{[t]}\right)\boldsymbol{\theta}^{[t]} + \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\boldsymbol{\varepsilon}^{[t]} \\ &= \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\left\{\mathbf{H}\left(\boldsymbol{\theta}^{[t]}\right)\boldsymbol{\theta}^{[t]} + \boldsymbol{\varepsilon}^{[t]}\right\} \\ &= \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\widehat{\mathbf{y}}^{[t]}, \end{aligned}$$

where

$$\widehat{\mathbf{y}}^{[t]} = \mathbf{H}\left(\boldsymbol{\theta}^{[t]}\right)\boldsymbol{\theta}^{[t]} + \boldsymbol{\varepsilon}^{[t]} \quad (15.52)$$

is a "pseudo-data" vector that changes values iteratively. Collecting (15.48), (15.51), and (15.52) to form the scoring algorithm, the iteration can be represented as

$$\begin{bmatrix} \boldsymbol{\beta}^{[t+1]} \\ \mathbf{u}_M^{[t+1]} \\ \mathbf{u}_{\overline{M}}^{[t+1]} \\ \mathbf{e}^{[t+1]} \end{bmatrix} = \left(\mathbf{C}^{[t]}\right)^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{d}^{[t]} + \Gamma^{-1}\boldsymbol{\alpha} \\ \mathbf{H}'\mathbf{d}^{[t]} \\ \mathbf{0} \\ \mathbf{d}^{[t]} \end{bmatrix}, \quad (15.53)$$

where

$$\mathbf{d}^{[t]} = \mathbf{H}'\left(\boldsymbol{\theta}^{[t]}\right)\mathbf{R}^{-1}\widehat{\mathbf{y}}^{[t]}.$$

This is in the form of an iteratively reweighted system of linear mixed model equations for a Gaussian process. The coefficient matrix and the vector of the right-hand sides change from iterate to iterate. If the scoring algorithm converges to a global maximum (this being extremely difficult or impossible to check), we denote the maximum a posteriori estimates of $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$ as $\widehat{\boldsymbol{\beta}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, $\widehat{\mathbf{u}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, and $\widehat{\mathbf{e}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, respectively, to emphasize the dependence on the dispersion parameters.

15.3.2 Estimating $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\gamma}$ from Their Marginal Posterior Distribution

In the preceding, methods for inferring unknowns from the conditional posterior distribution

$$[\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}|\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}]$$

were outlined, using Gaussian approximation (15.46). This requires knowledge of the nuisance parameters $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\gamma}$, a situation seldom encountered in practice. In approximate Bayesian analysis (e.g., Box and Tiao, 1973; Gianola and Fernando, 1986), the usual approach consists of finding the maximizers of the marginal posterior distribution $[\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}]$ of the dispersion components, say

$$\widehat{\mathbf{G}}_0, \widehat{\boldsymbol{\Sigma}}_e, \widehat{\boldsymbol{\gamma}} = Arg \max_{\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}} p\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right). \tag{15.54}$$

If uniform priors are adopted for $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$, as in the developments presented here, $\widehat{\mathbf{G}}_0$, $\widehat{\boldsymbol{\Sigma}}_e$ and $\widehat{\boldsymbol{\gamma}}$ are usually known as "marginal ML estimates". Subsequently, these estimates can be used to obtain inferences based on the distribution

$$\left[\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}|\mathbf{G}_0 = \widehat{\mathbf{G}}_0, \boldsymbol{\Sigma}_e = \widehat{\boldsymbol{\Sigma}}_e, \boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right].$$

It must be emphasized that such an analysis does not take into account the error of estimation of the nuisance parameters. The main difficulty, specially for a nonlinear first-stage model, is that $p\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right)$ cannot be arrived at in closed form, and the marginal ML estimates cannot be written explicitly. However, these estimates can be approximated as described below.

Consider again the first-stage model in (15.3):

$$\mathbf{y} = \mathbf{f}\left(\boldsymbol{\theta}, \mathbf{t}\right) + \boldsymbol{\varepsilon}$$

and expand its structure $\mathbf{f}\left(\boldsymbol{\theta}, \mathbf{t}\right)$ about provisional estimates

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right), \\ \widehat{\mathbf{u}} &= \widehat{\mathbf{u}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right), \end{aligned}$$

and

$$\widehat{\mathbf{e}} = \widehat{\mathbf{e}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right),$$

obtained as in the preceding section for some sensible starting values of the dispersion parameters. Putting $\boldsymbol{\eta} = \left[\boldsymbol{\beta}', \mathbf{u}', \mathbf{e}'\right]'$, a first order Taylor series approximation about $\widehat{\boldsymbol{\eta}}$ yields

$$\mathbf{y} \approx \mathbf{f}\left[\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}} + \widehat{\mathbf{e}}, \mathbf{t}\right] + \left\{\left[\frac{\partial \mathbf{f}\left(\boldsymbol{\theta}, \mathbf{t}\right)}{\partial \boldsymbol{\theta}'}\right]\left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}'}\right]\right\}_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) + \boldsymbol{\varepsilon}$$

$$= \mathbf{f}\left[\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}} + \widehat{\mathbf{e}}, \mathbf{t}\right] + \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right)\left[\begin{array}{ccc} \mathbf{X} & \mathbf{Z} & \mathbf{I} \end{array}\right]\left[\begin{array}{c} \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \\ \mathbf{u} - \widehat{\mathbf{u}} \\ \mathbf{e} - \widehat{\mathbf{e}} \end{array}\right] + \boldsymbol{\varepsilon}$$

$$= \mathbf{f}\left[\widehat{\boldsymbol{\theta}}, \mathbf{t}\right] - \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right)\left(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}} + \widehat{\mathbf{e}}\right) + \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right)(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}) + \boldsymbol{\varepsilon}. \quad (15.55)$$

In the preceding, now let

$$\begin{aligned} \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{X} &= \widehat{\mathbf{X}}, \\ \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{Z} &= \widehat{\mathbf{Z}}, \\ \mathbf{H}\left(\widehat{\boldsymbol{\theta}}\right) &= \widehat{\mathbf{H}}, \end{aligned}$$

and

$$\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{f}\left[\widehat{\boldsymbol{\theta}}, \mathbf{t}\right].$$

Recalling that the pseudo-data vector is

$$\widehat{\mathbf{y}} = \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{Z}}\widehat{\mathbf{u}} + \widehat{\mathbf{H}}\widehat{\mathbf{e}} + \widehat{\boldsymbol{\varepsilon}},$$

then, (15.55) can be rearranged as

$$\widehat{\mathbf{y}} \approx \widehat{\mathbf{X}}\boldsymbol{\beta} + \widehat{\mathbf{Z}}\mathbf{u} + \widehat{\mathbf{H}}\mathbf{e} + \boldsymbol{\varepsilon}, \quad (15.56)$$

so the expansion leads to a linear "pseudo-model" for "pseudo-data" from which an update for the dispersion parameters $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$ can be obtained by some standard method suitable for linear models, such as those based on maximizing likelihoods. With this update, a new point estimate of $\boldsymbol{\theta}$ can be computed with the scoring algorithm (15.53).

Under the linear pseudo-model, one can adopt the Bayesian hierarchy

$$\widehat{\mathbf{y}}|\boldsymbol{\beta}, \mathbf{u}, \widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \boldsymbol{\gamma} \approx N\left(\widehat{\mathbf{X}}\boldsymbol{\beta} + \widehat{\mathbf{Z}}\mathbf{u} + \widehat{\mathbf{H}}\mathbf{e}, \mathbf{R}\left(\boldsymbol{\gamma}\right)\right) \quad (15.57)$$

with priors as in (15.16)–(15.21). If flat, improper priors are adopted for $\boldsymbol{\beta}$ and for the dispersion parameters, any available algorithm for REML estimation of dispersion parameters can be used to obtain the revised values for $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\gamma}$. This is because with such flat priors, the REML estimates are the modal values of the distribution $[\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\widehat{\mathbf{y}}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}]$ in a Gaussian linear model (Harville, 1974). Such algorithms do not consider the situation where it is assumed that $\boldsymbol{\beta}$ has the informative distribution in (15.17), so this would work only if $p(\boldsymbol{\beta})$ is taken to be proportional to a constant. In this case, one proceeds to iterate between the REML algorithm applied to $\widehat{\mathbf{y}}$, to obtain new values of $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, $\boldsymbol{\gamma}$, the scoring algorithm (15.53) to obtain new values $\widehat{\boldsymbol{\beta}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, $\widehat{\mathbf{u}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, and $\widehat{\mathbf{e}}\left(\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\right)$, and the Taylor series expansion leading to the linear pseudo-model for $\widehat{\mathbf{y}}$ as in (15.56) and (15.57). The cycle is repeated until $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e, \boldsymbol{\gamma}$

stabilize, these being the modal values sought. Finally, inferences about $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{e}$ are obtained from the Gaussian approximation (15.46). In particular, since

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} = \left[\begin{array}{ccc} \mathbf{X} & \mathbf{Z} & \mathbf{I} \end{array} \right] \left[\begin{array}{c} \boldsymbol{\beta} \\ \mathbf{u} \\ \mathbf{e} \end{array} \right] = \mathbf{K} \left[\begin{array}{c} \boldsymbol{\beta} \\ \mathbf{u} \\ \mathbf{e} \end{array} \right],$$

so $\boldsymbol{\theta}$ is a linear combination of $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{e}$, then approximately,

$$\boldsymbol{\theta}|\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma} \sim N(\mathbf{K}\mathbf{c}, \mathbf{K}\mathbf{C}_\mathbf{c}^{-1}\mathbf{K}'), \tag{15.58}$$

with the understanding that if an REML algorithm is employed, then there is the implicit assumption that $\boldsymbol{\Gamma}^{-1} \to \mathbf{0}$, to make the prior distribution of $\boldsymbol{\beta}$ flat (and improper).

An alternative approach to locating the modal values of the dispersion parameters of $\mathbf{G}_0$, $\boldsymbol{\Sigma}_e$, $\boldsymbol{\gamma}$ would be using a Laplacian approximation, but this is not discussed here.

15.3.3 Special Case: Linear First Stage

If the trajectory is linear in the parameters, as in (15.12), the entire vector of observations can be written as

$$\mathbf{y} = \mathbf{T}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{15.59}$$

for $\mathbf{T} = \mathbf{T}_1 \oplus \mathbf{T}_2 \oplus \cdots \oplus \mathbf{T}_M$, where $\mathbf{T}_i$ is a known matrix, and

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

as before. For this linear model, the matrix of derivatives of the conditional expectation vector with respect to $\boldsymbol{\theta}$ is

$$\mathbf{H}'(\boldsymbol{\theta}) = \frac{\partial\left(\boldsymbol{\theta}'\mathbf{T}'\right)}{\partial\boldsymbol{\theta}} = \mathbf{T}'$$

which is independent of the parameters. Also, note that the pseudo-data vector $\widehat{\mathbf{y}}$ becomes

$$\begin{aligned} \widehat{\mathbf{y}}^{[t]} &= \mathbf{H}\left(\boldsymbol{\theta}^{[t]}\right)\boldsymbol{\theta}^{[t]} + \left[\mathbf{y} - \mathbf{f}\left(\boldsymbol{\theta}^{[t]}, \mathbf{t}\right)\right] \\ &= \mathbf{T}\boldsymbol{\theta}^{[t]} + \left[\mathbf{y} - \mathbf{T}\boldsymbol{\theta}^{[t]}\right] = \mathbf{y}, \end{aligned}$$

so it is identical to the vector of observations $\mathbf{y}$. In this situation, as seen in Chapter 6, the posterior distribution

$$[\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}|\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}]$$

is exactly Gaussian, with mean vector $\mathbf{m} = \mathbf{M}^{-1}\mathbf{r}$, and posterior variance–covariance matrix $\mathbf{M}^{-1}$, where

$$\mathbf{M} = \begin{bmatrix} \mathbf{X}'\mathbf{Q}\mathbf{X} + \Gamma^{-1} & \mathbf{X}'\mathbf{Q} & \mathbf{0} & \mathbf{X}'\mathbf{Q} \\ . & \mathbf{Q} + \mathbf{G}^{MM} & \mathbf{G}^{M\overline{M}} & \mathbf{Q} \\ . & . & \mathbf{G}^{\overline{M}\overline{M}} & \mathbf{0} \\ . & . & . & \mathbf{Q} + \mathbf{I} \otimes \boldsymbol{\Sigma}_e^{-1} \end{bmatrix}, \quad (15.60)$$

for $\mathbf{Q} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{T}$, and

$$\mathbf{r} = \begin{bmatrix} \mathbf{X}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y} + \Gamma^{-1}\boldsymbol{\alpha} \\ \mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \\ \mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \quad (15.61)$$

If a diffuse prior for $\boldsymbol{\beta}$ is adopted, such that $\boldsymbol{\Gamma}^{-1} \to \mathbf{0}$, the $\boldsymbol{\beta}$ component of the mean vector tends toward the best linear unbiased estimator of $\boldsymbol{\beta}$, whereas the $\mathbf{u}$ and $\mathbf{e}$ components tend to the BLUP of $\mathbf{u}$ and of $\mathbf{e}$, respectively. Further, in the approximate first-stage distribution in (15.57), one has that

$$\widehat{\mathbf{X}}\boldsymbol{\beta} + \widehat{\mathbf{Z}}\mathbf{u} + \widehat{\mathbf{H}}\mathbf{e} = \mathbf{T}\mathbf{X}\boldsymbol{\beta} + \mathbf{T}\mathbf{Z}\mathbf{u} + \mathbf{T}\mathbf{e} = \mathbf{T}\boldsymbol{\theta}.$$

Thus:

$$\begin{aligned} \widehat{\mathbf{y}}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}, \widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \boldsymbol{\gamma} &\approx N\left(\widehat{\mathbf{X}}\boldsymbol{\beta} + \widehat{\mathbf{Z}}\mathbf{u} + \widehat{\mathbf{H}}\mathbf{e}, \mathbf{R}\right) \\ &\equiv \mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{e}, \mathbf{X}, \mathbf{Z}, \mathbf{T}, \boldsymbol{\gamma} \sim N\left[\mathbf{T}\boldsymbol{\theta}, \mathbf{R}\right] \end{aligned} \quad (15.62)$$

which is exactly the first-stage distribution. Therefore, as $\Gamma^{-1} \to \mathbf{0}$, the mode of the posterior distribution $[\mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}]$ tends to the REML estimates, provided the prior for $\mathbf{u}$ is as in (15.16).

15.4 Computation via Markov Chain Monte Carlo

The approximate Bayesian method described in the preceding section is, undoubtedly, computationally involved, specially for nonlinear trajectories. However, it is statistically and algorithmically equivalent to what might be termed a "quasi-BLUP coupled with REML" analysis of a nonlinear mixed effects model. Actually, the latter represents the state of the art from a likelihood-frequentist perspective (e.g., Wolfinger, 1993; Wolfinger and Lin, 1997), and relies on asymptotic arguments as well. The approximate Bayesian approach, however, has the additional flexibility conferred by the possibility of introducing prior information, as noted earlier. It cannot be overemphasized, however, that in the approximate analysis:

(1) the estimates of the dispersion parameters $\mathbf{G}_0$, $\mathbf{\Sigma}_e$, $\boldsymbol{\gamma}$ are from a joint mode;
(2) inferences obtained for $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{e}$ or $\boldsymbol{\theta}$ are conditional on such modal values, and
(3) an asymptotic approximation to a conditional posterior distribution, as in (15.46), is used.

An alternative, and probably simpler from a computational point of view, is to carry out a fully Bayesian analysis by sampling from marginal posterior distributions of interest. This must be contrasted at the onset with the conditioning arguments employed in the approximate method. These reveal that uncertainty about nuisance parameters that one should integrate out is not taken into account. Although procedures for sampling from the posterior probably require more computer time, these are easier to program. In addition, an entire distribution can be estimated, as opposed to just its location and (perhaps) dispersion.

An implementation, based possibly on a combination of several sampling techniques, is discussed in this section since there is seldom a unique, optimal approach for drawing samples from posterior distributions. The starting point is to consider constructing a Gibbs sampler, that is, attempting to form (and identify) all possible fully conditional posterior distributions, and then looping through all such conditionals by iterative updating of the conditioning unknowns. As noted in Chapter 11, a complete pass or scan through all fully conditional distributions defines an iteration of the Gibbs sampler, at least in its best-known form. The scan can be done in a fixed or randomized order, or in an "up and down" direction, that is, when the order of updating in the iteration is exactly the reverse of that in the preceding one. Also, some "sites" (conditional distributions) may not be visited at all in a given iteration, although all sites must be updated, eventually, and visited an infinite number of times, theoretically. For example, in a statistical model with five unknowns, a scan pattern of 1, 2, 3, 1, 2, 3, 1, 2, 3, 4, 5 may be repeated indefinitely (Neal, personal communication). One might do this if variables 4 and 5 are thought to be almost independent of the others, and mix faster, so that it is better to spend more time working on 1, 2, 3. For example, a suitable reparameterization of the trajectory parameters or of the location vector $\boldsymbol{\beta}$, may enhance orthogonality, so one can then spend more time updating the dispersion parameters and the $\mathbf{u}$ and $\mathbf{e}$ vectors. In genetic applications, these two vectors may be highly colinear, and may not be identifiable distinctly from each other when the additive genetic relationship matrix ($\mathbf{A}$) is nearly an identity matrix and $\mathbf{G}_0$ and $\mathbf{\Sigma}_e$ are unknown. On the other hand, when dispersion parameters are known, $\mathbf{u}$ and $\mathbf{e}$ have distinct conditional posterior distributions, although strong posterior inter-correlations still may exist. The implementation described here operates as follows: if the fully conditional distribution of a scalar or vector can be identified, the corresponding sample is drawn directly, as in

standard Gibbs sampling. Otherwise, a Metropolis–Hastings, acceptance–rejection or importance sampling (with resampling) step can be adopted.

15.4.1 Fully Conditional Posterior Distributions

The starting point for identifying the needed conditional posterior distributions is the augmented joint posterior density (15.24). This is restated here to facilitate references to the expression in subsequent developments. The joint density is

$$\begin{aligned} &p\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma} | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \\ &\propto \left\{\prod_{i=1}^{M}\left|\mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i' \mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon}_i\right] p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right)\right\} \\ &\qquad p\left(\boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u} | \mathbf{G}_0\right). \end{aligned} \tag{15.63}$$

The conditional posterior densities can be deduced by retaining the part of (15.63) that is a function of the pertinent unknown, and treating the remaining portion as fixed, becoming, thus, a part of the integration constant of the conditional distribution sought. Such distributions will be examined systematically in what follows.

Trajectory Parameters

From (15.63):

$$\begin{aligned} &p\left(\boldsymbol{\theta} | \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \\ &\propto \left\{\prod_{i=1}^{M} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i' \mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon}_i\right] p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right)\right\}. \end{aligned} \tag{15.64}$$

This indicates that, given all other parameters, the trajectory coefficients $\boldsymbol{\theta}_i$ of different individuals are mutually independent of each other and, hence, can be sampled independently. Thus

$$\begin{aligned} p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) &= p\left(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}_i\right) \\ &\propto \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i - \mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}\right)\right]' \mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\left[\mathbf{y}_i - \mathbf{f}_i\left(\boldsymbol{\theta}_i, \mathbf{t}\right)\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{u}_i\right]' \boldsymbol{\Sigma}_e^{-1}\left[\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{u}_i\right]\right\} \end{aligned} \tag{15.65}$$

for $i = 1, 2, ..., M$. The sampling method to be used depends on whether or not the first-stage model is linear in the trajectory parameters or not.

Linear First-Stage: Gibbs Sampling

In the case of a linear first-stage model

$$\mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t}) = \mathbf{T}_i \boldsymbol{\theta}_i, \quad i = 1, 2, \ldots, M,$$

for some known matrix $\mathbf{T}_i$. Standard results for Bayesian linear models with known dispersion parameters give as conditional posterior distribution,

$$\boldsymbol{\theta}_i | \boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}_i \sim N\left(\overline{\boldsymbol{\theta}}_i, \mathbf{V}_i\right), \quad i = 1, 2, \ldots, M, \tag{15.66}$$

where

$$\overline{\boldsymbol{\theta}}_i = \left[\mathbf{T}_i' \mathbf{R}_i^{-1}(\boldsymbol{\gamma}) \mathbf{T}_i + \boldsymbol{\Sigma}_e^{-1}\right]^{-1} \left[\mathbf{T}_i' \mathbf{R}_i^{-1}(\boldsymbol{\gamma}) \mathbf{y}_i + \boldsymbol{\Sigma}_e^{-1}(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i)\right], \tag{15.67}$$

and

$$\mathbf{V}_i = \left[\mathbf{T}_i' \mathbf{R}_i^{-1}(\boldsymbol{\gamma}) \mathbf{T}_i + \boldsymbol{\Sigma}_e^{-1}\right]^{-1}. \tag{15.68}$$

Hence, collecting samples from the conditional posterior of the trajectory parameters is straightforward, as the draws involve sampling from M independent r-variate normal distributions at each iteration of the MCMC process. Most often, r is small, e.g., $r = 3$ in a Gompertz growth function.

Nonlinear First Stage: Metropolis–Hastings

When the trajectory is not linear in $\boldsymbol{\theta}_i$, the kernel of the distribution (15.65) does not have a recognizable form, so direct drawing is not feasible. A first possibility here consists of setting up a Metropolis–Hastings scheme. The main difficulty is the fine-tuning of a proposal distribution that does not result either in too frequent rejection or in a large acceptance rate. In the first situation, the sampler does not visit effectively the parameter space; in the second one, the algorithm moves very slowly, so that the ensuing effective sample size is small. Additional details on Metropolis–Hastings computations are in the chapter on MCMC procedures.

A reasonable candidate-generating distribution might be a multivariate normal process such as that in (15.66), but using the pseudo-data (instead of the observed data vector) resulting from the Taylor series expansion

$$\widehat{\mathbf{y}}_i = \widehat{\mathbf{X}}_i \widehat{\boldsymbol{\beta}} + \widehat{\mathbf{Z}}_i \widehat{\mathbf{u}} + \widehat{\mathbf{H}} \widehat{\mathbf{e}}_i + \widehat{\boldsymbol{\varepsilon}}_i$$

evaluated at the current values of the unknowns in the course of iteration. A computationally simpler (although probably less effective) proposal distribution could be a normal process with fixed mean vector and variance–covariance matrix. For example, the proposal could be centered at the maximum likelihood estimates of individual $\boldsymbol{\theta}_i$ parameters, $i = 1, 2, ..., M$. The covariance matrix could be taken to be equal to the inverse of Fisher's expected information (given $\boldsymbol{\theta}_i$). This may work well if trajectories are "long enough", but it may be unsatisfactory for individuals having sparse longitudinal information.

Nonlinear First Stage: Acceptance–Rejection

Another option in the nonlinear situation is using a rejection scheme (Ripley, 1987), also known as "acceptance and rejection". Here the basic idea is to cover the conditional posterior density of interest by an envelope, this being the product of some sampling density $S(\boldsymbol{\theta}_i)$ times a positive constant Q_i, such that density of the conditional posterior is smaller or equal than the envelope at all values of $\boldsymbol{\theta}_i$. In our context, the required condition is, from (15.65), that $C(\boldsymbol{\theta}_i) \leq 1$, where

$$C(\boldsymbol{\theta}_i) = \frac{\exp\left\{-\frac{1}{2}\left[\boldsymbol{\varepsilon}_i'(\boldsymbol{\theta}_i)\,\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\,\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_i) + \mathbf{e}_i'(\boldsymbol{\theta}_i)\,\boldsymbol{\Sigma}_e^{-1}\mathbf{e}_i(\boldsymbol{\theta}_i)\right]\right\}}{S(\boldsymbol{\theta}_i)\,Q_i} \tag{15.69}$$

$i = 1, 2, ..., M$. The first- and second-stage residuals are written as

$$\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_i) = \mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t}),$$

and

$$\mathbf{e}_i(\boldsymbol{\theta}_i) = \boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{u}_i,$$

respectively, to emphasize the dependence on $\boldsymbol{\theta}_i$; the integration constant of the conditional posterior is unimportant because it can be absorbed in Q_i. Now take

$$S(\boldsymbol{\theta}_i) \propto \exp\left[-\frac{1}{2}\mathbf{e}_i'(\boldsymbol{\theta}_i)\,\boldsymbol{\Sigma}_e^{-1}\mathbf{e}_i(\boldsymbol{\theta}_i)\right], \tag{15.70}$$

that is, the sampling density is the conditional (given all other parameters) prior of $\boldsymbol{\theta}_i$. In short, $S(\boldsymbol{\theta}_i)$ is the density of the normal process

$$\boldsymbol{\theta}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i, \boldsymbol{\Sigma}_e). \tag{15.71}$$

Further, let $\widetilde{\boldsymbol{\theta}}_i$ be the conditional ML estimator of $\boldsymbol{\theta}_i$ obtained from the first-stage of the model, assuming $\mathbf{R}_i(\boldsymbol{\gamma})$ is known (here, $\boldsymbol{\gamma}$ would be fixed at the current value of the MCMC scheme). Then, set

$$Q_i = \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'\left(\widetilde{\boldsymbol{\theta}}_i\right)\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\,\boldsymbol{\varepsilon}_i\left(\widetilde{\boldsymbol{\theta}}_i\right)\right].$$

It follows that in (15.69):

$$C(\boldsymbol{\theta}_i) = \frac{\exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'(\boldsymbol{\theta}_i)\,\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\,\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_i)\right]}{\exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'\left(\widetilde{\boldsymbol{\theta}}_i\right)\mathbf{R}_i^{-1}(\boldsymbol{\gamma})\,\boldsymbol{\varepsilon}_i\left(\widetilde{\boldsymbol{\theta}}_i\right)\right]} \leq 1, \tag{15.72}$$

for all i. This is so because the conditional likelihood, being proportional to the numerator, is maximized when evaluated at $\widetilde{\boldsymbol{\theta}}_i$, so the denominator must be at least as large as the numerator for any value of the trajectory parameters, with this being true for each individual. The sampling scheme is conducted as follows:

(a) draw the parameters from the conditional prior (15.71);
(b) evaluate (15.72), and
(c) extract a random deviate from an uniform $U(0,1)$ process.

If this deviate is smaller than $C(\boldsymbol{\theta}_i)$, the value sampled from the conditional prior is accepted as belonging to the conditional posterior distribution having density as in (15.65); otherwise, the sample value is rejected, so the process must be repeated until acceptance. A potential problem with this scheme is that if the values drawn from the conditional prior have a small likelihood, then $C(\boldsymbol{\theta}_i)$ is always very small, causing a high rate of rejection.

The sampling scheme is dynamic, in the sense that the parameters of (15.71) change in the course of iteration. A more general treatment of adaptive rejection/sampling schemes is in Gilks and Wild (1992). The basic idea is that when a point is rejected, the envelope is updated, to correspond more closely to the target density. This reduces the chances of rejecting subsequent proposals, thus decreasing the number of functions that need to be evaluated.

Nonlinear First Stage: Importance Sampling

A third possibility for nonlinear first-stage models consists of employing an importance sampling/resampling scheme (e.g., Tanner, 1996). This topic was discussed in Chapter 12 in conjunction with a sensitivity analysis of the Bayesian model. Here it is described in a little more detail in the context of the longitudinal data problem. We consider first how the method can be used to compute a posterior expectation, and then see how the drawn samples can be resampled, to arrive at draws from the posterior distribution of interest.

Let $g(\boldsymbol{\theta})$ be the density of some posterior distribution of interest, and suppose one wishes to compute the posterior expectation of the parameter vector $\boldsymbol{\theta}$ (or of $h(\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$); that is:

$$\begin{aligned} E_g(\boldsymbol{\theta}) &= \int \boldsymbol{\theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{\int \boldsymbol{\theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \end{aligned}$$

Often, this computation is not feasible, because of analytical difficulties. Suppose there is some distribution having the same support as the posterior of interest, and which is easy to sample from. This will be called the importance distribution, and let its density be $I(\boldsymbol{\theta})$. For example, if the sampling space of $\boldsymbol{\theta}$ is $\Re^r$, perhaps a multivariate normal or multivariate$-t$ distribution of order r could be used. The posterior expectation of $\boldsymbol{\theta}$ is derived from (12.30):

$$E_g(\boldsymbol{\theta}) = \frac{\int \boldsymbol{\theta} w(\boldsymbol{\theta}) I(\boldsymbol{\theta}) d\boldsymbol{\theta}}{E_I[w(\boldsymbol{\theta})]} \tag{15.73}$$

where

$$w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})}{I(\boldsymbol{\theta})} \tag{15.74}$$

and

$$E_I\left[w(\boldsymbol{\theta})\right] = \int w(\boldsymbol{\theta}) I(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int g(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Defining the random variable

$$z(\boldsymbol{\theta}) = \frac{w(\boldsymbol{\theta})}{E_I\left[w(\boldsymbol{\theta})\right]}\boldsymbol{\theta}, \tag{15.75}$$

one can write

$$E_g(\boldsymbol{\theta}) = \int z(\boldsymbol{\theta}) I(\boldsymbol{\theta}) d\boldsymbol{\theta} = E_I\left[z(\boldsymbol{\theta})\right].$$

Now suppose that m independent samples are drawn from the distribution with density $I(\boldsymbol{\theta})$, also called the "importance sampling" function, and let such draws be $\boldsymbol{\theta}^{[k]}$ $(k = 1, 2, \ldots, m)$. Then a simulation consistent estimator of the posterior expectation $E_g(\boldsymbol{\theta})$ or, equivalently, of $E_I\left[z(\boldsymbol{\theta})\right]$, is given by (12.31)

$$\widehat{E}_I\left[z(\boldsymbol{\theta})\right] = \sum_{k=1}^{m} \frac{w(\boldsymbol{\theta}^{[k]})}{\sum_{k=1}^{m} w(\boldsymbol{\theta}^{[k]})} \boldsymbol{\theta}^{[k]}. \tag{15.76}$$

In this expression, the denominator in (15.73) has been replaced by its consistent estimator

$$\widehat{E}_I\left[w(\boldsymbol{\theta})\right] = \frac{1}{m} \sum_{k=1}^{m} w(\boldsymbol{\theta}^{[k]}). \tag{15.77}$$

The "importance" weights are

$$\begin{aligned} \frac{w(\boldsymbol{\theta}^{[k]})}{\sum_{k=1}^{m} w(\boldsymbol{\theta}^{[k]})} &= \frac{\frac{g(\boldsymbol{\theta}^{[k]})}{I(\boldsymbol{\theta}^{[k]})}}{\sum_{k=1}^{m} \frac{g(\boldsymbol{\theta}^{[k]})}{I(\boldsymbol{\theta}^{[k]})}} \\ &= \frac{\frac{cp\left(\mathbf{y}|\boldsymbol{\theta}^{[k]}\right)p\left(\boldsymbol{\theta}^{[k]}\right)}{I(\boldsymbol{\theta}^{[k]})}}{\sum_{k=1}^{m} \frac{cp\left(\mathbf{y}|\boldsymbol{\theta}^{[k]}\right)p\left(\boldsymbol{\theta}^{[k]}\right)}{I(\boldsymbol{\theta}^{[k]})}} \\ &= \frac{\frac{p\left(\mathbf{y}|\boldsymbol{\theta}^{[k]}\right)p\left(\boldsymbol{\theta}^{[k]}\right)}{I(\boldsymbol{\theta}^{[k]})}}{\sum_{k=1}^{m} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}^{[k]}\right)p\left(\boldsymbol{\theta}^{[k]}\right)}{I(\boldsymbol{\theta}^{[k]})}}, \end{aligned} \tag{15.78}$$

where $p\left(\boldsymbol{\theta}\right)$ is the prior density, $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$ is the density of the sampling model, and c is the integration constant of the posterior density. It follows that

knowledge of the constant of integration is not needed to carry out the importance sampling process, as it cancels out in the numerator and denominator. Note, incidentally, that for the "new" weight

$$w^*(\boldsymbol{\theta}) = \frac{p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{I(\boldsymbol{\theta})},$$

$$\begin{aligned} E_I\left[w^*(\boldsymbol{\theta})\right] &= \int \frac{p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{I(\boldsymbol{\theta})} I(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} = c^{-1}. \end{aligned}$$

Hence, a simulation consistent estimator of the integration constant (or of its reciprocal) is given by

$$\widehat{c} = \frac{m}{\sum\limits_{k=1}^{m} \frac{p\left(\mathbf{y}|\boldsymbol{\theta}^{[k]}\right) p\left(\boldsymbol{\theta}^{[k]}\right)}{I(\boldsymbol{\theta}^{[k]})}}.$$

As an important tuning issue, observe that an importance sampling density having thinner tails than the unnormalized posterior may cause some weights to "blow up", with the consequence that a few values dominate others in the weighted average. This is a reason why a multivariate-t distribution, perhaps with six to eight degrees of freedom may be a better importance function than the multivariate normal. For example, in the context of drawing the nonlinear parameters of the individual trajectories, one could use the following r-variate t distribution, constructed from (15.66):

$$\boldsymbol{\theta}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}, \mathbf{y}_i \sim t_r\left(\overline{\boldsymbol{\theta}}_i, \mathbf{V}_i \frac{v}{v-2}, v\right)$$

with v degrees of freedom, and where $\mathbf{V}_i \frac{v}{v-2}$ is the variance–covariance matrix of the t process.

Note in (15.76) that when the weights are constant

$$\widehat{E}_I\left[z(\boldsymbol{\theta}_i)\right] = \sum_{k=1}^{m} \frac{w(\boldsymbol{\theta}_i^{[k]})}{\sum\limits_{k=1}^{m} w(\boldsymbol{\theta}_i^{[k]})} \boldsymbol{\theta}_i^{[k]} = \frac{1}{m} \sum_{k=1}^{m} \boldsymbol{\theta}_i^{[k]},$$

which is the posterior expectation calculated directly from the importance sampling distribution. This implies that as the coefficient of variation of the weights goes to 0, the importance distribution "gets closer" to the posterior distribution. This follows from (15.74), that is, if $g(\boldsymbol{\theta}_i)/I(\boldsymbol{\theta}_i)$ is a constant that does not depend on $\boldsymbol{\theta}_i$, it cancels out in the expression, with the result that the importance distribution is identical to the posterior distribution of interest.

The preceding indicates how a posterior expectation can be computed, but does not give guidance on how a sample is to be drawn from the conditional posterior distribution with density (15.65), so that the MCMC procedure can continue. Now (15.73) implies that the random variable $\boldsymbol{\theta}_i$ with posterior density (15.65) has the same distribution as $\boldsymbol{\theta}_i$ with density

$$\frac{w(\boldsymbol{\theta})I(\boldsymbol{\theta})}{E_I\left[w(\boldsymbol{\theta})\right]},$$

which clearly integrates to 1. This observation is the basis of the importance sampling/resampling algorithm of Rubin (1988):
(1) Draw $\boldsymbol{\theta}_i^{[k]}$, $(k = 1, 2, ..., m)$, from the importance distribution.
(2) Calculate the weights

$$w^*(\boldsymbol{\theta}_i^{[k]}) = \frac{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[k]}\right) p\left(\boldsymbol{\theta}_i^{[k]}\right)}{I(\boldsymbol{\theta}_i^{[k]})},$$

and the relative weights

$$q^{[k]} = \frac{w^*(\boldsymbol{\theta}_i^{[k]})}{\sum_{k=1}^{m} w^*(\boldsymbol{\theta}_i^{[k]})}.$$

(3) Draw $\boldsymbol{\theta}_i^*$ from a discrete distribution with $\Pr\left(\boldsymbol{\theta}_i^* = \boldsymbol{\theta}_i^{[k]}\right) = q^{[k]}$, so the sample space of $\boldsymbol{\theta}_i^*$ is the set

$$\left\{\boldsymbol{\theta}_i^{[1]}, \boldsymbol{\theta}_i^{[2]}, ..., \boldsymbol{\theta}_i^{[m]}\right\}.$$

As $m \to \infty$, then $\boldsymbol{\theta}_i^*$ is a draw from the target posterior distribution. In our context, this requires drawing a large number of importance samples from the conditional posterior distribution of each of the trajectory parameters, and then resampling one value at random, with probability $q^{[k]}$. This value would be retained to continue the MCMC algorithm.

Second-Stage Location Effects

Return to the joint posterior density in (15.63) and consider the part that varies with $\boldsymbol{\beta}$ and $\mathbf{u}$ only. The conditional posterior density of $\boldsymbol{\beta}$ and $\mathbf{u}$ is then

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \mathbf{u}|\boldsymbol{\theta}, \mathbf{G_0}, \boldsymbol{\Sigma_e}, \boldsymbol{\gamma}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \\ &\propto \left[\prod_{i=1}^{M} p\left(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right)\right] p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u}|\mathbf{G}_0\right) \\ &\quad \propto p\left(\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}_e\right) p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) p\left(\mathbf{u}|\mathbf{G}_0\right). \end{aligned} \tag{15.79}$$

This is precisely the posterior distribution of "fixed" and "random" effects in a multivariate Gaussian mixed effects linear model with known dispersion parameters, but with $\boldsymbol{\theta}_i$ in lieu of the observations taken in individual i. Here $\boldsymbol{\theta}_i$ can be viewed as r attributes measured simultaneously on individual i. As developed in previous chapters, the conditional posterior distribution is Gaussian, with mean vector

$$\begin{bmatrix} \overleftrightarrow{\boldsymbol{\beta}} \\ \overleftrightarrow{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X}+\boldsymbol{\Gamma}^{-1} & \mathbf{X}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{Z} \\ \mathbf{Z}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X} & \mathbf{Z}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{Z}+\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1} \end{bmatrix}^{-1} \times \begin{bmatrix} \mathbf{X}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\boldsymbol{\theta}+\boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha} \\ \mathbf{Z}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\boldsymbol{\theta} \end{bmatrix}, \tag{15.80}$$

and variance–covariance matrix

$$\begin{bmatrix} \mathbf{X}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X}+\boldsymbol{\Gamma}^{-1} & \mathbf{X}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{Z} \\ \mathbf{Z}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X} & \mathbf{Z}'\left(\mathbf{I}\otimes\boldsymbol{\Sigma}_e^{-1}\right)\mathbf{Z}+\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1} \end{bmatrix}^{-1}. \tag{15.81}$$

It has been seen already that the draws from a Gaussian posterior distribution can be effected either in a piecewise, blockwise, or multivariate manner, with the only consequence of the method chosen being on the mixing rate of the MCMC scheme. Naturally, if the order of $\boldsymbol{\theta}$ allows one to do so, samples can be obtained by standard methods for drawing from multivariate Gaussian distribution.

First-Stage Dispersion Parameters

From (15.63), the conditional posterior density of the first-stage dispersion parameter vector $\boldsymbol{\gamma}$ is

$$\begin{aligned} &p\left(\boldsymbol{\gamma}|\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{u},\mathbf{G}_0,\boldsymbol{\Sigma}_e,\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_M,\boldsymbol{\alpha},\boldsymbol{\Gamma}\right) \\ &\propto \prod_{i=1}^{M}\left|\mathbf{R}_i\left(\boldsymbol{\gamma}\right)\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}_i'\mathbf{R}_i^{-1}\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon}_i\right]. \end{aligned} \tag{15.82}$$

The form of this distribution depends on the specification of the covariance matrix $\mathbf{R}_i\left(\boldsymbol{\gamma}\right)$ in (15.4). For example, if the first-stage residuals are assumed to be independently distributed and homoscedastic, then $\mathbf{R}_i\left(\boldsymbol{\gamma}\right)=\mathbf{I}_{n_i}\gamma$. Using this in (15.82) gives, as conditional posterior,

$$\begin{aligned} p\left(\boldsymbol{\gamma}|\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{u},\mathbf{G}_0,\boldsymbol{\Sigma}_e,\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_M,\boldsymbol{\alpha},\boldsymbol{\Gamma}\right) &\propto \prod_{i=1}^{M}\gamma^{-\frac{n_i}{2}}\exp\left(-\frac{1}{2\gamma}\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i\right) \\ &\propto \gamma^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\gamma}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left[y_{ij}-f_{ij}\left(\boldsymbol{\theta}_i,t_{ij}\right)\right]^2\right\}, \end{aligned} \tag{15.83}$$

where $N = \sum_{i=1}^{M} n_i$ is the total number of observations taken. This density is that of the scaled inverted chi-square random variable

$$\gamma|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}$$
$$\sim \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left[y_{ij} - f_{ij} \left(\boldsymbol{\theta}_i, t_{ij} \right) \right]^2 \right\} \chi^{-2}_{N-2}, \quad \gamma \in \Re^{+}_{\gamma} \tag{15.84}$$

which may be truncated if limits are placed on the parameter space $\Re^{+}_{\gamma}$. On the other hand, if the residuals are independent but heteroscedastic across individuals, such that $\mathbf{R}_i(\boldsymbol{\gamma}) = \mathbf{I}_{n_i}\gamma_i$ for $i = 1, 2, ..., M$, so $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, ..., \gamma_M]'$, then the first-stage dispersion parameters are conditionally independent of each other, with each having the conditional posterior distribution

$$\gamma_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\Sigma}_e, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}$$
$$\sim \left\{ \sum_{j=1}^{n_i} \left[y_{ij} - f_{ij} \left(\boldsymbol{\theta}_i, t_{ij} \right) \right]^2 \right\} \chi^{-2}_{n_i-2}, \quad \gamma_i \in \Re_{\gamma_i}. \tag{15.85}$$

While in the preceding cases the conditional posterior distributions can be identified and are easy to sample from, this is not so when there is a more complex structure in the residual variance–covariance matrix. For example, with auto-regressive processes, Markov-type dependencies, or with a structural model for $\mathbf{R}_i(\boldsymbol{\gamma})$, the corresponding distributions cannot be recognized. Hence, a Metropolis–Hastings step, for example, must be incorporated.

Second-Stage Dispersion Parameters

The conditional posterior density of the second-stage residual variance-covariance matrix can be deduced directly from (15.63) yielding

$$p\left(\boldsymbol{\Sigma}_e|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{G}_0, \boldsymbol{\gamma}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}\right) \propto \prod_{i=1}^{M} p\left(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \mathbf{u}_i, \boldsymbol{\Sigma}_e\right)$$
$$\propto \left|\boldsymbol{\Sigma}_e\right|^{-\frac{M}{2}} \exp\left[-\frac{1}{2} tr\left(\boldsymbol{\Sigma}_e^{-1}\mathbf{B}\right)\right], \tag{15.86}$$

where, as before, $\mathbf{B} = \sum_{i=1}^{M} \mathbf{e}_i\mathbf{e}_i'$ is an $r \times r$ matrix and $\mathbf{e}_i = \boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{u}_i$. The preceding density is the kernel of a scaled inverted Wishart distribution of order r, scale matrix $\mathbf{B}$, and "degrees of freedom" parameter equal to $M - r - 1$. If the parameter space of $\boldsymbol{\Sigma}_e$ is subject to restrictions beyond $|\boldsymbol{\Sigma}_e| > 0$, then the distribution is a truncated scaled inverted Wishart. As seen in a previous chapter, it is relatively easy to sample from standard or truncated scaled inverted Wishart distributions.

Reorder now the additive genetic effects by nesting individuals within parameters, and let the ensuing vector be $\mathbf{u}^*$, such that $\mathbf{u}_1^*$ contains additive genetic effects for parameter 1, and so on. Thus, the conditional posterior density of the second-stage additive genetic variance–covariance matrix $\mathbf{G}_0$ is

$$\begin{aligned} p\left(\mathbf{G}_0|\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\Sigma}_e,\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_M,\boldsymbol{\alpha},\boldsymbol{\Gamma}\right) &\propto p\left(\mathbf{u}|\mathbf{G}_0\right) = p\left(\mathbf{u}^*|\ \mathbf{G}_0\right) \\ &\propto \left|\mathbf{G}_0\otimes\mathbf{A}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{u}^{*\prime}\left(\mathbf{G}_0\otimes\mathbf{A}\right)^{-1}\mathbf{u}^*\right] \\ &\propto \left|\mathbf{G}_0\right|^{-\frac{q}{2}}\exp\left[-\frac{1}{2}\mathbf{u}^{*\prime}\left(\mathbf{G}_0^{-1}\otimes\mathbf{A}^{-1}\right)\mathbf{u}^*\right] \\ &\propto \left|\mathbf{G}_0\right|^{-\frac{q}{2}}\exp\left[-\frac{1}{2}\,tr\left(\mathbf{G}_0^{-1}\mathbf{U}^*\right)\right], \end{aligned} \tag{15.87}$$

where, as before, q is the order of the additive genetic relationship matrix $\mathbf{A}$, and $\mathbf{U}^*$ is the $r\times r$ symmetric matrix

$$\mathbf{U}^*=\begin{bmatrix} \mathbf{u}_1^{*\prime}\mathbf{A}^{-1}\mathbf{u}_1^* & \mathbf{u}_1^{*\prime}\mathbf{A}^{-1}\mathbf{u}_2^* & \cdots & \mathbf{u}_1^{*\prime}\mathbf{A}^{-1}\mathbf{u}_r^* \\ \mathbf{u}_2^{*\prime}\mathbf{A}_1^{-1}\mathbf{u}_1^* & \mathbf{u}_2^{*\prime}\mathbf{A}^{-1}\mathbf{u}_2^* & \cdots & \mathbf{u}_1^{*\prime}\mathbf{A}^{-1}\mathbf{u}_r^* \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_r^{*\prime}\mathbf{A}^{-1}\mathbf{u}_1^* & \mathbf{u}_r^{*\prime}\mathbf{A}^{-1}\mathbf{u}_2^* & \cdots & \mathbf{u}_r^{*\prime}\mathbf{A}^{-1}\mathbf{u}_r^* \end{bmatrix}.$$

Density (15.87) is the kernel of a scaled inverted Wishart distribution of order r, with scale matrix $\mathbf{U}^*$ and $q-r-1$ degrees of freedom.

This completes the description of all conditional posterior distributions needed to implement the MCMC procedure. Once a chain of appropriate length is run, and convergence to the equilibrium distribution seems to have been attained, the analysis of the output proceeds in a standard manner. Output analysis is discussed in Chapter 12.

15.5 Analysis with Thick-Tailed Distributions

It is known that the normal distribution is sensitive to departures from assumptions, so one may wish to consider models based on "robust" distributions, and one of these is the t process, either in its univariate or multivariate form (e.g., Rogers and Tukey, 1972; Zellner, 1976; Lange and Sinsheimer, 1993; Strandén and Gianola, 1999). Regression and cross-sectional models with t distributed errors have been discussed in Chapter 13, Section 13.6.

In the treatment of longitudinal data given so far, normality has been assumed for the residuals of the first two stages of the models. In what follows, the assumption of normality of residuals at the first stage will be replaced

by one of i.i.d. errors having a univariate-t distribution with unknown degrees of freedom. At the second stage, a multivariate-t distribution will be employed for the second-stage residuals, instead of an r-variate normal distribution. For the sake of simplicity, it will be assumed that individuals are genetically unrelated to each other. While this assumption is not tenable in animal breeding, it is used frequently in biostatistics (Laird and Ware, 1982).

15.5.1 First- and Second-Stage Models

Recall that a t distribution, either univariate or multivariate, arises from mixing a normal distribution over a gamma (equivalently, over an inverted gamma or scale inverted chi-squared) process. This can be used to advantage in an MCMC implementation by augmenting the joint posterior with some unobservable "weights" following the appropriate gamma distributions.

The first-stage model will be as in (15.2), but amended as

$$\begin{aligned} y_{ij} &= f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right) + \frac{\varepsilon_{ij}}{\sqrt{w_{ij}}} \\ &= f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right) + \varepsilon_{ij}^*, \end{aligned} \tag{15.88}$$

where $\varepsilon_{ij} \sim N\left(0, \gamma\right)$ and $w_{ij} \sim Ga\left(\frac{\nu_\varepsilon}{2}, \frac{\nu_\varepsilon}{2}\right)$ are independently distributed random variables. As seen before, the distribution of the "new" residual ε_{ij}^* can be shown to be $t\left(0, \gamma, \nu_\varepsilon\right)$, where γ is the scale parameter of the t distribution and ν_ε are the degrees of freedom, which will be treated as unknown. The new residuals will be assumed to be mutually independent, both within and between individuals. Note that, given w_{ij}, then

$$y_{ij}|\boldsymbol{\theta}_i, t_{ij}, \gamma, w_{ij} \sim N\left(f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right), \frac{\gamma}{w_{ij}}\right).$$

Likewise, the second-stage model will be taken to be

$$\boldsymbol{\theta}_i = \mathbf{X}_i\boldsymbol{\beta} + \frac{\mathbf{e}_i}{\sqrt{w_i}} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i^*, \quad i = 1, 2, \ldots, M, \tag{15.89}$$

where $\mathbf{e}_i \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_e\right)$ and $w_i \sim Ga\left(\frac{\nu_e}{2}, \frac{\nu_e}{2}\right)$ are independently distributed, for all i, as well as between individuals. Under this assumption, the distribution of the "new" residual is the r-dimensional process $t_r\left(\mathbf{0}, \boldsymbol{\Sigma}_e, \nu_e\right)$, where $\boldsymbol{\Sigma}_e$ is the scale matrix and ν_ε are the degrees of freedom, which will also be treated as unknown. The trajectory parameters will be assumed to be independent across individuals, a priori. As stated above, a genetic effect is not included in the specification of the second-stage model. Here, the second-stage residual reflects both genetic and nongenetic sources of

variation (and covariation) between parameters, plus any error in the specification of the model. Note that, given w_i, then

$$\boldsymbol{\theta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, w_i \sim N\left(\mathbf{X}_i\boldsymbol{\beta}, \frac{\boldsymbol{\Sigma}_e}{w_i}\right).$$

The prior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_e$, and γ will be as in (15.17), (15.20), and (15.21), and the prior distributions of the two degrees of freedom parameters will be assumed to be independent a priori, and uniform within some bounded interval of the positive part of the real line. After augmenting the prior with the trajectory parameters (as before) and with all weights w_{ij} and w_i, the joint prior density of all unknowns can be written as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma, \boldsymbol{\Sigma}_e, \mathbf{w}_\varepsilon, \mathbf{w}_e, \nu_\varepsilon, \nu_e|\boldsymbol{\alpha}, \boldsymbol{\Gamma}) &= p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \mathbf{w}_e)\, p(\boldsymbol{\beta}|\alpha, \Gamma)\, p(\gamma, \boldsymbol{\Sigma}_e) \\ &\quad \times p(\mathbf{w}_\varepsilon|\nu_\varepsilon)\, p(\mathbf{w}_e|\nu_e)\, p(\nu_\varepsilon)\, p(\nu_e) \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, \mathbf{w}_e)\, p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma})\, p(\mathbf{w}_\varepsilon|\nu_\varepsilon)\, p(\mathbf{w}_e|\nu_e), \end{aligned} \tag{15.90}$$

where $\mathbf{w}_\varepsilon = \{w_{ij}\}$ is an $\sum_{i=1}^{M} n_i \times 1$ vector and $\mathbf{w}_e = \{w_i\}$ is an $M \times 1$ vector of second-stage "weights". Furthermore, in view of the independence assumption for parameters and weights,

$$\begin{aligned} &p(\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma, \boldsymbol{\Sigma}_e, \mathbf{w}_\varepsilon, \mathbf{w}_e, \nu_\varepsilon, \nu_e|\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \\ &\propto \left[\prod_{i=1}^{M} p(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, w_i)\, p(w_i|\nu_e)\right] p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \prod_{i=1}^{M}\prod_{j=1}^{n_i} p(w_{ij}|\nu_\varepsilon). \end{aligned} \tag{15.91}$$

The joint posterior density is expressible as

$$\begin{aligned} &p(\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma, \boldsymbol{\Sigma}_e, \mathbf{w}_\varepsilon, \mathbf{w}_e, \nu_\varepsilon, \nu_e|\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) \\ &\quad \propto \left[\prod_{i=1}^{M}\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\theta}_i, \gamma, w_{ij})\, p(w_{ij}|\nu_\varepsilon)\right] \\ &\quad \left[\prod_{i=1}^{M} p(\boldsymbol{\theta}_i|\boldsymbol{\beta},\ \boldsymbol{\Sigma}_e, w_i)\, p(w_i|\nu_e)\right] p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma}). \end{aligned} \tag{15.92}$$

15.5.2 Fully Conditional Posterior Distributions

Trajectory Parameters

As in the purely normal hierarchical model, all trajectory parameters are conditionally independent from each other. Using the

$$[\text{parameter}|ELSE]$$

notation to represent a fully conditional posterior distribution, one has that

$$p(\boldsymbol{\theta}_i|ELSE) \propto \exp\left\{-\frac{1}{2\gamma}\left[\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t})\right]_i' \mathbf{W}_{\varepsilon i}\left[\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t})\right]\right\}$$
$$\times \exp\left[-\frac{w_i}{2}(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}_e^{-1}(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta})\right], \tag{15.93}$$

where $\mathbf{W}_{\varepsilon i} = Diag\{w_{i1}, w_{i2}, \ldots, w_{in_i}\}$ is an $n_i \times n_i$ diagonal matrix, $i = 1, 2, \ldots, M$.

If the model is nonlinear in the parameters, the distribution is not recognizable, and the samples must be drawn by, for example, the Metropolis–Hastings algorithm. On the other hand, if the model is linear, with

$$\mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{t}) = \mathbf{T}_i\boldsymbol{\theta}_i,$$

then the conditional posterior distribution is normal, so one can use Gibbs sampling since draws can be effected easily. Employing standard results for combining quadratic forms, one can arrive at

$$\boldsymbol{\theta}_i|ELSE \sim N\left(\overline{\boldsymbol{\theta}}_i, \mathbf{V}_i\right), \quad i = 1, 2, \ldots, M. \tag{15.94}$$

The parameters of this normal distribution are

$$\overline{\boldsymbol{\theta}}_i = \left[\mathbf{T}_i'\frac{\mathbf{W}_{\varepsilon i}}{\gamma}\mathbf{T}_i + w_i\boldsymbol{\Sigma}_e^{-1}\right]^{-1}\left[\mathbf{T}_i'\frac{\mathbf{W}_{\varepsilon i}}{\gamma}\mathbf{y}_i + w_i\boldsymbol{\Sigma}_e^{-1}\mathbf{X}_i\boldsymbol{\beta}\right],$$

and

$$\mathbf{V}_i = \left[\mathbf{T}_i'\frac{\mathbf{W}_{\varepsilon i}}{\gamma}\mathbf{T}_i + w_i\boldsymbol{\Sigma}_e^{-1}\right]^{-1}.$$

Second-Stage Location Effects

From the joint posterior density (15.92) one arrives at

$$p(\boldsymbol{\beta}|ELSE) \propto \left[\prod_{i=1}^{M} p(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_e, w_i)\right] p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Gamma})$$
$$\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{M}(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta})' w_i\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\alpha})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\alpha})\right]\right\}$$
$$\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})'\left(\mathbf{W}_e \otimes \boldsymbol{\Sigma}_e^{-1}\right)(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\alpha})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\alpha})\right]\right\}, \tag{15.95}$$

where $\mathbf{W}_e = Diag(w_i)$ is an $M \times M$ matrix of second stage weights. Since the two intervening densities are in Gaussian form, this implies that the conditional posterior distribution of $\boldsymbol{\beta}$ is normal, with mean vector

$$\left[\mathbf{X}'\left(\mathbf{W}_e \otimes \boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X} + \boldsymbol{\Gamma}^{-1}\right]^{-1}\left[\mathbf{X}'\left(\mathbf{W}_e \otimes \boldsymbol{\Sigma}_e^{-1}\right)\boldsymbol{\theta} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\alpha}\right], \tag{15.96}$$

and covariance matrix

$$\left[\mathbf{X}'\left(\mathbf{W}_e \otimes \boldsymbol{\Sigma}_e^{-1}\right)\mathbf{X} + \boldsymbol{\Gamma}^{-1}\right]^{-1}. \tag{15.97}$$

Hence, Gibbs sampling is straightforward.

Scale Parameter of the First Stage Distribution

Retaining, in the joint posterior density (15.92), only the terms that depend on γ leads to

$$\begin{aligned} p(\gamma|ELSE) &\propto \prod_{i=1}^{M}\prod_{j=1}^{n_i} p\left(y_{ij}|\boldsymbol{\theta}_i, \gamma, w_{ij}\right) \\ &\propto \gamma^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\gamma}\sum_{i=1}^{M}\sum_{j=1}^{n_i} w_{ij}\left[y_{ij} - f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)\right]^2\right\}. \end{aligned} \tag{15.98}$$

This is the density of the distribution

$$\gamma|ELSE \sim \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i} w_{ij}\left[y_{ij} - f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)\right]^2\right\}\chi_{N-2}^{-2}, \quad \gamma \in \Re_{\gamma}^{+},$$

which is straightforward to sample from in the context of Gibbs sampling.

Scale Parameter of the Second-Stage Distribution

Similarly,

$$\begin{aligned} p\left(\boldsymbol{\Sigma}_e|ELSE\right) &\propto \prod_{i=1}^{M}\left\{\left|\boldsymbol{\Sigma}_e\right|^{-\frac{1}{2}} \exp\left[-\frac{w_i}{2}\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}_e^{-1}\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\right]\right\} \\ &\propto \left|\boldsymbol{\Sigma}_e\right|^{-\frac{M}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{M} w_i'\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\Sigma}_e^{-1}\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\right] \\ &\propto \left|\boldsymbol{\Sigma}_e\right|^{-\frac{M}{2}} \exp\left[-\frac{1}{2}tr\left(\boldsymbol{\Sigma}_e^{-1}\mathbf{B}_e\right)\right], \end{aligned} \tag{15.99}$$

where

$$\mathbf{B}_e = \sum_{i=1}^{M} w_i\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)'.$$

It can be readily ascertained that (15.99) is the kernel of a scaled inverted Wishart distribution of order r, scale matrix $\mathbf{B}_e$, and "degrees of freedom" parameter equal to $M-r-1$. If the parameter space of the scaled matrix $\boldsymbol{\Sigma}_e$ is subject to restrictions beyond $|\boldsymbol{\Sigma}_e| > 0$, then the resulting distribution is truncated scaled-inverted Wishart.

Weight Parameters

The preceding conditional posterior distributions indicate that, so far, the sampling process is as in the purely normal case, with the only novelty being the appearance of the weights, which need to be sampled from the corresponding conditional distributions. Consider the joint posterior density as a function of the first-stage weights w_{ij}, yielding

$$p\left(w_{ij}|ELSE\right) \propto p\left(y_{ij}|\boldsymbol{\theta}_i, \gamma, w_{ij}\right) p\left(w_{ij}|\nu_\varepsilon\right)$$
$$\propto w_{ij}^{\frac{1}{2}} \exp\left\{-\frac{w_{ij}\left[y_{ij} - f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)\right]^2}{2\gamma}\right\} \left[w_{ij}^{\frac{\nu_\varepsilon}{2}-1} \exp\left(-\frac{\nu_\varepsilon w_{ij}}{2}\right)\right]$$

for $i = 1, 2, \ldots, M$ and $j = 1, 2, , \ldots, n_i$. The expression in brackets is the contribution from the gamma prior distribution of the first-stage weights. Rearrangement leads to

$$p\left(w_{ij}|ELSE\right) \propto w_{ij}^{\frac{\nu_\varepsilon+1}{2}-1} \exp\left(-\frac{w_{ij}S_{ij}}{2}\right), \tag{15.100}$$

where

$$S_{ij} = \frac{\left[y_{ij} - f_{ij}\left(\boldsymbol{\theta}_i, t_{ij}\right)\right]^2 + \nu_\varepsilon\gamma}{\gamma}.$$

Hence, the conditional posterior distribution of each w_{ij} weight is the gamma process

$$w_{ij}|ELSE \sim Ga\left(\frac{\nu_\varepsilon + 1}{2}, \frac{S_{ij}}{2}\right), \quad i = 1, 2, \ldots, M, \ j = 1, 2, \ldots, n_i. \tag{15.101}$$

Thus, the samples can be drawn without difficulty.

Similarly, the density of the conditional posterior distribution of the second-stage weights can be put, after some algebra, as

$$p\left(w_i|ELSE\right) \propto w_i^{\frac{\nu_e+1}{2}-1} \exp\left\{-\frac{w_i S_i}{2}\right\},$$

where

$$S_i = \left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' \boldsymbol{\Sigma}_e^{-1} \left(\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta}\right) + \nu_e.$$

Thus

$$w_i|ELSE \sim Ga\left(\frac{\nu_e + 1}{2}, \frac{S_i}{2}\right), \quad i = 1, 2, \ldots, M. \tag{15.102}$$

Degrees of Freedom

Inspection of the joint posterior density of the degrees of freedom parameters reveals that, given all other parameters, the degrees of freedom of the

first- and second-stage distributions are mutually independent. For the first stage distribution, one has

$$p(\nu_\varepsilon|ELSE) \propto \prod_{i=1}^{M}\prod_{j=1}^{n_i} p(w_{ij}|\nu_\varepsilon)$$
$$\propto \left[\frac{\left(\frac{\nu_\varepsilon}{2}\right)^{\left(\frac{\nu_\varepsilon}{2}\right)}}{\Gamma\left(\frac{\nu_\varepsilon}{2}\right)}\right]^N \prod_{i=1}^{M}\prod_{j=1}^{n_i}\left[w_{ij}^{\frac{\nu_\varepsilon}{2}-1}\exp\left(-\frac{\nu_\varepsilon w_{ij}}{2}\right)\right]. \tag{15.103}$$

For the second-stage degrees of freedom parameters, the resulting conditional posterior distribution is

$$p(\nu_e|ELSE) \propto \prod_{i=1}^{M} p(w_i|\nu_e)$$
$$\propto \left[\frac{\left(\frac{\nu_e}{2}\right)^{\left(\frac{\nu_e}{2}\right)}}{\Gamma\left(\frac{\nu_e}{2}\right)}\right]^M \prod_{i=1}^{M}\left[w_{ij}^{\frac{\nu_e}{2}-1}\exp\left(-\frac{\nu_e w_i}{2}\right)\right]. \tag{15.104}$$

None of the two distributions has a recognizable form. Hence, either a Metropolis–Hastings, rejection, or importance sampling with resampling step needs to be tailored, to draw the degrees of freedom. This is probably the most difficult part of the implementation, since it is not easy to arrive at suitable proposal distributions or rejection envelopes (e.g., Strandén, 1996). An alternative might be to set the degrees of freedom parameters to some fixed values, and then vary these, to study sensitivity of inferences (Rodriguez-Zas, 1998; Rosa et al., 2001). In brief, the distributions with densities as in (15.93), (15.95),(15.98), (15.99), and (15.101)–(15.104), complete the specification of a possible MCMC sampler for a longitudinal data model with two tiers of robustness.

In conclusion, an MCMC analysis is particularly attractive for linear and nonlinear first-stage models for longitudinal data, relative to the two-step approximate Bayesian analysis, where a number of (sometimes dubious) approximations must be employed. In the linear case, the MCMC computations are equivalent to those needed to undertake a Bayesian mixed effects linear model analysis. In the nonlinear situations, computations are more involved. When robust distributions are used to describe the uncertainty about the first- and second-stage models, additional difficulties arise, due to the need to tune proposal distributions for sampling, e.g., the degrees of freedom in the case of the t distributions.

16

Introduction to Segregation and Quantitative Trait Loci Analysis

16.1 Introduction

The genetic model assumed so far is based on a very large number of independent loci with each locus contributing additively with an infinitesimally small effect to the additive genetic value of an individual. This has been termed the infinitesimal model, as noted earlier in this book. Like all models, this is an intellectual abstraction. Box (1976) pointed out that all models are wrong but that some are useful, and this is certainly the case of the infinitesimal model. It has been shown, however, to be remarkably effective for predicting expected response to artificial selection programmes, for predicting breeding values of candidates for selection, for estimating genetic variances and, interestingly enough, for interpreting results from selection experiments (Martinez et al., 2000).

In many traits, in contrast, part of the genetic variance can be attributed to one or more major genes segregating in the population. Many reasons can be advanced for studying the number, location, mode of action and magnitude of such gene effects. The so called mixed inheritance model poses that the genetic variance is partly due to many loci of infinitesimally small effect, and partly due to the presence of a finite number of loci of relatively large effect. Prior to the availability of molecular markers, a method known as complex segregation analysis was one of the most important tools for detection of major genes. Influential papers were Elston and Stewart (1971), Morton and MacLean (1974) and Lange and Elston (1975). This approach was the basis for the successful mapping of many Mendelian genes in the

1980s. Shortcomings that have been pointed out include the limited power of the method for finding major genes and a lack of robustness, leading to false detection (Go et al., 1978). Both problems are much alleviated if information on genetic markers is incorporated into the analysis. The explosion of molecular polymorphisms in the last twenty years or so has stimulated the development and successful application of many methods for major gene or, more generally, for quantitative trait loci (QTL) detection.

This chapter provides an introduction to some models that can be used for inferring the presence of one or more major genes. The chapter is organized as follows. The first section introduces the topic of segregation analysis, and a Bayesian implementation is presented. The second section discusses models for QTL detection. First, a model postulating a single QTL is presented and both likelihood and Bayesian inferences are illustrated. Second, models with an unknown number of QTL are introduced. The chapter ends with an application of reversible jump MCMC for making inferences about the number of QTL segregating in a population.

16.2 Segregation Analysis Models

The simplest mixed inheritance model postulates that there is a single major locus. Applications of the mixed inheritance model using MCMC can be found, for example, in Guo and Thompson (1994), Janss et al. (1995), and Lund and Jensen (1999). A useful recent review including many topics in pedigree analysis, can be found in Thompson (2001).

16.2.1 Notation and Model

Assume a major locus with two alleles, A_1 and A_2, with respective gene frequencies $(1-q)$ and q, and that the base (founding) population from which base individuals were conceptually sampled was in Hardy–Weinberg and linkage equilibrium. The genotypes at the major locus are A_1A_1, A_1A_2, and A_2A_2. Due to Hardy–Weinberg equilibrium, alleles combine independently to form these genotypes with frequencies $(1-q)^2$, $2q(1-q)$, and q^2, respectively. No distinction is made between maternally and paternally inherited alleles. Define a vector $\mathbf{m} = (m_1, m_2, m_3)'$, whose elements describe the effects that genotypes A_1A_1, A_1A_2, and A_2A_2 have on the phenotypic scale.

The genealogy is assumed to consist of n_q individuals. For each individual in the pedigree, define an unknown random variable $\mathbf{w}_i$ $(i = 1, 2, \ldots, n_q)$, taking the values $(1,0,0)$, $(0,1,0)$ or $(0,0,1)$ associated with genotypes A_1A_1, A_1A_2, or A_2A_2, respectively. Also let

$$\Pr(\mathbf{w}_i = k|q), \quad k = 1, 2, 3,$$

Genotype of father:	$\mathbf{w}_{fi} = 2$		
Genotype of mother:	$\mathbf{w}_{mi} = 1$	$\mathbf{w}_{mi} = 2$	$\mathbf{w}_{mi} = 3$
$\mathbf{w}_i = 1$	1/2	1/4	0
$\mathbf{w}_i = 2$	1/2	1/2	1/2
$\mathbf{w}_i = 3$	0	1/4	1/2

TABLE 16.1. Probability of offspring genotypes given parental genotypes, when the father is heterozygote at the major locus.

be the probability that $\mathbf{w}_i$ takes values $(1,0,0)$, $(0,1,0)$, or $(0,0,1)$, respectively.

In the pedigree, there are individuals with unidentified fathers and mothers. These are defined as founders. On the other hand, the nonfounders are individuals with both parents identified. When an individual has only one parent identified, a phantom parent is created. This leads to simpler notation and simpler expressions later on.

Let $\mathbf{W}$ be a matrix of order $n_q \times 3$ such that its ith row is $\mathbf{w}_i'$; thus, $\mathbf{W}$ denotes the configuration of the underlying genotypes at the major locus. The p.m.f. of the genotypic configuration is

$$p(\mathbf{W}|q) = \prod_{\text{founders } i} p(\mathbf{w}_i|q) \prod_{\text{nonfounders } j} p(\mathbf{w}_j|\mathbf{w}_{mj}, \mathbf{w}_{fj}), \tag{16.1}$$

where subscript m (f) stands for the mother (father) of j. The product decomposition of the first term on the right-hand side arises because genotypes of founders are assumed to be a function of gene frequencies, and the model here postulates independence of the two alleles at the locus. The model of genetic transmission gives rise to the product decomposition of the second term on the right-hand side. This postulates that offspring genotypes are conditionally independent, given parental genotypes. Example 1.16 from Chapter 1 illustrates (16.1) for a pedigree with loops.

The first term on the right-hand of (16.1) can take one of the following forms:

$$\Pr[\mathbf{w}_i = (1,0,0)|q] = (1-q)^2,$$

$$\Pr[\mathbf{w}_i = (0,1,0)|q] = 2q(1-q),$$

and

$$\Pr[\mathbf{w}_i = (0,0,1)|q] = q^2.$$

The second term on the right-hand side of (16.1) is a model for Mendelian segregation. To illustrate, Table 16.1 shows the probabilities of offspring genotypes, given the genotypes of both parents, for one of the three possible paternal genotypes (the heterozygote) and the three possible maternal genotypes. The notation has been simplified as follows: $\mathbf{w}_i = (1,0,0)$ becomes $\mathbf{w}_i = 1$; $\mathbf{w}_i = (0,1,0)$ becomes $\mathbf{w}_i = 2$, and $\mathbf{w}_i = (0,0,1)$ becomes $\mathbf{w}_i = 3$.

In models with more than one locus, the p.m.f. $p(\mathbf{w}_j|\mathbf{w}_{mj}, \mathbf{w}_{fj})$ is a function of the recombination fraction between the loci involved, provided these loci are linked.

The model for the data assumes the following linear additive structure:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{W}\mathbf{m} + \mathbf{e}. \tag{16.2}$$

Note that matrix $\mathbf{W}$ is not observed. The other elements of the model have been defined in Section 13.2 of Chapter 13. The (conditional) sampling distribution of the data is assumed to be Gaussian, with the form

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \sigma_e^2 \sim N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{W}\mathbf{m}, \mathbf{I}\sigma_e^2\right),$$

where σ_e^2 is the residual variance. That is, the elements of this vector are assumed to be conditionally independent, given the parameters; the associated densities are referred to as penetrances in the linkage analysis literature.

The prior distribution of the additive genetic values is

$$\mathbf{a}|\mathbf{A}, \sigma_a^2 \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right),$$

where $\mathbf{A}$ and σ_a^2 are the known relationship matrix and the unknown additive genetic variance in the conceptual base population from which base individuals were sampled, respectively. Parameters $\boldsymbol{\beta}$, σ_a^2, and σ_e^2 are assigned priors of the form in (13.4) and (13.5). The vector $\mathbf{m}$ is assigned a proper uniform prior in $\mathbb{R}^3$, of the same form as in (13.4). Finally, a beta distribution with known parameters e and f, $Be(e, f)$, is assigned a priori to describe previous knowledge of the gene frequency; thus,

$$p(q|e, f) \propto q^{e-1}(1-q)^{f-1}. \tag{16.3}$$

After augmentation with $\mathbf{W}$, the joint prior density admits the form

$$p\left(\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, q, \sigma_e^2, \sigma_a^2\right) \propto p(\mathbf{W}|q)\, p(q)\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right).$$

Given the model, the joint posterior distribution of the parameters (we leave implicit the conditioning on $\mathbf{A}$, the known incidence matrices and hyperparameters) is given by

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, q, \sigma_e^2, \sigma_a^2|\mathbf{y}\right) \\ &\propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \sigma_e^2\right) p\left(\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, q, \sigma_e^2, \sigma_a^2\right) \\ &\propto p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \sigma_e^2\right) p(\mathbf{W}|q)\, p(q)\, p\left(\mathbf{a}|\sigma_a^2\right) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right). \end{aligned} \tag{16.4}$$

A classical full likelihood approach involves the joint maximization of

$$L(\Omega|\mathbf{y}) \propto \sum_{\mathbf{W}} \left[\int p\left(\mathbf{y}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \sigma_e^2\right) p\left(\mathbf{a}|\sigma_a^2\right) d\mathbf{a}\right] p(\mathbf{W}|q),$$

over $\boldsymbol{\Omega} = \left(\boldsymbol{\beta}', \mathbf{m}', q, \sigma_e^2, \sigma_a^2\right)'$. This shows that the likelihood is the expected value of the conditional likelihood given $\mathbf{W}$, with the distribution $[\mathbf{W}|q]$ acting as the mixing process. The sum is taken over all possible genotypic configurations of the pedigree, and the dimension of the integral is of the order of the number of elements in $\mathbf{a}$. In the Gaussian model the integration can be performed analytically, but the summation over all possible genotype configurations, in the case of complex pedigrees, is an insurmountable task. In contrast, the MCMC Bayesian model can be implemented in a relatively straightforward manner, as shown below.

16.2.2 Fully Conditional Posterior Distributions

The first step of the Gibbs sampling algorithm is to find a starting value for the genotypic configuration $\mathbf{W}$. A simple way of achieving this is to sample genes/genotypes from the prior distribution of founder individuals, and then sample the nonfounder genotypes from the conditional probability distribution of the nonfounder, given the genotype of its parents. This is known as "gene dropping" (MacCluer et al., 1986) and is essentially a Monte Carlo implementation of (16.1). Since the data contain information on the major gene effects, Guo and Thompson (1994) propose what they call a "posterior gene dropping" approach: the major gene is dropped down from the top of the pedigree conditionally on the current values of parameters and the data on each individual. This method works well in the case of the model under consideration with a continuous penetrance function in the absence of typed genotypic information. However, with other penetrance functions and data structures, as in pedigree studies where the data consist of genotypes of some members of a pedigree (usually the younger ones) and the objects of inference are the genotypes of the remaining members, the approach can be exceedingly inefficient. This is so, because the availability of partial genotypic information imposes rigid compatibility constraints: every proposed configuration must be checked for consistency with the data at hand, and rejection rates can be in the neighborhood of 100%. With this type of data structure, a more efficient approach can be found in Lange and Goradia (1987) and in Lange (1997). Sheehan (2000) provides a good discussion about this and other issues involved in the application of MCMC to genetic analyses on complex pedigrees.

The joint posterior distribution (16.4) has the same form as the joint posterior (13.6). Therefore, allowing for the extra terms involving $\mathbf{W}$ and $\mathbf{m}$, the fully conditional posterior distribution of $\boldsymbol{\theta}' = \left(\boldsymbol{\beta}', \mathbf{m}', \mathbf{a}'\right)$ is identical to (13.9) or (13.11). For example, a simple manipulation of (13.11) shows that the fully conditional posterior distribution of $\mathbf{m}$ is

$$\mathbf{m}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{W}, \sigma_e^2, \mathbf{y} \sim N\left(\widehat{\mathbf{m}}, \left(\mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}\right)^{-1}\sigma_e^2\right),$$

where

$$\widehat{\mathbf{m}} = \left(\mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a}\right).$$

The term $\mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}$ is a diagonal matrix, with elements equal to the number of A_1A_1 genotypes, A_1A_2 genotypes, and A_2A_2 genotypes, among those individuals with records. Having sampled from the fully conditional posterior distribution of $\mathbf{m}$, one may wish to store contrasts like

$$\mathbf{k}' = \begin{bmatrix} 1 & 0 & -1 \\ -0.5 & 1 & -0.5 \end{bmatrix}.$$

The first line in $\mathbf{k}'$ represents the difference between homozygotes at the major locus and the second represents the degree of dominance.

Likewise, the fully conditional posterior distributions of σ_e^2 and σ_a^2 have the same forms as in (13.14) and (13.16). Thus

$$\sigma_a^2|\mathbf{a},\mathbf{y} \sim \left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + \nu_a S_a\right)\chi^{-2}_{n_q+\nu_a} \tag{16.5}$$

and

$$\begin{aligned}&\sigma_e^2|\boldsymbol{\beta},\mathbf{m},\mathbf{a},\mathbf{W},\mathbf{y}\\ &\sim \left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a} - \mathbf{Z}\mathbf{W}\mathbf{m}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{a} - \mathbf{Z}\mathbf{W}\mathbf{m}\right) + \nu_e S_e\right]\chi^{-2}_{n+\nu_e}.\end{aligned} \tag{16.6}$$

The preceding assumes independent scale inverted chi-square prior distributions for σ_a^2 and σ_e^2.

We derive now the fully conditional posterior distribution of $\mathbf{w}_i$. Following Guo and Thompson (1994), define $\{\mathbf{w}_{ij}\}$ as the genotypes of the mates of individual i, $\{\mathbf{w}_{ijl}\}$ as the genotypes of the offspring of individuals i and j, $\mathbf{w}_{mi}$ as the genotype of the mother of i and, finally, $\mathbf{w}_{fi}$ as the genotype of the father of i. From the joint posterior (16.4) we can write

$$\begin{aligned}p\left(\mathbf{w}_i|\boldsymbol{\beta},\mathbf{a},\mathbf{W}_{-i},\mathbf{m},\sigma_e^2,\mathbf{y}\right) &\propto p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{a},\mathbf{W},\mathbf{m},\sigma_e^2\right)p\left(\mathbf{W}\right)\\ &\propto p\left(y_i|\boldsymbol{\beta},\mathbf{a},\mathbf{w}_i,\mathbf{m},\sigma_e^2\right)p\left(\mathbf{w}_i|\mathbf{W}_{-i}\right).\end{aligned} \tag{16.7}$$

The terms that include $\mathbf{w}_i$ in $p\left(\mathbf{w}_i|\mathbf{W}_{-i}\right)$ are

$$\begin{aligned}p\left(\mathbf{w}_i|\mathbf{W}_{-i}\right) &\propto p\left(\mathbf{w}_i|\left\{\mathbf{w}_{ij}\right\},\left\{\mathbf{w}_{ijl}\right\},\mathbf{w}_{mi},\mathbf{w}_{fi}\right)\\ &\propto p\left(\mathbf{w}_i,\left\{\mathbf{w}_{ij}\right\},\left\{\mathbf{w}_{ijl}\right\},\mathbf{w}_{mi},\mathbf{w}_{fi}\right)\\ &= \prod_{j=1}^{n_i} p\left(\mathbf{w}_{ijl}|\mathbf{w}_i,\mathbf{w}_{ij},\mathbf{w}_{mi},\mathbf{w}_{fi}\right)p\left(\mathbf{w}_i,\mathbf{w}_{ij},\mathbf{w}_{mi},\mathbf{w}_{fi}\right)\\ &\propto \prod_{j=1}^{n_i} p\left(\mathbf{w}_{ijl}|\mathbf{w}_i,\mathbf{w}_{ij}\right)p\left(\mathbf{w}_i|\mathbf{w}_{mi},\mathbf{w}_{fi}\right),\end{aligned} \tag{16.8}$$

where n_i are the number of mates of i and n_{ij} (below) are the number of offspring that i has with mate j. Substituting in (16.7) yields

$$p\left(\mathbf{w}_i|\boldsymbol{\beta},\mathbf{a},\mathbf{W}_{-i},\mathbf{m},\sigma_e^2,\mathbf{y}\right) \propto p\left(y_i|\boldsymbol{\beta},\mathbf{a},\mathbf{w}_i,\mathbf{m},\sigma_e^2\right) \times \prod_{j=1}^{n_i}\prod_{l=1}^{n_{ij}} p\left(\mathbf{w}_{ijl}|\mathbf{w}_i,\mathbf{w}_{ij}\right) p\left(\mathbf{w}_i|\mathbf{w}_{mi},\mathbf{w}_{fi}\right). \tag{16.9}$$

This is a discrete distribution of unknown analytical form, but which is easy to sample from. Expression (16.9) is evaluated for the three possible values that $\mathbf{w}_i$ can take and, after normalization, $\mathbf{w}_i$ is accepted with probability equal to the appropriate normalized value. This is done by drawing from a uniform distribution $Un\,(0,1)$.

If individual i has no phenotypic record, the fully conditional distribution is proportional to $p\,(\mathbf{w}_i|\mathbf{W}_{-i})$. If individual i is a founder, the term $p\,(\mathbf{w}_i|\mathbf{w}_{mi},\mathbf{w}_{fi})$ in (16.9) is replaced by $p\,(\mathbf{w}_i|q)$, which is a function of gene frequency, as indicated in (16.1).

Finally, the fully conditional posterior distribution of the gene frequency is obtained as follows. From (16.4),

$$\begin{aligned} p\,(q|\mathbf{W},\mathbf{y}) &\propto p\,(\mathbf{W}|q)\,p\,(q) \\ &\propto p\,(q)\prod_{\text{founders } i} p\,(\mathbf{w}_i|q)\,. \end{aligned} \tag{16.10}$$

Seen as a function of q, the second term has the form

$$\prod_{\text{founders } i} p\,(\mathbf{w}_i|q) \propto (1-q)^{n_{A_1}}\,q^{n_{A_2}},$$

where n_{A_1} and n_{A_2} are the number of A_1 and A_2 alleles among founder individuals. Given the prior density $Be\,(q|e,f)$ in (16.3), (16.10) becomes

$$p\,(q|\mathbf{W},\mathbf{y}) \propto (1-q)^{\left(n_{A_1}+f-1\right)}\,q^{\left(n_{A_2}+e-1\right)},$$

which is the kernel of a beta distribution with parameters $n_{A_1}+f$ and $n_{A_2}+e$. That is,

$$q|\mathbf{W},\mathbf{y} \sim Be\,(n_{A_1}+f, n_{A_2}+e)\,. \tag{16.11}$$

If a uniform prior $Un\,(0,1)$ is assumed for q, instead of (16.3), the fully conditional posterior distribution for q has density

$$q|\mathbf{W},\mathbf{y} \sim Be\,(q|n_{A_1},n_{A_2})\,.$$

16.2.3 Some Implementation Issues

It was discussed in Chapter 10 that a finite-state space discrete Markov chain converges to a unique stationary distribution, provided it is irreducible and aperiodic. The sampling scheme for the major locus genotype

defines a discrete chain. Starting from a particular "legal" genotypic configuration, one wishes to know whether it is possible to visit any other legal configuration by updating individual genotypes one at a time.

In the case of a diallelic locus with alleles A_1 and A_2, Sheehan and Thomas (1993) show that irreducibility of the Markov chain is guaranteed, provided that, if for all y satisfying

$$p\,(y|A_1A_1)\,p\,(y|A_2A_2) > 0, \tag{16.12}$$

it also holds that

$$p\,(y|A_1A_2) > 0. \tag{16.13}$$

In other words, irreducibility is established unless the genetic model allows for a phenotype compatible with both homozygous genotypes, to be incompatible with the heterozygote state. This is the only situation in which the Gibbs sampler may define a reducible Markov chain for a diallelic system.

To illustrate, consider the following example taken from Sheehan (2000). Individuals are classified into "affected" or "normal", depending on whether they are homozygotes or heterozygotes, respectively. Data are available on a mother-daughter pair, where both are "affected". If the starting legal configuration assigns genotype A_1A_1 to the mother, then the updating scheme based on the genotypic distribution of the daughter conditional on the mother's genotype assigns probability 1 to the event "genotype of daughter is A_1A_1". The daughter can never change to the other homozygote genotype, given that the mother has genotype A_1A_1.

With more than two alleles at a single locus, irreducibility is no longer ensured by imposing a simple condition such as that defined by (16.12) and (16.13). Reducibility depends on the data and method of sampling. For example, consider the following human ABO blood group example, taken from Sheehan and Thomas (1993). The data ($\mathbf{y}$) consist of the genotypes of two offspring, which are A_1B and OO, respectively. This implies that the genotype of the parents must be (A_1O, BO) or (BO, A_1O). Consider an updating scheme consisting of drawing from the conditional posterior distribution of a parent, given the genotype of the other parent and the data. For example, once the father's genotype (f) has been assigned (A_1O, say), the maternal genotype (m) is forced to be the complementary type (BO) with probability 1. Given the data and the sampling mechanism, this updating scheme creates a Markov chain with two noncommunicating states.

The way around this toy problem is to sample the parental genotypes jointly, in one block, from the joint distribution

$$\Pr\,(f = A_1O, m = BO|\mathbf{y}) = \frac{1}{2},$$

$$\Pr\,(f = BO, m = A_1O|\mathbf{y}) = \frac{1}{2}.$$

These examples emphasize that care must be taken about inferences derived from a single site updating Gibbs scheme, in multiallelic systems. In fact, there is at present no method that can guarantee irreducibility in large complex pedigrees. More importantly, even if the chain is irreducible, mixing may be slow; consequently, a limited range of the support of the posterior distribution may be visited by the Monte Carlo draws. This will result in poor inferences. Slow mixing of the chain affects convergence of the time-average of draws to the expectation of the function under the equilibrium distribution, even if the chain starts in the equilibrium distribution.

Various methods for dealing with reducibility and slow mixing have been proposed in recent years, and many of these are reviewed by Sheehan (2000) and by Thompson (2001). Several of the approaches are based on a variety of joint-updating schemes. For example, Jensen et al. (1995) update genotypes of blocks of individuals jointly at several loci. The method was extended by Lund and Jensen (1999) to cope with the mixed inheritance model. Janss et al. (1995) also propose a blocking scheme, whereby a sire with all its final offspring are updated in one pass. Other forms of joint updating were discussed by Heath (1997), Thompson and Heath (2000) and Sheehan et al. (2002). In general, the joint updating sampling strategies are a great improvement over single-site methods.

16.3 Models for the Detection of Quantitative Trait Loci

In the last decade, and due to the availability of information on molecular markers, there has been much interest in detecting chromosomal regions responsible for some of the variation observed for quantitative traits. In particular, an extensive literature on methods for detecting QTL using marked regions has accumulated. The objective here is to provide an introduction to statistical aspects of the subject and to illustrate how likelihood or MCMC-based methods can be applied for drawing inferences concerning effects and positions of such QTL. The presentation is restricted to the analysis of a backcross design involving inbred lines; full marker information in a single chromosome is assumed.

First, a model with a single QTL flanked by two marker loci is introduced. This is then extended to models with an arbitrary number of QTL using information on a large number of genetic markers. In this final section, ways in which models can be compared using their posterior probabilities are outlined.

16.3.1 Models with a Single QTL

In the single QTL model, the marker at the first locus has alleles M and m, and at the second locus, the marker alleles are N and n. At each locus, markers are assumed to be codominant. Interest focuses on whether there is evidence for the presence of a QTL, with unknown alleles Q and q, placed between the two (known) markers. That is, it is assumed here that the locus order is MQN. (The frequency of the QTL alleles does not feature in this section; therefore the symbol q in this section is always associated with the QTL allele, and not with allele frequency).

There may also be genetic variation contributed by genes of small effect. Due to the nature of the experimental design, this variation cannot be estimated and is part of the residual variance.

The recombination fraction between locus M and the QTL will be denoted by r_m, and the recombination fraction between the QTL and locus N by r_n. It is assumed that recombination in the M-Q interval is independent of recombination in the Q-N interval. Therefore the probability of a double recombinant is $r_m r_n$. Another simplifying assumption is that the recombination fraction between the two markers, r, is known, and that it is less than 0.5. A recombination occurs, if the number of crossovers between the loci involved is odd. Recombination between M and N, which is the probability of an odd number of crossovers, arises as follows. A recombination occurs between M and Q and not between Q and N, or a recombination occurs between Q and N and not between M and Q. With no interference, the probability of recombination is the sum of the probabilities of these two mutually exclusive events. Thus,

$$r = r_m\,(1 - r_n) + (1 - r_m)\,r_n = r_m + r_n - 2r_m r_n$$

(e.g., Ott, 1999), which implies that

$$r_m = (r - r_n)/\,(1 - 2r_n)\,, \quad 0 < r_m < r,\ 0 < r_n < r. \tag{16.14}$$

Thus, with r known, there is only one recombination fraction to be estimated, because r_m can be written as a function of r_n or vice-versa.

It is often convenient to parameterize the model in terms of genetic map distances rather than in terms of recombination fractions. The genetic map distance between two loci is defined as the expected number of crossovers occurring on a given chromosome (in a gamete) between the loci (Ott, 1999). Genetic map distances are expressed in centimorgans (cM). The advantage of genetic map distances is that these are additive, whereas recombination fractions are not. Parameterization with genetic map distances requires mapping functions; there are several such functions (see Ott (1999) for an overview). Here the one proposed by Haldane (1919) is used, which assumes that recombination between any two markers is independent of that occurring at other marker intervals (i.e., no chiasma interference is assumed). Using this mapping function, the distance λ (measured in units of

Morgans, a positive quantity) and the recombination fraction r are related by the equation

$$r = f(\lambda) = \frac{1}{2}\left[1 - \exp(-2\lambda)\right], \tag{16.15}$$

with inverse function

$$\lambda = f^{-1}(r) = -\frac{1}{2}\ln(1 - 2r). \tag{16.16}$$

Thus, if the genetic map distance between loci i and $i+1$ is $|\lambda_{i+1} - \lambda_i|$, the recombination fraction is

$$r = f(\lambda_{i+1} - \lambda_i) = \frac{1}{2}\left[1 - \exp(-2|\lambda_{i+1} - \lambda_i|)\right].$$

Data from a backcross design are assumed to be generated as follows. Consider two completely inbred lines, one with genotype MQN/MQN and the other with genotype mqn/mqn. These lines are crossed to produce F_1 individuals (generation 1), all having genotype MQN/mqn. Notationally, the haplotype to the left of the slanted line in MQN/mqn represents the paternal gamete, and the one on the right (mqn) represents the maternal gamete. The inbred lines are typically chosen on the basis of some phenotypic attribute, as opposed to being randomly sampled.

The F_1 individuals are crossed back to individuals from one of the inbred lines, those carrying genotype MQN/MQN say, to produce the backcross generation (generation 2). The marker genotype for the ith individual of generation 2 is denoted $\mathcal{M}_i$, which can take values $\mathcal{G}_1 = MN/MN$, $\mathcal{G}_2 = Mn/MN$, $\mathcal{G}_3 = mN/MN$, and $\mathcal{G}_4 = mn/MN$. The QTL genotype of individual i from generation 2, $\mathcal{Q}_i$, is a random variable which can take values $\mathcal{Q}_i = QQ$ or $\mathcal{Q}_i = Qq$.

In a backcross design, individuals are genetically uncorrelated. To see this, let a_i and a_j denote additive genetic values of individuals i and j randomly sampled from generation 2. These additive genetic values arise due to the presence of many genes each of small effect acting additively on the genotype. If the original lines are completely inbred, there is no variation among the paternal additive genetic values (denoted as a_f) nor among the maternal additive genetic values (a_m) of individuals i and j. Then,

$$\begin{aligned} &Cov(a_i, a_j) \\ &= E\left[Cov(a_i, a_j | a_f, a_m)\right] + Cov\left[E(a_i | a_f, a_m), E(a_j | a_f, a_m)\right] \\ &= 0 + Cov\left[\frac{1}{2}(a_f + a_m), \frac{1}{2}(a_f + a_m)\right] = \frac{1}{4}Var(a_f + a_m) = 0, \\ &\qquad i \neq j. \end{aligned}$$

The term $E\left[Cov(a_i, a_j | a_f, a_m)\right]$ is zero because, given the additive genetic values of the parents, additive genetic values in the offspring are uncorrelated. Absence of genetic variation in the parents implies absence of genetic

covariation in the offspring, despite there being genetic variation among the latter.

Conditionally on the unknown QTL genotype, it is assumed that the phenotypic record of the ith individual is normal of the form

$$y_i|\mathcal{Q}_i = QQ \sim N\left(\mu_1, \sigma^2\right), \tag{16.17}$$

$$y_i|\mathcal{Q}_i = Qq \sim N\left(\mu_2, \sigma^2\right), \tag{16.18}$$

where μ_1 is the mean phenotype of individuals carrying QTL genotype QQ, and μ_2 is the mean phenotype of individuals carrying QTL genotype Qq. The variance term σ^2 typically contains a contribution from polygenic effects at many other loci.

Likelihood Inference

Writing the Likelihood

The parameters of interest are $\boldsymbol{\theta}' = \left(\mu_1, \mu_2, r_m, \sigma^2\right)$. The data consist of records on a particular trait, represented by the vector $\mathbf{y}$ of length n_{obs} and by information on the marker genotypes from generation 2. The likelihood is proportional to the density of the data, given marker information, which for record i is denoted by $p\left(y_i|\boldsymbol{\theta}, \mathcal{M}_i\right)$. The contribution to the likelihood from the record of individual i can be written as

$$\begin{aligned} &L\left(\boldsymbol{\theta}|y_i, \mathcal{M}_i\right) \\ &\propto p\left(y_i|QQ, \mathcal{M}_i\right) \Pr\left(QQ|\mathcal{M}_i, r_m\right) + p\left(y_i|Qq, \mathcal{M}_i\right) \Pr\left(Qq|\mathcal{M}_i, r_m\right) \\ &= p\left(y_i|QQ\right) \Pr\left(QQ|\mathcal{M}_i, r_m\right) + p\left(y_i|Qq\right) \Pr\left(Qq|\mathcal{M}_i, r_m\right), \end{aligned} \tag{16.19}$$

which is a mixture of normal distributions. (Notationally, terms of the form $p\left(y_i|\mathcal{Q}_i = QQ, \mathcal{M}_i\right)$ or $\Pr\left(\mathcal{Q}_i = QQ|\mathcal{M}_i, r_m\right)$ say, are written here as

$$p\left(y_i|QQ, \mathcal{M}_i\right),$$

or

$$\Pr\left(QQ|\mathcal{M}_i, r_m\right),$$

respectively). The equality in the second line of (16.19) arises because, given the QTL genotype, the marker genotype does not contribute with additional information to the probability of observing phenotype y_i. In view of the fact that records are independently distributed, and denoting the complete marker information by the vector $\mathcal{M}$, the likelihood is given by

$$L\left(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}\right) \propto \prod_{i=1}^{n} L\left(\boldsymbol{\theta}|y_i, \mathcal{M}_i\right). \tag{16.20}$$

In (16.19), the terms $\Pr\left(QQ|\mathcal{M}_i\right)$ and $\Pr\left(Qq|\mathcal{M}_i\right)$ are the conditional probabilities of observing a particular QTL genotype given marker information. As shown below, these are functions of the recombination fractions.

Genotype	Pr (Genotype)	Marker	QTL
MqN/MQN	$r_m r_n/2$	MN/MN	Qq
mQn/MQN	$r_m r_n/2$	mn/MN	QQ
Mqn/MQN	$r_m\left(1-r_n\right)/2$	Mn/MN	Qq
mQN/MQN	$r_m\left(1-r_n\right)/2$	mN/MN	QQ
MQn/MQN	$r_n\left(1-r_m\right)/2$	Mn/MN	QQ
mqN/MQN	$r_n\left(1-r_m\right)/2$	mN/MN	Qq
MQN/MQN	$\left(1-r_m\right)\left(1-r_n\right)/2$	MN/MN	QQ
mqn/MQN	$\left(1-r_m\right)\left(1-r_n\right)/2$	mn/MN	Qq

TABLE 16.2. Distribution of genotypes in the backcross design.

The genotypic distribution among generation 2 individuals, together with the observed marker information and the associated putative QTL genotypes, are shown in Table 16.2. It is emphasized that Haldane's mapping function, which assumes no interference, is assumed for these calculations. The marker genotype is observed in a backcross offspring, but the genotype (first column) is unknown, because the QTL genotype is not observed. Given that marker genotypes and phase of parents are known, the probability of a given genotype (Column 2) in the offspring can be readily calculated.

To illustrate, consider the term $\Pr\left(QQ|\mathcal{M}_i=\mathcal{G}_3,r_m\right)$ where, for individual i, the observed marker genotype is $\mathcal{G}_3 = mN/MN$, say. This is computed as

$$\Pr\left(QQ|\mathcal{M}_i=\mathcal{G}_3,r_m\right)=\Pr\left(QQ,\mathcal{M}_i=\mathcal{G}_3|r_m\right)/\Pr\left(\mathcal{M}_i=\mathcal{G}_3|r_m\right).$$

The term $\Pr\left(QQ,\mathcal{M}_i=\mathcal{G}_3|r_m\right)$ is equal to $\frac{1}{2}r_m\left(1-r_n\right)$, associated with genotype mQN/MQN in the 4th row in the body of Table 16.2. The denominator

$$\Pr\left(\mathcal{M}_i=\mathcal{G}_3|r_m\right)=\Pr\left(QQ,\mathcal{M}_i=\mathcal{G}_3|r_m\right)+\Pr\left(Qq,\mathcal{M}_i=\mathcal{G}_3|r_m\right)$$

is equal to the sum of the genotype probabilities in rows 4 and 6 in the body of Table 16.2,

$$\frac{1}{2}r_m\left(1-r_n\right)+\frac{1}{2}\left(1-r_m\right)r_n.$$

Therefore,

$$\begin{aligned}\Pr\left(QQ|\mathcal{M}_i=\mathcal{G}_3,r_m\right) &= \frac{r_m\left(1-r_n\right)}{r_m\left(1-r_n\right)+\left(1-r_m\right)r_n}\\ &= \frac{r_m\left(1-r_n\right)}{r}.\end{aligned}$$

Likewise,

$$\Pr\left(Qq|\mathcal{M}_i=\mathcal{G}_3,r_m\right)=\frac{\left(1-r_m\right)r_n}{r}.$$

Marker	$\Pr(QQ\|\mathcal{M}_i, r_m)$	$\Pr(Qq\|\mathcal{M}_i, r_m)$
MN/MN	$(1-r_m)(1-r_n)/(1-r)$	$r_m r_n/(1-r)$
Mn/MN	$(1-r_m)r_n/r$	$r_m(1-r_n)/r$
mN/MN	$r_m(1-r_n)/r$	$(1-r_m)r_n/r$
mn/MN	$r_m r_n/(1-r)$	$(1-r_m)(1-r_n)/(1-r)$

TABLE 16.3. Conditional probabilities of QTL genotypes given marker genotypes in the backcross generation.

Table 16.3 shows the terms $\Pr(QQ|\mathcal{M}_i, r_m)$ and $\Pr(Qq|\mathcal{M}_i, r_m)$ for all possible values of $\mathcal{M}_i$.

If individual i has observed marker information $\mathcal{M}_i = mN/MN$, from (16.19), its contribution to the overall likelihood is

$$L(\boldsymbol{\theta}|y_i, \mathcal{M}_i) \propto p(y_i|QQ)\, r_m(1-r_n) + p(y_i|Qq)(1-r_m)\, r_n. \qquad (16.21)$$

The recombination fractions satisfy (16.14).

Hypotheses Tests

The test for the presence of a QTL between marker loci M and N, versus absence of a QTL segregating, consists of computing the likelihood ratio. This observed likelihood ratio is (Knott and Haley, 1992)

$$LR = \frac{L\left(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma^2}, \widehat{r}_m|\mathbf{y}, \mathcal{M}\right)}{L\left(\widetilde{\mu}, \widetilde{\sigma^2}|\mathbf{y}, \mathcal{M}\right)}, \qquad (16.22)$$

where $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma^2}, \widehat{r}_m$ are the values obtained by joint maximization of

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}),$$

and $\widetilde{\mu}, \widetilde{\sigma^2}$ are the values obtained by maximizing the likelihood restricted by the null hypothesis (no QTL segregating).

Let $T(\mathbf{Y})$ be equal to (16.22), but with the important difference that $T(\mathbf{Y})$ is a function of the random variable $\mathbf{Y}$ rather than of the observed data $\mathbf{y}$, which is a realized value of this random variable. A test of the null hypothesis may require computing the so-called p-value

$$\Pr[T(\mathbf{Y}) \geq LR] = \int I[T(\mathbf{y}) \geq LR]\, p\left(\mathbf{y}|\mu, \sigma^2, \mathcal{M}\right) d\mathbf{y}. \qquad (16.23)$$

In a classical test of hypothesis, the null hypothesis is rejected if (16.23) is smaller than or equal to $\Pr[\textit{Type I error}]$. Because μ and σ^2 are usually not known, (16.23) cannot be computed exactly. In this situation one often appeals to asymptotic results. Under regularity conditions

$$\Pr[2\ln T(\mathbf{Y}) \geq 2\ln LR] = \Pr\left[\chi^2_2 \geq 2\ln LR\right].$$

That is, asymptotically, $2\ln T(\mathbf{Y})$ has a chi-square distribution, with two degrees of freedom in this case.

An alternative to asymptotic theory is to obtain a Monte Carlo approximation to (16.23). Since μ and σ^2 are not known, these are replaced by their maximum likelihood estimates $\widetilde{\mu}, \widetilde{\sigma^2}$. The required probability is now approximated by

$$\Pr[T(\mathbf{Y}) \geq LR] = \int I[T(\mathbf{y}) \geq LR]\, p\left(\mathbf{y}|\widetilde{\mu}, \widetilde{\sigma^2}, \mathbf{M}\right) d\mathbf{y}. \tag{16.24}$$

This integration can be approximated drawing data vectors $\mathbf{y}_i$ $(i = 1, \ldots, N)$ from $p\left(\mathbf{y}_i|\widetilde{\mu}, \widetilde{\sigma^2}, \mathbf{M}\right)$ and computing

$$\Pr[T(\mathbf{Y}) \geq LR] \approx \frac{1}{N}\sum_i I(T(\mathbf{y}_i) \geq LR), \tag{16.25}$$

where N represents the number of samples drawn. Other suggested procedures are based on permutation tests (Doerge and Churchill, 1996).

Rather than maximizing (16.22), a profile likelihood is often computed, whereby $\log_{10}$ of the ratios of the form in (16.22) are calculated over a grid of values of r_m. The term $2\log_{10}$ is known as the LOD score (see Ott, 1999 for a detailed discussion). The maximum LOD score indicates the grid value of r_m closest to the maximum likelihood estimate of r_m. A smooth curve is fitted to the set of LOD score values, and a measure of uncertainty, in conceptual repeated sampling, is obtained by a $2(LOD)$ interval. This interval is the set of values of r_m at which the LOD is not smaller than its maximum value minus two. The LOD score can be multiplied by a factor $2\ln(10) = 4.605$ for it to have the convenient property of being asymptotically distributed as a chi-square random variable.

We end this section with a word of caution. Setting up the correct test of hypothesis in a conventional likelihood scenario is a contentious issue. A test often entertained is based on

$$\frac{L\left(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma^2}, \widehat{r}_m|\mathbf{y}, \mathbf{M}\right)}{L\left(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\sigma^2}, r_m = 0.5|\mathbf{y}, \mathbf{M}\right)}.$$

The null hypothesis assumes the absence of a QTL between the region flanked by the markers, but allows for the fact that a QTL may be present elsewhere in the genome. This form of the likelihood ratio places the parameter $r_m \in [0, 0.5]$ at the boundary of the parameter space and, as a consequence, the necessary regularity conditions associated with classical asymptotic theory are not satisfied. Other possible tests are discussed in Knott and Haley (1992).

Bayesian Inference

The analysis in the previous section is now performed from a Bayesian perspective. Except for the additional presence of prior distributions, the model here is similar to the one in the previous section. As before, it is assumed that the genetic distance between the two markers is known.

Let $\mathcal{Q}$ represent the unknown QTL genotype of all individuals and let $\mathcal{Q}_i$ represent the unknown QTL genotype of individual i. The parameters of the Bayesian model are $(\mathcal{Q}, r_m)$ and $\boldsymbol{\theta}' = \left(\mu_1, \mu_2, \sigma^2\right)$.

The prior distributions of the parameters are assumed to be as follows. First, the elements of $\boldsymbol{\theta}$ are taken to be a priori independently distributed. As prior for μ_i a normal distribution is invoked, with zero mean and variance 10, say, to allow for large QTL effects; thus $\mu_i \sim N\left(0, \sigma_0^2 = 10\right)$, $(i = 1, 2)$. The prior for the residual variance is assumed to be a scaled inverted chi-square distribution with known parameters ν and S; that is, $\sigma^2 \sim \nu S \chi_\nu^{-2}$. It is also assumed that $(\mathcal{Q}, r_m)$ and $\boldsymbol{\theta}$ are a priori independently distributed.

Prior information about the recombination fraction could be incorporated as follows. Consider the beta distributed random variable with density

$$\eta \sim Be\left(\eta | a, b\right), \quad (0 < \eta < 1),$$

where a and b are known parameters defining the shape of the distribution. The recombination fraction r_m must take positive probability in the set $]0, r[$ (see (16.14)). Now let $r_m = r\eta$. The inverse transformation is equal to $\eta = r^{-1} r_m$, the Jacobian of the transformation is r^{-1}, and therefore the probability density of r_m has the form

$$p\left(r_m | a, b\right) = \begin{cases} C\left(r^{-1} r_m\right)^{a-1} \left(1 - r^{-1} r_m\right)^{b-1} r^{-1}, & 0 < r_m < r, \\ 0, \quad \text{otherwise}, \end{cases} \tag{16.26}$$

where C is a constant that does not depend on r_m.

The posterior distribution is given by

$$\begin{aligned} p\left(\mathcal{Q}, r_m, \boldsymbol{\theta} | \mathbf{y}, \mathcal{M}\right) &\propto p\left(\mathcal{Q}, r_m\right) p\left(\boldsymbol{\theta}\right) p\left(\mathbf{y}, \mathcal{M} | \mathcal{Q}, r_m, \boldsymbol{\theta}\right) \\ &= p\left(\mathcal{Q}, r_m\right) p\left(\boldsymbol{\theta}\right) \Pr\left(\mathcal{M} | \mathcal{Q}, r_m, \boldsymbol{\theta}\right) p\left(\mathbf{y} | \mathcal{M}, \mathcal{Q}, r_m, \boldsymbol{\theta}\right) \\ &\propto p\left(r_m\right) p\left(\mu_1\right) p\left(\mu_2\right) p\left(\sigma^2\right) \Pr\left(\mathcal{Q} | r_m, \mathcal{M}\right) p\left(\mathbf{y} | \mathcal{Q}, \boldsymbol{\theta}\right), \end{aligned} \tag{16.27}$$

where

$$p\left(\mathbf{y} | \mathcal{Q}, \boldsymbol{\theta}\right) = \prod_{i=1}^{n} p\left(y_i | \mathcal{Q}_i, \boldsymbol{\theta}\right), \tag{16.28}$$

and

$$\Pr\left(\mathcal{Q} | r_m, \mathcal{M}\right) = \prod_{i=1}^{n} \Pr\left(\mathcal{Q}_i | r_m, \mathcal{M}_i\right). \tag{16.29}$$

Expressions (16.28) and (16.29) imply that the QTL genotype can be drawn for each individual at a time. This conditional independence property is a consequence of the experimental design. In the backcross design, both the phase and the QTL genotype of parents are known. Therefore, the parental origin of an offspring's gamete can be unambiguously assigned. This is not the case with data from outbred populations.

The joint posterior is defined within the range of values of the parameters. Marginalization of (16.27) with respect to $\mathcal{Q}$, assuming uniform prior distributions for r_m and $\boldsymbol{\theta}$, retrieves a posterior distribution which has the same form as likelihood (16.20) derived from the mixture (16.19).

Fully Conditional Posterior Distributions

Extracting the terms containing μ_1 from (16.27) yields

$$p\left(\mu_1|.,data\right) \propto p\left(\mu_1\right)\prod_{i=1}^{n} p\left(y_i|\mathcal{Q}_i,\mu_1,\sigma^2\right)^{I(\mathcal{Q}_i=QQ)}$$

$$\propto \exp\left(-\frac{\mu_1^2}{2\sigma_0^2}\right)\exp\left[-\frac{\sum_{i=1}^{n} I\left(\mathcal{Q}_i=QQ\right)\left(y_i-\mu_1\right)^2}{2\sigma^2}\right], \tag{16.30}$$

where $I(\mathcal{Q}_i = QQ)$ is the indicator function that takes the value one when the individual is QQ, and zero otherwise. Let $\widehat{\mu}_1 = [\sum_{i=1}^{n} I(\mathcal{Q}_i = QQ)y_i]/n_{QQ}$ be the mean of the records on individuals whose QTL genotype is QQ, where n_{QQ} is the number of such individuals. Then (16.30) can be expressed as

$$p\left(\mu_1|.,data\right)$$

$$\propto \exp\left[-\frac{\sigma^2\mu_1^2+\sigma_0^2\sum_{i=1}^{n} I\left(\mathcal{Q}_i=QQ\right)\left[\left(y_i-\widehat{\mu}_1\right)+\left(\widehat{\mu}_1-\mu_1\right)\right]^2}{2\sigma^2\sigma_0^2}\right]$$

$$\propto \exp\left[-\frac{\sigma^2\mu_1^2+\sigma_0^2 n_{QQ}\left(\widehat{\mu}_1-\mu_1\right)^2}{2\sigma^2\sigma_0^2}\right]. \tag{16.31}$$

Making use of the identity (Box and Tiao, 1973, page 74):

$$A\left(z-a\right)^2+B\left(z-b\right)^2=\left(A+B\right)\left(z-c\right)^2+\frac{AB}{A+B}\left(a-b\right)^2$$

and letting $c=(Aa+Bb)/(A+B)$, $A=\sigma^2$, $B=\sigma_0^2 n_{QQ}$, $z=\mu_1$, $a=0$, and $b=\widehat{\mu}_1$ allows expressing (16.31) as

$$p\left(\mu_1|.,data\right) \propto \exp\left[-\frac{\left(\sigma^2+n_{QQ}\sigma_0^2\right)\left(\mu_1-c\right)^2}{2\sigma^2\sigma_0^2}\right],$$

where $c=n_{QQ}\sigma_0^2\widehat{\mu}_1/(\sigma^2+n_{QQ}\sigma_0^2)$. Therefore,

$$\mu_1|.,data \sim N\left(\frac{n_{QQ}\sigma_0^2\widehat{\mu}_1}{\sigma^2+n_{QQ}\sigma_0^2},\frac{\sigma^2\sigma_0^2}{\sigma^2+n_{QQ}\sigma_0^2}\right). \tag{16.32}$$

By symmetry considerations,

$$\mu_2|.,data \sim N\left(\frac{n_{Qq}\sigma_0^2\widehat{\mu}_2}{\sigma^2+n_{Qq}\sigma_0^2},\frac{\sigma^2\sigma_0^2}{\sigma^2+n_{Qq}\sigma_0^2}\right), \tag{16.33}$$

where n_{Qq} is the number of individuals with QTL genotype Qq, and $\widehat{\mu}_2 = \sum_{i=1}^{n} I(\mathcal{Q}_i = Qq)y_i/n_{Qq}$.

To derive $p\left(\sigma^2|.,data\right)$ the terms containing σ^2 are extracted from the joint posterior (16.27). This yields

$$\begin{aligned}
p\left(\sigma^2|.,data\right) &\propto p\left(\sigma^2\right)\prod_{i=1}^{n}\left[p\left(y_i|\mathcal{Q}_i,\mu_1,\sigma^2\right)^{I(\mathcal{Q}_i=QQ)}\right.\\
&\qquad\left.\times\, p\left(y_i|\mathcal{Q}_i,\mu_2,\sigma^2\right)^{I(\mathcal{Q}_i=Qq)}\right]\\
&\propto p\left(\sigma^2\right)\prod_{i=1}^{n}\left\{\left(\sigma^2\right)^{-\frac{1}{2}}\exp\left[-\frac{I\left(\mathcal{Q}_i=QQ\right)\left(y_i-\mu_1\right)^2}{2\sigma^2}\right]\right.\\
&\qquad\left.\times\exp\left[-\frac{I\left(\mathcal{Q}_i=Qq\right)\left(y_i-\mu_2\right)^2}{2\sigma^2}\right]\right\}.
\end{aligned}$$

Let

$$\mathcal{V}_{QQ}=\frac{\sum_{i=1}^{n}I\left(\mathcal{Q}_i=QQ\right)\left(y_i-\widehat{\mu}_1\right)^2}{n_{QQ}}$$

and

$$\mathcal{V}_{Qq}=\frac{\sum_{i=1}^{n}I\left(\mathcal{Q}_i=Qq\right)\left(y_i-\widehat{\mu}_2\right)^2}{n_{Qq}}.$$

Then $p\left(\sigma^2|.,data\right)$ can be written as

$$\begin{aligned}
p\left(\sigma^2|.,data\right) &\propto \left(\sigma^2\right)^{-\left(\frac{\nu+n}{2}+1\right)}\\
\times\exp\left\{-\frac{n_{QQ}\left[\mathcal{V}_{QQ}+\left(\mu_1-\widehat{\mu}_1\right)^2\right]+n_{Qq}\left[\mathcal{V}_{Qq}+\left(\mu_2-\widehat{\mu}_2\right)^2\right]+\nu S}{2\sigma^2}\right\}&\\
=\left(\sigma^2\right)^{-\left(\frac{\widetilde{\nu}}{2}+1\right)}\exp\left(-\frac{\widetilde{\nu}\widetilde{S}}{2\sigma^2}\right),&
\end{aligned}$$

where $\widetilde{\nu}=\nu+n$ and

$$\widetilde{S}=\left\{n_{QQ}\left[\mathcal{V}_{QQ}+\left(\mu_1-\widehat{\mu}_1\right)^2\right]+n_{Qq}\left[\mathcal{V}_{Qq}+\left(\mu_2-\widehat{\mu}_2\right)^2\right]+\nu S\right\}\Big/\widetilde{\nu}.$$

This is recognized as the density of a scaled inverted chi-square distribution with parameters $\widetilde{S}$ and $\widetilde{\nu}$

$$\sigma^2|., data \sim \widetilde{\nu}\widetilde{S}\chi^{-2}_{\widetilde{\nu}}. \tag{16.34}$$

From the joint posterior (16.27) the fully conditional posterior distribution of the QTL genotypes is proportional to

$$\begin{aligned}\Pr(\mathcal{Q}|., data) &\propto \Pr(\mathcal{Q}|r_m, \mathcal{M})\, p(\mathbf{y}|\mathcal{Q}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \Pr(\mathcal{Q}_i|r_m, \mathcal{M}_i)\, p(y_i|\mathcal{Q}_i, \boldsymbol{\theta}),\end{aligned} \tag{16.35}$$

which implies that sampling can proceed separately for each individual. For individual i

$$\Pr(\mathcal{Q}_i = QQ|., data) = \frac{\Pr(\mathcal{Q}_i = QQ|r_m, \mathcal{M}_i)\, p(y_i|\mathcal{Q}_i = QQ, \boldsymbol{\theta})}{\sum_{\omega} \Pr(\mathcal{Q}_i = \omega|r_m, \mathcal{M}_i)\, p(y_i|\mathcal{Q}_i = \omega, \boldsymbol{\theta})}, \tag{16.36}$$

where ω denotes QQ or Qq. Similarly,

$$\Pr(\mathcal{Q}_i = Qq|., data) = \frac{\Pr(\mathcal{Q}_i = Qq|r_m, \mathcal{M}_i)\, p(y_i|\mathcal{Q}_i = Qq, \boldsymbol{\theta})}{\sum_{\omega} \Pr(\mathcal{Q}_i = \omega|r_m, \mathcal{M}_i)\, p(y_i|\mathcal{Q}_i = \omega, \boldsymbol{\theta})}. \tag{16.37}$$

Finally, the fully conditional posterior distribution of the recombination fraction r_m is proportional to

$$p(r_m|., data) \propto p(r_m) \prod_{i=1}^{n} \Pr(\mathcal{Q}_i|r_m, \mathcal{M}_i). \tag{16.38}$$

In (16.38), $p(r_m)$ is given by (16.26). Expression (16.38) does not have a standard form; therefore a univariate Metropolis-Hastings algorithm can be used for drawing samples r_m. Let r_m^* denote a candidate value generated by the candidate generating density $u(r_m^*|r_m)$. Then the proposal is accepted with probability $\alpha(r_m^*, r_m)$ given by

$$\alpha(r_m^*, r_m) = \begin{cases} \min\left[\frac{p(r_m^*|., data)u(r_m|r_m^*)}{p(r_m|., data)u(r_m^*|r_m)}, 1\right], \text{ if } p(r_m|., data) > 0, \\ 1, \text{ otherwise.} \end{cases} \tag{16.39}$$

The candidate generating density could be a uniform distribution on the interval $(r_m - d, r_m + d)$, where d is chosen such that the acceptance rate is in the range 20% to 50% (Chib and Greenberg, 1995). If a uniform density is chosen, $Un(r_m^*|r_m - d, r_m + d) = Un(r_m|r_m^* - d, r_m^* + d) = 1/2d$. Then, only the ratio $p(r_m^*|., data)/p(r_m|., data)$ needs to be computed in (16.39). This is known as the Metropolis algorithm (Metropolis et al., 1953).

Implementation of the MCMC approach described above generates Monte Carlo samples from the joint posterior distribution (16.27). Specific parameters of the model can be inferred from their marginal posterior distribution.

Model Selection

A Bayesian counterpart of (16.22) is the Bayes factor that was discussed in Chapter 8. One may wish to compare the model that assumes one QTL is segregating, labeled M_1, versus a model that assumes that no QTL is segregating, labeled M_0. Under M_0, the posterior distribution is

$$p\left(\mu, \sigma_e^2|\mathbf{y}, M_0\right) \propto p\left(\mu, \sigma_e^2|M_0\right) p\left(\mathbf{y}|\mu, \sigma_e^2, M_0\right).$$

The Bayes factor of model M_1 relative to model M_0 requires computation of

$$\begin{aligned} B_{10} &= \frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_0\right)} \\ &= \frac{\sum\limits_{\mathcal{Q}} \int p\left(\mathbf{y}|\mathcal{Q}, \boldsymbol{\theta}, M_1\right) \Pr\left(\mathcal{Q}|r_m, \mathcal{M}, M_1\right) p\left(r_m, \boldsymbol{\theta}|M_1\right) d\boldsymbol{\theta}}{\int p\left(\mathbf{y}|\mu, \sigma_e^2, M_0\right) p\left(\mu, \sigma_e^2|M_0\right) d\mu d\sigma_e^2}. \end{aligned}$$

As discussed in Chapter 8, B_{10} can be computed using standard MCMC output with the approach suggested by Newton and Raftery (1994). A Monte Carlo estimate of B_{10} is given by

$$\widehat{B}_{10} = \frac{\sum\limits_{j=1}^{N} p^{-1}\left(\mathbf{y}|\mu^{(j)}, \sigma_e^{2(j)}, M_0\right)}{\sum\limits_{j=1}^{N} p^{-1}\left(\mathbf{y}|\mathcal{Q}^{(j)}, \boldsymbol{\theta}^{(j)}, M_1\right)},$$

where N is the length of the Monte Carlo chain, $\mu^{(j)}, \sigma_e^{2(j)}$ is the jth draw from $\left[\mu, \sigma_e^2|\mathbf{y}, M_0\right]$, and $\mathcal{Q}^{(j)}, \boldsymbol{\theta}^{(j)}$ is the jth draw from $[\mathcal{Q}, \boldsymbol{\theta}|\mathbf{y}, M_1]$.

Another approach to Bayesian model choice is based on a Monte Carlo estimate of the posterior probability of the model; this is discussed at the end of the next section.

As a final word of warning, we remind the reader that tests involving specific values of continuous parameters ($r = 0.5$, say), must build on a prior specification which assigns probability mass to that specific value. Otherwise, by definition, the posterior probability that the continuous parameter takes the specified value is 0. A point null hypothesis cannot be tested under a continuous prior distribution. For example, the prior (16.26) cannot be used for testing a specific value of the recombination fraction.

16.3.2 Models with an Arbitrary Number of QTL

In this section the model is extended to include an arbitrary number of QTL. A model indicator M is introduced, which can take values $1, 2, \ldots, I$, where I is an integer. Inference is restricted to the Bayesian approach only. Suppose that $(1, 2, \ldots, K)$ ordered markers and phenotypic observations y_i,

$(i = 1, \ldots, n_{obs})$ are available from a backcross line. The marker genotype information for the K loci of individual i is denoted $\mathcal{M}_i = \{\mathcal{M}_{ij}\}_{j=1,\ldots,K}$, where $\mathcal{M}_{ij}$ is the information on the jth marker. The known positions of the K markers are collected in the vector $D = \{D_l\}_{l=1,\ldots,K}$, where D_l is the genetic map distance between markers 1 and l and $D_1 = 0$.

Suppose that m QTL are present at locations $\lambda = (\lambda_1, \ldots, \lambda_m)$, where $D_1 < \lambda_i < D_K$. More than one QTL may be present in the region defined by two flanking markers. Let the matrix

$$\mathcal{Q} = \{\mathcal{Q}_{ij}\}_{i=1,\ldots,n_{obs},j=1,\ldots,m}$$

represent a genotype configuration where $\mathcal{Q}_{ij}$ is the QTL genotype at location λ_j for individual i. Let the ith row of $\mathcal{Q}$, $\mathcal{Q}_i = \{\mathcal{Q}_{ij}\}_{j=1,\ldots,m}$, represent the QTL genotypes for individual i, and let $\mathcal{Q}^j = \{\mathcal{Q}_{ij}\}_{i=1,\ldots,n_{obs}}$ represent the jth column of $\mathcal{Q}$ with QTL genotypes for the n_{obs} individuals at location λ_j.

In the backcross design, conditionally on marker information $\mathcal{M}$ and D and on number and QTL locations, the QTL genotypes of individuals are independent, so

$$\Pr(\mathcal{Q}|m, \mathcal{M}, \lambda, D) = \prod_{i=1}^{n_{obs}} \Pr(\mathcal{Q}_i|m, \mathcal{M}_i, \lambda, D). \tag{16.40}$$

Following Sillanpää and Arjas (1998), the generic term "object" will be used for any marker or QTL in the linkage group. Let $\mathcal{G}^j_{i,L}$ and $\mathcal{G}^j_{i,R}$ represent the genotypes of two flanking objects (marker or QTL) in individual i, located, respectively, to the left and right of the jth QTL in individual i. These flanking objects represent a subset of the parameters one conditions on in the conditional probability distribution of the QTL genotypes of individual i. In the case of the present backcross design, the conditional probability distribution of the QTL genotypes of individual i, $\Pr(\mathcal{Q}_i|m, \mathcal{M}_i, \lambda, D)$, given the QTL locations and the genotypes and locations of other objects (markers or QTL), can be written as

$$\Pr(\mathcal{Q}_i|m, \mathcal{M}_i, \lambda, D) = \prod_{j=1}^{m} \Pr\left(\mathcal{Q}_{ij}|\mathcal{G}^j_{i,L}, \mathcal{G}^j_{i,R}, \lambda_{j-1}, \lambda_j, \lambda_{j+1}\right), \tag{16.41}$$

where λ_{j-1} is the map distance between $\mathcal{G}^j_{i,L}$ and D_1, λ_j is the map distance between QTL genotype $\mathcal{Q}_{ij}$ and D_1, and λ_{j+1} is the map distance between $\mathcal{G}^j_{i,R}$ and D_1. In order to generate the correct joint prior distribution from (16.41), it is important to choose the objects judiciously. If the object is a marker, it is chosen among the complete set of markers in the chromosome; if it is a QTL, it is chosen among the set of QTL whose index is lower than the one being considered (Sillanpää and Arjas, 1998).

It is also possible to specify the conditional distribution of a particular QTL genotype across the n_{obs} individuals, given the genotypic configuration for the remaining QTL genotypes, the QTL locations λ, and the marker information $\mathcal{M}$. For the present design,

$$\Pr\left(\mathcal{Q}^j|\mathcal{Q}^1,\ldots,\mathcal{Q}^{j-1},\mathcal{M},\lambda,D\right)=\prod_{i=1}^{n_{obs}}\Pr\left(\mathcal{Q}_{ij}|\mathcal{G}_{i,L}^j,\mathcal{G}_{i,R}^j,\lambda_{j-1},\lambda_j,\lambda_{j+1}\right), \tag{16.42}$$

which leads to the alternative form for $\Pr\left(\mathcal{Q}|m,\mathcal{M},\lambda,D\right)$

$$\begin{aligned}\Pr\left(\mathcal{Q}|m,\mathcal{M},\lambda,D\right)&=\prod_{j=1}^{m}\Pr\left(\mathcal{Q}^j|\mathcal{Q}^1,\ldots,\mathcal{Q}^{j-1},\mathcal{M},\lambda,D\right)\\&=\prod_{j=1}^{m}\prod_{i=1}^{n_{obs}}\Pr\left(\mathcal{Q}_{ij}|\mathcal{G}_{i,L}^j,\mathcal{G}_{i,R}^j,\lambda_{j-1},\lambda_j,\lambda_{j+1}\right).\end{aligned} \tag{16.43}$$

In the backcross design, at each of the m loci only two genotypes are possible, with effects represented by real parameters μ_{j1} or μ_{j2}, $(j=1,\ldots,m)$. The data $\mathbf{y}=\{y_i\}$ are assumed to be a realized value from $\mathbf{Y}$, where

$$\mathbf{Y}|m,\boldsymbol{\mu},\mathcal{Q},\sigma^2\sim N\left(\mathbf{X}\boldsymbol{\mu},\mathbf{I}\sigma^2\right). \tag{16.44}$$

In (16.44), $\boldsymbol{\mu}=\left\{\mu_{ji}\right\}_{j=1,\ldots,m,i=1,2}$, $\mathbf{X}$ is an incidence matrix associating QTL effects to the data, $\mathbf{I}$ is the identity matrix, and $\sigma^2>0$ is a residual variance which may include a polygenic contribution from other loci affecting the trait.

It will be assumed that the prior distribution of the parameters can be factorized as follows (ignoring the dependence on hyperparameters to simplify notation)

$$\begin{aligned}p\left(\boldsymbol{\mu},\sigma^2,\lambda,\mathcal{Q},m|\mathcal{M},D\right)&=\Pr\left(m\right)p\left(\boldsymbol{\mu},\sigma^2,\lambda,\mathcal{Q}|m,\mathcal{M},D\right)\\&=\Pr\left(m\right)p\left(\boldsymbol{\mu}|m\right)p\left(\sigma^2|m\right)\Pr\left(\mathcal{Q}|m,\mathcal{M},\lambda,D\right)p\left(\lambda|m,D\right).\end{aligned} \tag{16.45}$$

The posterior distribution is given by

$$\begin{aligned}p\left(\boldsymbol{\mu},\sigma^2,\lambda,\mathcal{Q},m|\mathbf{y},\mathcal{M},D\right)&\propto\Pr\left(m\right)p\left(\boldsymbol{\mu},\sigma^2,\lambda,\mathcal{Q}|m,\mathcal{M},D\right)\\&\quad\times p\left(\mathbf{y}|m,\boldsymbol{\mu},\mathcal{Q},\sigma^2\right).\end{aligned} \tag{16.46}$$

The prior distribution of the parameters of the model could be specified as follows. Given $M=m$, the locations $\lambda_1,\ldots,\lambda_m$ are assumed to be independent and uniformly distributed in the interval $\Delta=(D_1,D_K)$. For m, the number of QTL, a Poisson distribution with mean α can be posited. The elements of $\boldsymbol{\mu}$ can be assumed to be a priori independently and normally distributed with mean zero and variance σ_0^2. Finally, the residual variance can be assumed to follow a scaled inverted chi-square distribution with known parameters ν and S: $\sigma^2\sim\nu S\chi_\nu^{-2}$.

The posterior distribution (16.46) does not have a fixed dimension because m varies according to the unknown number of QTL. Reversible jump MCMC provides a flexible method for drawing samples from such a posterior distribution. The algorithm consists of moves that lead to a change of dimension and thereby a change of model, increasing or decreasing the number of QTL, and of updates within models. Updates within a model are very similar to those discussed in connection with the model assuming one QTL and two flanking markers and are briefly dealt with first.

Fully Conditional Posterior Distributions with Fixed Number of QTL

Updating $\boldsymbol{\mu}$

From (16.45) and (16.46) the fully conditional posterior density of $\boldsymbol{\mu}$ is

$$\begin{aligned} p\left(\boldsymbol{\mu}|.,data\right) &\propto p\left(\boldsymbol{\mu}|m\right) p\left(\mathbf{y}|m,\boldsymbol{\mu},\mathcal{Q},\sigma^2\right) \\ &\propto \exp\left(-\frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{2\sigma_0^2}\right) \exp\left[-\frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right)}{2\sigma^2}\right]. \end{aligned} \tag{16.47}$$

The quadratic term, viewed as a function of $\boldsymbol{\mu}$, can be shown to be proportional to

$$\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right) \propto \left(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\right),$$

where

$$\widehat{\boldsymbol{\mu}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

Substituting in (16.47) and combining quadratic forms as indicated in Box and Tiao (1973), page 418, it can be shown that

$$p\left(\boldsymbol{\mu}|.,data\right) \propto \exp\left[-\frac{\left(\boldsymbol{\mu}-\mathbf{c}\right)'\left(\mathbf{I}\sigma^2+\mathbf{X}'\mathbf{X}\sigma_0^2\right)\left(\boldsymbol{\mu}-\mathbf{c}\right)}{2\sigma^2\sigma_0^2}\right],$$

where

$$\mathbf{c} = \left(\mathbf{I}\sigma^2+\mathbf{X}'\mathbf{X}\sigma_0^2\right)^{-1}\sigma_0^2\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\mu}}.$$

Therefore,

$$\boldsymbol{\mu}|.,data \sim N\left(\mathbf{c},\left(\mathbf{I}\sigma^2+\mathbf{X}'\mathbf{X}\sigma_0^2\right)^{-1}\sigma_0^2\sigma^2\right). \tag{16.48}$$

Elements of $\boldsymbol{\mu}$ have the same form as (16.32) and (16.33).

Updating σ^2

The fully conditional posterior distribution of σ^2 is given by

$$p\left(\sigma^2|.,data\right) \propto p\left(\sigma^2\right) p\left(\mathbf{y}|m,\boldsymbol{\mu},\mathcal{Q},\sigma^2\right),$$

which is in the form of a scaled inverted chi-square distribution

$$\sigma^2|.,data \sim \widetilde{v}\widetilde{S}\chi_{\widetilde{v}}^{-2} \tag{16.49}$$

with $\widetilde{v} = \nu + n$ and $\widetilde{S} = \left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\right)+\nu S\right]/\widetilde{v}$.

Updating $\mathcal{Q}$

The conditional independence of (16.43) can be exploited to update $\mathcal{Q}$, sampling each genotype separately, one individual at a time. The fully conditional posterior distribution is

$$\Pr\left(\mathcal{Q}|.,data\right) \propto \prod_{j=1}^{m}\prod_{i=1}^{n_{obs}} \Pr\left(\mathcal{Q}_{ij}|\mathcal{G}_{i,L}^{j},\mathcal{G}_{i,R}^{j},\lambda_{j-1},\lambda_{j},\lambda_{j+1}\right) p\left(y_{ij}|m,\mu_{j1},\mu_{j2}\right) \quad (16.50)$$

and drawing from it can be performed after appropriate scaling. The form of (16.50) resembles (16.36) and (16.37).

Updating λ

Elements of λ can be updated one at a time using a Metropolis-Hastings step. The fully conditional posterior distribution is

$$p\left(\lambda_j|.,data\right) \propto p\left(\lambda_j|m,D\right) \prod_{j=1}^{m} \Pr\left(\mathcal{Q}_{ij}|\mathcal{G}_{i,L}^{j},\mathcal{G}_{i,R}^{j},\lambda_{j-1},\lambda_{j},\lambda_{j+1}\right).$$

Following a similar strategy used in (16.39), the proposal λ_j^* could be generated from a uniform distribution with density $Un\left(\lambda_j^*|a,b\right)$ where $a=\max\left(\lambda_{j-1},\lambda_j-d\right)$, $b=\min\left(\lambda_{j+1},\lambda_j+d\right)$, and d is a tuning parameter. This proposal is accepted with probability

$$\alpha\left(\lambda_j^*,\lambda_j\right)=\begin{cases} \min\left[\frac{p\left(\lambda_j^*|.,data\right)Un(\lambda_j|a,b)}{p(\lambda_j|.,data)Un\left(\lambda_j^*|a,b\right)},1\right], \text{ if } p\left(\lambda_j|.,data\right)>0, \\ 1, \text{ otherwise.}\end{cases}$$

The proposal distribution maintains ordering of the loci.

Updates Leading to Changes of the Model

This section derives the acceptance probability for a reversible jump MCMC algorithm for moves involving a change in the number of QTL. The approach builds on that in Section 11.7 of Chapter 11, where the acceptance probability was derived and illustrated for models with continuous parameters. The discrete nature of the QTL genotypes calls for some minor modifications. The material is largely taken from Waagepetersen and Sorensen (2001). Applications of reversible jump in QTL studies can be found in Heath (1997), Uimari and Hoeschele (1997), Stephens and Fisch (1998), Sillanpää and Arjas (1998, 1999), George et al. (2000), Lee and Thomas (2000), and Yi and Xu (2000). In this last part, following the notation used in Chapter 11, no distinction is made among vectors, matrices, and scalars, unless otherwise stated.

Suppose that the current state of the Markov chain has $M = m$ QTL. With probability $p_{m,m}$ it is proposed to update parameters within the current model, with probability $p_{m,m-1}$ it is proposed to decrease the number of QTL by one, and with probability $p_{m,m+1}$ it is proposed to increase the number of QTL by one (if it is proposed to decrease the number of QTL the chain remains at the current state if $m = 0$).

Given m, define

$$z = \left(\mu_{11}, \mu_{12}, \lambda_1, \mathcal{Q}^1, \ldots, \mu_{m1}, \mu_{m2}, \lambda_m, \mathcal{Q}^m\right).$$

In dimension changing moves, the residual variance σ^2 is not updated; it is updated in moves within models. Therefore to economize on notation, σ^2 is omitted from the formulas in the remaining of this section. Let (m, z) represent the state of the Markov chain, where $z = (z_1, \ldots, z_m)$ is a vector of QTL configurations $z_j = \left(\mu_{j1}, \mu_{j2}, \lambda_j, \mathcal{Q}^j\right)$ $(j = 1, \ldots, m)$. Thus, each QTL configuration z_j consists of the QTL effects, together with the associated QTL location and n_{obs} genotypes. A QTL configuration belongs in the space $C_{conf} = \mathbb{R}^2 \times \Delta \times \{0, 1\}^{n_{obs}}$, where 0 and 1 are labels for genotypes Qq and QQ. The vector z of the m QTL configurations belongs in the space $C_m = C_{conf}^m$.

Removal of a QTL

Suppose there are $m \geq 1$ QTL configurations and that this number is proposed to be reduced by 1 with probability $p_{m,m-1}$. The current state of the Markov chain is $X_n = (m, z)$ where $z = (z_1, \ldots, z_m)$. A move reducing the number of QTL may be accomplished by deterministically removing the last (in terms of the position in the vector z, and not in terms of the physical position of the locus on the chromosome) QTL configuration in z. As in Section 11.7.2 of Chapter 11, denote the proposal $Y_{n+1} = \left(Y_{n+1}^{ind}, Y_{n+1}^{par}\right)$, where $Y_{n+1}^{ind} = m - 1$ and $Y_{n+1}^{par} = g_{1m,m-1}(z_1, \ldots, z_{m-1}) = (z_1, \ldots, z_{m-1})$, since $g_{1m,m-1}$ is the identity mapping. Suppose that A_m is a subset of C_m and that B_{m-1} is a subset of C_{m-1}. The left-hand side of the reversibility condition (11.67) is

$$\begin{aligned} & P\left(M_n = m, Z_n \in A_m, Y_{n+1}^{ind} = m - 1, Y_{n+1}^{par} \in B_{m-1}\right. \\ & \quad \left. \text{and } Y_{n+1} \text{ accepted}\right) \\ = & \int_{\Delta^m} \int_{\mathbb{R}^{2m}} \sum_{\mathcal{Q} \in \{0,1\}^{m n_{obs}}} \mathsf{Q}_{m,m-1}^{a}(z, B_{m-1})\, I(z \in A_m) \\ & \times f(m, z|\mathbf{y})\, d\boldsymbol{\lambda}\, d\boldsymbol{\mu}, \end{aligned} \tag{16.51}$$

where

$$f(m, z|\mathbf{y}) \propto \Pr(M = m|\mathbf{y})\, p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mathcal{Q}|m, \mathbf{y})$$

and

$$\begin{aligned}&\mathsf{Q}^a_{m,m-1}(z, B_{m-1})\\ &= P\left(Y^{ind}_{n+1} = m-1, Y^{par}_{n+1} \in B_{m-1}, Y_{n+1} \text{ accepted} | X_n = (m,z)\right). \quad (16.52)\end{aligned}$$

Since the proposal Y_{n+1} is generated deterministically, in analogy with (11.77),

$$\begin{aligned}\mathsf{Q}^a_{m,m-1}(z, B_{m-1}) &= p_{m,m-1} I\left((z_1, \ldots, z_{m-1}) \in B_{m-1}\right)\\ &\quad \times a_{m,m-1}\left(z, (z_1, \ldots, z_{m-1})\right). \quad (16.53)\end{aligned}$$

Substituting in (16.51) yields

$$\begin{aligned}&p_{m,m-1} \int_{\Delta^m} \int_{\mathbb{R}^{2m}} \sum_{\mathcal{Q} \in \{0,1\}^{mn_{obs}}} f(m, z|\mathbf{y})\\ &\times I\left(z \in A_m, (z_1, \ldots, z_{m-1}) \in B_{m-1}\right) a_{m,m-1}\left(z, (z_1, \ldots, z_{m-1})\right) d\boldsymbol{\lambda}\, d\boldsymbol{\mu}. \quad (16.54)\end{aligned}$$

In these expressions,

$$\mathcal{Q} = \left(\mathcal{Q}^1, \ldots, \mathcal{Q}^m\right),$$
$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m),$$

and

$$\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \ldots, \mu_{m1}, \mu_{m2}).$$

Further, $d\boldsymbol{\lambda}$ and $d\boldsymbol{\mu}$ are shorthand for

$$d\lambda_1, d\lambda_2, \ldots, d\lambda_m$$

and for

$$d\mu_{11}, d\mu_{12}, \ldots, d\mu_{m1}, d\mu_{m2},$$

respectively.

Addition of a QTL

Suppose now that there are $m-1$ QTL and that this number is to be increased by one with probability $p_{m-1,m}$. The current state of the Markov chain is $X_n = (m-1, z') = (m-1, z_1, \ldots, z_{m-1})$. The right-hand side of the reversibility condition (11.67) is

$$\begin{aligned}&P\left(M_n = m-1, Z_n \in B_{m-1}, Y^{ind}_{n+1} = m, Y^{par}_{n+1} \in A_m\right.\\ &\left.\text{and } Y_{n+1} \text{ accepted}\right)\\ = &\int_{\Delta^{m-1}} \int_{\mathbb{R}^{2(m-1)}} \sum_{\mathcal{Q}' \in \{0,1\}^{(m-1)n_{obs}}} \mathsf{Q}^a_{m-1,m}(z', A_m)\, I\left(z' \in B_{m-1}\right)\\ &\times f\left(m-1, z'|\mathbf{y}\right) d\boldsymbol{\lambda}'\, d\boldsymbol{\mu}' \quad (16.55)\end{aligned}$$

where

$$f\left(m-1,z'|\mathbf{y}\right)\propto \Pr\left(M=m-1|\mathbf{y}\right)p\left(\boldsymbol{\mu}',\boldsymbol{\lambda}',\mathcal{Q}'|m-1,\mathbf{y}\right),$$

and

$$\begin{aligned}&\mathsf{Q}^{a}_{m-1,m}\left(z',A_m\right)\\&=P\left(Y^{ind}_{n+1}=m,Y^{par}_{n+1}\in A_m \text{ and } Y_{n+1} \text{ accepted}|X_n=(m-1,z')\right).\end{aligned}\tag{16.56}$$

In this move, the proposal

$$\begin{aligned}Y^{par}_{n+1}&=g_{1m-1,m}\left(z_1,\ldots,z_{m-1},(\mu_{m1},\mu_{m2},\lambda_m,\mathcal{Q}^m)\right)\\&=\left(z_1,\ldots,z_{m-1},(\mu_{m1},\mu_{m2},\lambda_m,\mathcal{Q}^m)\right)\end{aligned}$$

is generated stochastically: one must draw

$$z_m=(\mu_{m1},\mu_{m2},\lambda_m,\mathcal{Q}^m)$$

from the proposal density $q_{m-1,m}\left(z',z_m\right)$. The elements in z_m are placed immediately after z_{m-1}.

A simple approach to generate z_m could be to draw each QTL effect from $\mu_{m,i}\sim N\left(0,\tau^2\right)$ where τ^2 is properly tuned, the location λ_m from a uniform distribution between D_1 and D_K, and the vector of QTL genotypes from the conditional probability

$$\Pr\left(\mathcal{Q}^m|\mathcal{Q}',\lambda_m,\lambda',D,\mathcal{M},m\right)=\prod_{i=1}^{n_{obs}}\Pr\left(\mathcal{Q}_{im}|\mathcal{G}^m_{i,L},\mathcal{G}^m_{i,R},\lambda_m,\lambda'\right),$$

where $\mathcal{G}^m_{i,L}$ $\left(\mathcal{G}^m_{i,R}\right)$ is the flanking object to the left (right) of λ_m in individual i. Therefore, the proposal density is

$$\begin{aligned}q_{m-1,m}\left(z',z_m\right)&=f\left(\mu_{m1}|0,\tau^2\right)f\left(\mu_{m2}|0,\tau^2\right)\frac{1}{D_K-D_1}\\&\times\Pr\left(\mathcal{Q}^m|\mathcal{Q}',\lambda_m,\lambda',D,\mathcal{M},m\right).\end{aligned}\tag{16.57}$$

In the expressions above,

$$\mathcal{Q}'=\left(\mathcal{Q}^1,\ldots,\mathcal{Q}^{m-1}\right),$$

$$\boldsymbol{\lambda}'=\left(\lambda_1,\ldots,\lambda_{m-1}\right),$$

and

$$\boldsymbol{\mu}'=\left(\mu_{11},\mu_{12},\ldots,\mu_{(m-1)1},\mu_{(m-1)2}\right).$$

Due to the stochastic nature of the proposal, (16.56) can be written as

$$\begin{aligned}\mathsf{Q}^{a}_{m-1,m}\left(z',A_m\right)&=p_{m-1,m}\int_{\Delta}\int_{\mathbb{R}^2}\sum_{\mathcal{Q}^m\in\{0,1\}^{n_{obs}}}I\left((z',z_m)\in A_m\right)\\&\times a_{m-1,m}\left(z',(z',z_m)\right)q_{m-1,m}\left(z',z_m\right)d\lambda_m\,d\mu_{m1}\,d\mu_{m2}.\end{aligned}\tag{16.58}$$

Substituting in (16.55) leads to the following form for the right hand side of the reversibility condition (11.67):

$$\begin{aligned} & p_{m-1,m} \int_{\Delta^m} \int_{\mathbb{R}^{2m}} \sum_{\mathcal{Q} \in \{0,1\}^{mn_{obs}}} I\left((z', z_m) \in A_m, z' \in B_{m-1}\right) \\ & \times f\left(m-1, z'|\mathbf{y}\right) a_{m-1,m}\left(z', (z', z_m)\right) \\ & \times q_{m-1,m}\left(z', z_m\right) d\lambda_m \, d\mu_{m1} \, d\mu_{m2} \, d\boldsymbol{\lambda}' \, d\boldsymbol{\mu}'. \end{aligned} \tag{16.59}$$

In the expressions above, $d\boldsymbol{\lambda}'$ and $d\boldsymbol{\mu}'$ are shorthand for

$$d\lambda_1, d\lambda_2, \ldots, d\lambda_{m-1}$$

and for

$$d\mu_{11}, d\mu_{12}, \ldots, d\mu_{(m-1)1}, d\mu_{(m-1)2},$$

respectively.

Derivation of the Acceptance Probability

The dimensions of (16.54) and (16.59) are equal. In terms of the dimension matching expression (11.60), here, $n_m = m$, $n_{m,m-1} = 0$, $n_{m-1} = m - (2 + 1 + n_{obs})$, and $n_{m-1,m} = 2 + 1 + n_{obs}$, so that

$$n_m + n_{m,m-1} = m = n_{m-1} + n_{m-1,m}.$$

For reversibility to hold, (16.54) must equal (16.59). The equality is satisfied if

$$\begin{aligned} & p_{m,m-1} f\left(m, z|\mathbf{y}\right) a_{m,m-1}\left(z, (z_1, \ldots, z_{m-1})\right) \\ = \; & p_{m-1,m} f\left(m-1, z'|\mathbf{y}\right) a_{m-1,m}\left(z', (z', z_m)\right) q_{m-1,m}\left(z', z_m\right), \end{aligned}$$

which leads to the following expression for the acceptance probability:

$$\begin{aligned} & a_{m,m-1}\left(z, z'\right) \\ = & \min\left[1, \frac{p_{m-1,m} f\left(m-1, z'|\mathbf{y}\right) q_{m-1,m}\left(z', z_m\right)}{p_{m,m-1} f\left(m, z|\mathbf{y}\right)}\right]. \end{aligned} \tag{16.60}$$

The acceptance probability (16.60) holds when the dimension changing moves are based on the strategy of deleting deterministically the last (in terms of the position in the vector z) QTL and appending the new QTL in the last position in z. The diligent reader may wish to confirm that the same acceptance probability is arrived at when the QTL to be deleted is randomly chosen with probability $1/m$ among the m existing positions in z, and the QTL to be added is inserted randomly with probability $1/m$ among the m available positions in z.

Model Selection

The above reversible jump algorithm generates samples from the posterior distribution $p(\boldsymbol{\theta}_i|M=i,\mathbf{y})\Pr(M=i|\mathbf{y})$, where $\boldsymbol{\theta}_i$ are parameters of model i. Models can be compared by means of their posterior probabilities. These can be estimated from

$$\widehat{\Pr}(M=i|\mathbf{y}) = \frac{1}{N}\sum_{j=1}^{N} I(m_j = i)$$

where m_j is the jth Monte Carlo sample of a variable that takes a particular value for each model and N is the number of samples.

This page intentionally left blank

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions*. Dover Publications.

Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statistics in Medicine 8*, 1209–1224.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley.

Aitken, A. C. (1934). A note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society 4*, 106–110.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado, Budapest.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Albert, J. H. and S. Chib (1995). Bayesian residual analysis for binary response regression models. *Biometrika 82*, 747–759.

Anderson, D. A. and M. Aitkin (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society Series B 47*, 203–210.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Applebaum, D. (1996). *Probability and Information - An Integrated Approach*. Cambridge University Press.

Baldi, P. and S. Brunak (1998). *Bioinformatics: The Machine Learning Approach*. MIT Press.

Barndorff-Nielsen, O. E. (1983). On a formula for a distribution of the maximum likelihood estimator. *Biometrika 70*, 343–365.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized log likelihood ratio. *Biometrika 73*, 307–322.

Barndorff-Nielsen, O. E. (1991). Likelihood theory. In D. V. Hinkley, N. Reid, and E. J. Snell (Eds.), *Statistical Theory and Modelling*, Chapter 10, pp. 232–264. Chapman and Hall.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and Asymptotics*. Chapman and Hall.

Barnett, V. (1999). *Comparative Statistical Inference*. Wiley.

Barnett, V. and T. Lewis (1995). *Outliers in Statistical Data*. Wiley.

Bates, D. and D. G. Watts (1988). *Nonlinear Regression Analysis and its Applications*. Wiley.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trabajos de Estadística 32*, 18–31.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Becker, W. A. (1984). *Manual of Quantitative Genetics*. Academic Enterprises.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer–Verlag.

Berger, J. O. and J. M. Bernardo (1992). On the development of reference priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 35–60. Oxford University Press.

Berger, J. O. and L. R. Pericchi (1996). The intrinsic Bayes Factor for model selection and prediction. *Journal of the American Statistical Association 91*, 109–122.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B 41*, 113–147.

Bernardo, J. M. (2001). Bayesian statistics. Submitted manuscript.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Wiley.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B 36*, 192–326.

Besag, J. (1994). Contribution to the discussion paper by Grenander and Miller. *Journal of the Royal Statistical Society Series B 56*, 591–592.

Bibby, J. and H. Toutenburg (1977). *Prediction and Improved Estimation in Linear Models*. Wiley.

Blasco, A. (2001). The Bayesian controversy in animal breeding. *Journal of Animal Science 79*, 2023–2046.

Blasco, A. and L. Varona (1999). Ajuste y comparación de curvas de crecimiento. *ITEA 95A*, 131–142.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology 22*, 134–167.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association 71*, 791–799.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society Series A 143*, 383–430.

Box, G. E. P. and M. E. Muller (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics 29*, 610–611.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Wiley.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computer Graphics and Statistics 7*, 434–455.

Brooks, S. P., P. Giudici, and G. O. Roberts (2001). Efficient construction of reversible jump MCMC proposal distributions. Submitted manuscript.

Brooks, S. P. and G. O. Roberts (1998). Diagnosing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing 8*, 319–335.

Brown, L. D., T. T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Statistical Science 16*, 101–133.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist 105*, 201–211.

Bulmer, M. G. (1979). *Principles of Statistics*. Dover Publications.

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press.

Bunke, O. (1975). Minimax linear, ridge and shrunken estimators for linear parameters. *Mathematische Operationsforschung und Statistik 6*, 697–701.

Carlin, B. P. and T. A. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall.

Casella, G. and R. L. Berger (1990). *Statistical Inference*. Brooks–Cole.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*, 167–170.

Casella, G., M. Lavine, and C. P. Robert (2001). Explaining the perfect sampler. *The American Statistician 55*, 299–305.

Casella, G. and C. P. Robert (1996). Rao-Blackwellisation of sampling schemes. *Biometrika 83*, 81–94.

Chen, M. H., Q. M. Shao, and J. G. Ibrahim (2000). *Monte Carlo Methods in Bayesian Computation*. Springer–Verlag.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–335.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*, 270–281.

Cohen, M. D. (1986). Pseudo-random number generators. In S. Kotz, N. L. Johnson, and C. B. Read (Eds.), *Encyclopedia of Statistics, Vol. 7*, pp. 327–333. Wiley.

Collet, D. (1994). *Modelling Survival Data in Medical Research*. Chapman and Hall.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing 6*, 101–111.

Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association 91*, 883–904.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium 1*, 105–123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series B 24*, 406–424.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall.

Cox, D. R. and H. D. Miller (1965). *The Theory of Stochastic Processes*. Chapman and Hall.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society Series B 49*, 1–39.

Cox, D. R. and E. J. Snell (1989). *Analysis of Binary Data*. Chapman and Hall.

Crow, J. F. and M. Kimura (1970). *An Introduction to Population Genetics Theory*. Harper and Row.

Curnow, R. N. (1961). The estimation of repeatability and heritability from records subject to culling. *Biometrics 17*, 553–566.

Curnow, R. N. (1972). The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk. *Biometrics 28*, 931–946.

Curnow, R. N. and C. Smith (1975). Multifactorial models for familial diseases in man. *Journal of the Royal Statistical Society Series A 138*, 131–169.

Dahlquist, Å, B. and Å. Björck (1974). *Numerical Methods*. Prentice -Hall.

De Finetti, B. (1975a). *Theory of Probability, Vol. 1*. Wiley.

De Finetti, B. (1975b). *Theory of Probability, Vol. 2*. Wiley.

Dempster, A. P. (1974). The direct use of likelihood for significance testing. In O. E. Barndorff-Nielsen, P. Blæsild, and G. Schou (Eds.), *Proceedings of the Conference on the Foundational Questions in Statistical Inference*, pp. 335–352. Department of Theoretical Statistics, University of Aarhus.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing 7*, 247–252.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via de EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B 39*, 1–38.

Dempster, E. R. and I. M. Lerner (1950). Heritability of threshold characters. *Genetics 35*, 212–236.

Dennis, J. E. and R. B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice–Hall.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer–Verlag.

Dickey, J. M. (1968). Three multidimensional integral identities with Bayesian applications. *Annals of Statistics 39*, 1615–1627.

Doerge, R. W. and G. A. Churchill (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics 142*, 285–294.

Draper, N. R. and H. Smith (1981). *Applied Regression Analysis*. Wiley.

Ducrocq, V., R. L. Quaas, E. Pollak, and G. Casella (1988). Length of productive life of dairy cows. 2. Variance component estimation and sire evaluation. *Journal of Dairy Science 71*, 3071–3079.

Durbin, R., S. R. Eddy, A. Krogh, and G. J. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Earman, J. (1992). *Bayes or Bust*. The MIT Press.

Edwards, A. W. F. (1974). The history of likelihood. *International Statistical Review 42*, 9–15.

Edwards, A. W. F. (1992). *Likelihood*. The John Hopkins University Press.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika 80*, 3–26.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science 13*, 95–122.

Efron, B. and D. V. Hinkley (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika 65*, 457–482.

Elston, R. C. and J. Stewart (1971). A general model for the genetic analysis of pedigree data. *Human Heredity 21*, 523–542.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer–Verlag.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics 29*, 51–76.

Falconer, D. S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics 31*, 1–20.

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics*. Longman.

Famula, T. R. (1981). Exponential stayability model with censoring and covariates. *Journal of Dairy Science 64*, 538–545.

Fan, J., H. Hung, and W. Wong (2000). Geometric understanding of likelihood ratio statistics. *Journal of the American Statistical Association 95*, 836–841.

Feller, W. (1970). *An Introduction to Probability Theory and its Applications, Vol. 1*. Wiley.

Feng, Z. D. and C. E. McCulloch (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society Series B 58*, 609–617.

Fernandez, C. and M. F. J. Steel (1998). On Bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association 93*, 359–371.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh 52*, 399–433.

Fisher, R. A. (1920). A mathematical examination of determining accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society 80*, 758–770.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A 222*, 309–368.

Fisher, R. A. (1925). Theory of statistical information. *Proceedings of the Cambridge Philosophical Society 22*, 700–725.

Fishman, G. S. (1973). *Concepts and Methods in Discrete Event Digital Simulation*. Wiley.

Flury, B. and A. Zoppe (2000). Exercises in EM. *The American Statistician 54*, 207–209.

Foulley, J. L., D. Gianola, and R. Thompson (1983). Prediction of genetic merit from data on categorical and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Genetics, Selection, Evolution 25*, 407–424.

Foulley, J. L., S. Im, D. Gianola, and I. Hoeschele (1987). Empirical Bayes estimation of parameters for n polygenic binary traits. *Genetics, Selection, Evolution 19*, 197–224.

Foulley, J. L. and E. Manfredi (1991). Approches statistiques de l'évaluation génétiques des reproducteurs pour des caractères binaires à seuils. *Genetics, Selection, Evolution 23*, 309–338.

Fox, C. (1987). *An Introduction to Calculus of Variations*. Dover.

Galton, F. (1885). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute 15*, 246–263.

García-Cortés, L. A. and D. Sorensen (1996). On a multivariate implementation of the Gibbs sampler. *Genetics, Selection, Evolution 28*, 121–126.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*, 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 145–161. Chapman and Hall.

Gelfand, A. E. and D. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B 56*, 501–514.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford University Press.

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association 85*, 972–985.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parameterization for normal linear mixed models. *Biometrika 82*, 479–488.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1996). Efficient parameterizations for generalized linear mixed models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 165–180. Oxford University Press.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

Gelman, A., X. L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*, 457–511.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

George, A. W., K. L. Mengersen, and G. P. Davis (2000). Localization of a quantitative trait locus via a Bayesian approach. *Biometrics 56*, 40–51.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics 8*, S19–S40.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*, 473–511.

Gianola, D. (1982). Theory and analysis of threshold characters. *Journal of Animal Science 54*, 1079–1096.

Gianola, D. and R. L. Fernando (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science 63*, 217–244.

Gianola, D., R. L. Fernando, S. Im, and J. L. Foulley (1989). Likelihood estimation of quantitative genetic parameters when selection occurs: Models and problems. *Genome 31*, 768–777.

Gianola, D. and J. L. Foulley (1983). Sire evaluation for ordered categorical data with a threshold model. *Genetics, Selection, Evolution 15*, 201–223.

Gianola, D., S. Im, and F. W. Macedo (1990). A framework for prediction of breeding values. In D. Gianola and K. Hammond (Eds.), *Statistical Methods for Genetic Improvement of Livestock*, pp. 210–238. Springer–Verlag.

Gilks, W. R. and G. O. Roberts (1996). Strategies for improving MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 89–114. Chapman and Hall.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41*, 336–348.

Gilmour, A. R., R. D. Anderson, and A. L. Rae (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika 72*, 593–599.

Go, R. C. P., R. C. Elston, and E. B. Kaplan (1978). Efficiency and robustness of pedigree segregation analysis. *American Journal of Human Genetics 30*, 28–37.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B 14*, 107–114.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association 53*, 799–813.

Goodman, L. A. and H. O. Hartley (1958). The precision of unbiased ratio-type estimators. *Journal of the American Statistical Association 53*, 491–508.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika 82*, 711–732.

Grimmet, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes*. Clarendon Press.

Gross, A. J. and V. A. Clark (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley.

Grossman, S. I. and J. E. Turner (1974). *Mathematics for the Biological Sciences*. Macmillan.

Guo, S. W. and E. A. Thompson (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics 50*, 417–432.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.

Hager, W. H. (1988). *Applied Numerical Linear Algebra*. Prentice–Hall.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics 8*, 229–309.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika 35*, 297–303.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo Methods*. Wiley.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics*. Wiley.

Han, C. and B. P. Carlin (2001). Markov chain Monte Carlo methods for computing Bayes Factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

Harville, D. A. (1974). Bayesian inference of variance components using only error contrasts. *Biometrika 61*, 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–340.

Harville, D. A. and R. W. Mee (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics 40*, 393–408.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika 57*, 97–109.

Hazel, L. N. (1943). The genetic basis for constructing selection indices. *Genetics 28*, 476–490.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics 61*, 748–760.

Heisenberg, W. (1958). The representation of nature in contemporary physics. *Daedalus 87*, 95–108.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics 9*, 226–252.

Henderson, C. R. (1963). Selection index and expected selection advance. In W. D. Hanson and H. F. Robinson (Eds.), *Statistical Genetics and Plant Breeding*, pp. 141–163. National Academy of Sciences, National Research Council Publication No. 982, Washington, D C.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. J. L. Lush*, pp. 10–41. American Society of Animal Science.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics 31*, 423–447.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. N. Von Krosigk (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, M. and M. C. Meyer (2001). Exploring the confidence interval for a binomial parameter in a first course in statistical computing. *The American Statistician 55*, 337–344.

Heringstad, B., R. Rekaya, D. Gianola, G. Klemetsdal, and K. A. Weigel (2001). Bayesian analysis of liability of clinical mastitis in Norwegian cattle with a threshold model: Effects of data sampling and model specification. *Journal of Dairy Science 84*, 2337–2346.

Hills, S. E. and A. F. M. Smith (1992). Parameterization issues in Bayesian inference (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 227–246. Oxford University Press.

Hills, S. E. and A. F. M. Smith (1993). Diagnostic plots for improved parameterization in Bayesian inference. *Biometrika 80*, 61–74.

Hobert, J. P. and G. Casella (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *Journal of the American Statistical Association 91*, 1461–1473.

Hoel, P. G., S. C. Port, and C. J. Stone (1971). *Introduction to Probability Theory*. Houghton Mifflin.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression. *Technometrics 12*, 55–67; 69–82.

Hoeschele, I. and B. Tier (1995). Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling. *Genetics, Selection, Evolution 27*, 519–540.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*, 382–417.

Hogg, R. V. and A. T. Craig (1995). *Introduction to Mathematical Statistics*. Prentice Hall.

Howson, C. and P. Urbach (1989). *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle.

Jamrozik, J. and L. R. Schaeffer (1997). Estimates of genetic parameters for a test-day model with random regressions for production of first lactation Holsteins. *Journal of Dairy Science 80*, 762–770.

Janss, L. L. G., R. Thompson, and J. A. M. Van Arendonk (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics 91*, 1137–1147.

Jaynes, E. T. (1957). On the rationale of maximum-entropy methods. *Physics Review 106*, 620–630.

Jaynes, E. T. (1994). *Probability Theory: The Logic of Science*. http://omega.albany.edu:8008/JaynesBook.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press.

Jensen, C. S., U. Kjærulff, and A. Kong (1995). Blocking Gibbs sampling in very large probabilistic expert systems. *International Journal of Human Computer Studies 42*, 647–666.

Jensen, J. (1994). Bayesian analysis of bivariate mixed models with one continuous and one binary trait using the Gibbs sampler. In *Proceedings of the 5th World Congress of Genetics Applied to Livestock Production, Vol. 18*, pp. 333–336. University of Guelph.

Jensen, J., C. S. Wang, D. Sorensen, and D. Gianola (1994). Bayesian inference on variance and covariance components for traits influenced by maternal and direct genetic effects using the Gibbs sampler. *Acta Agricultura Scandinavica 44*, 193–201.

Johnson, N. L. and S. Kotz (1969). *Distributions in Statistics: Discrete Distributions*. Wiley.

Johnson, N. L. and S. Kotz (1970a). *Distributions in Statistics: Continuous Univariate Distributions, Vol. 1*. Wiley.

Johnson, N. L. and S. Kotz (1970b). *Distributions in Statistics: Continuous Univariate Distributions, Vol. 2*. Wiley.

Johnson, N. L. and S. Kotz (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley.

Kackar, R. N. and D. A. Harville (1981). Ubiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics Series A: Theory and Methods 10*, 1249–1261.

Kadarmideen, H. N., R. Rekaya, and D. Gianola (2002). Genetic parameters for clinical mastitis in Holstein Freisians in the United Kingdom: A Bayesian analysis. *Animal Science*. In press.

Kalbfleisch, J. D. and D. A. Sprott (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *Journal of the Royal Statistical Society Series B 32*, 175–208.

Kalbfleisch, J. D. and D. A. Sprott (1973). Marginal and conditional likelihoods. *Sankya A 35*, 311–328.

Kaplan, W. (1993). *Advanced Calculus*. Addison and Wesley.

Karlin, S. and H. M. Taylor (1975). *A First Course in Stochastic Processes*. Academic Press.

Kass, E. R. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kass, R. E. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician 52*, 93–100.

Keller, E. F. (2000). *The Century of the Gene*. Harvard University Press.

King, G. (1989). *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*. Cambridge University Press.

Kirkpatrick, M., W. G. Hill, and R. Thompson (1994). Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genetical Research 64*, 57–69.

Kleinbaum, D. G. (1996). *Survival Analysis*. Springer–Verlag.

Knott, S. A. and C. S. Haley (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research 60*, 139–151.

Koerkhuis, A. N. M. and R. Thompson (1997). Models to estimate maternal effects for juvenile body weights in broiler chickens. *Genetics, Selection, Evolution 29*, 225–249.

Korsgaard, I. R., A. H. Andersen, and D. Sorensen (1999). A useful reparameterisation to obtain samples from conditional inverse Wishart distributions. *Genetics, Selection, Evolution 31*, 177–181.

Korsgaard, I. R., M. S. Lund, D. Sorensen, D. Gianola, P. Madsen, and J. Jensen (2002). Multivariate Bayesian analysis of Gaussian, right censored Gaussian, ordered categorical and binary traits using Gibbs sampling. *Genetics, Selection, Evolution*. In press.

Kullback, S. (1968). *Information Theory and Statistics*. Wiley.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*, 963–974.

Lange, K. (1995). A Quasi-Newton acceleration of the EM algorithm. *Journal of the Royal Statistical Society Series B 44*, 226–233.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer–Verlag.

Lange, K. and R. C. Elston (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity 25*, 95–105.

Lange, K. and T. M. Goradia (1987). An algorithm for automatic genotype elimination. *American Journal of Human Genetics 40*, 250–256.

Lange, K. and J. S. Sinsheimer (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computer Graphics and Statistics 2*, 175–198.

Lee, J. K. and D. C. Thomas (2000). Performance of Markov chain Monte Carlo approaches for mapping genes in oligogenic models with unknown number of loci. *American Journal of Human Genetics 67*, 1232–1250.

Lee, P. M. (1989). *Bayesian Statistics: An Introduction*. Edward Arnold.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society Series B 58*, 619–678.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer–Verlag.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. Springer–Verlag.

Leonard, T. and J. S. Hsu (1999). *Bayesian Methods*. Cambridge University Press.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Annals of Mathematical Statistics 27*, 986–1005.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*, 187–192.

Lindley, D. V. and A. F. M. Smith (1972). Bayesian estimates for the linear model. *Journal of the Royal Statistical Society Series B 34*, 1–41.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

Liu, C. and D. B. Rubin (1995). ML estimation of the t distribution using EM and its extensions, ECM, and ECME. *Statistica Sinica 5*, 19–39.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association 89*, 958–966.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer–Verlag.

Liu, J. S., H. W. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika 81*, 27–40.

Lo, Y., N. R. Mendell, and D. B. Rubin (2001). Testing the number of components in a normal mixture. *Biometrika 88*, 767–778.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B 44*, 226–233.

Lund, M. S. and C. S. Jensen (1999). Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genetics, Selection, Evolution 31*, 3–24.

Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates.

MacCluer, J. W., J. L. Vandeburg, B. Read, and O. A. Ryder (1986). Pedigree analysis by computer simulation. *Zoo Biology 5*, 147–160.

Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*, 1535–1546.

Malécot, G. (1947). Annotated translation by D. Gianola of: Les criteres statistiques et la subjectivite de la connaisance scientifique (Statistical methods and the subjective basis of scientific knowledge), by G. Malécot, (1947), Annales de l'Universite de Lyon, X, 43-74. *Genetics, Selection, Evolution 31*, 269–298.

Malécot, G. (1969). *The Mathematics of Heredity*. W. H. Freeman. Originally published in 1948 by Masson et Cie.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press.

Marsaglia, G. and A. Zaman (1993). *The Kiss Generator*. Technical Report, Department of Statistics, University of Florida.

Martinez, V., L. Bünger, and W. G. Hill (2000). Analysis of response to 20 generations of selection for body composition in mice: Fit to infinitesimal model. *Genetics, Selection, Evolution 32*, 3–21.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*, 330–335.

McCulloch, R. E. and P. E. Rossi (1991). A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics 49*, 141–168.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.

Meeker, W. Q. and L. A. Escobar (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician 49*, 48–53.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society Series B 51*, 127–138.

Meng, X. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association 86*, 899–909.

Mengersen, K. L., C. P. Robert, and C. Guihenneuc-Jouyaux (1999). MCMC convergence diagnostics: A reviewww (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 6*, pp. 415–440. Oxford University Press.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Meyer, K. (1999). Estimates of genetic and phenotypic covariance functions for post-weaning growth and mature weight of beef cows. *Journal of Animal Breeding and Genetics 116*, 181–205.

Meyn, S. P. and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer–Verlag.

Milliken, G. A. and D. E. Johnson (1992). *Analysis of Messy Data, Vol. I: Designed Experiments*. Chapman and Hall.

Misztal, I., D. Gianola, and J. L. Foulley (1989). Computing aspects of nonlinear methods of sire evaluation for categorical data. *Journal of Dairy Science 72*, 1557–1568.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Moreno, C., D. Sorensen, L. A. García-Cortés, L. Varona, and J. Altarriba (1997). On biased inferences about variance components in the binary threshold model. *Genetics, Selection, Evolution 29*, 145–160.

Morton, N. E. and C. J. MacLean (1974). Analysis of family resemblance. III. Complex segregation of quantitative traits. *American Journal of Human Genetics 26*, 489–503.

Nandram, B. and M. H. Chen (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation 54*, 129–144.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A 135*, 370–384.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society Series B 56*, 1–48.

Neyman, J. and E. S. Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Parts I and II. *Biometrika 20A*, 175–240; 263–294.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society Series B 61*, 479–482.

Odell, P. L. and A. H. Feiveson (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association 61*, 198–203.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*. Edward Arnold.

Ott, J. (1999). *Analysis of Human Genetic Linkage*. John Hopkins University Press.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Pauler, D. K., J. C. Wakefield, and R. E. Kass (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association 94*, 1242–1253.

Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *The American Statistician 54*, 54–56.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VIII. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society of London Series A 195*, 79–121.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London Series A 200*, 1–66.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika 60*, 607–612.

Popper, K. R. (1972). *The Logic of Scientific Discovery*. Hutchinson.

Popper, K. R. (1982). *Quantum Theory and the Schism in Physics*. Routledge.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms 9*, 223–252.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society 44*, 50–57.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley.

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science 10*, 138–199.

Reid, N. (2000). Likelihood. *Journal of the American Statistical Association 95*, 1335–1340.

Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B 59*, 731–792.

Ripley, B. (1987). *Stochastic Simulation*. Wiley.

Robert, C. P. (1994). *The Bayesian Choice*. Springer–Verlag.

Robert, C. P. (1998). *Discretization and MCMC Convergence Assessment*. Lecture Notes in Statistics, Vol. 135. Springer–Verlag.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer–Verlag.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability 7*, 110–120.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society Series B 59*, 291–317.

Robertson, A. (1977). The effect of selection on the estimation of genetic parameters. *Journal of Animal Breeding and Genetics 94*, 131–135.

Robertson, A. and I. M. Lerner (1949). The heritability of all-or-none traits: Viability of poultry. *Genetics 34*, 395–411.

Rodriguez-Zas, S. L. (1998). *Bayesian Analysis of Somatic Cell Score Lactation Patterns in Holstein Cows Using Nonlinear Mixed Effects Models*. Ph. D. thesis, University of Wisconsin-Madison.

Roff, D. E. (1997). *Evolutionary Quantitative Genetics*. Chapman and Hall.

Rogers, W. H. and J. W. Tukey (1972). Understanding some long-tailed distributions. *Statistica Neerlandica 26*, 211–226.

Rosa, G. J. M. (1998). *Analise Bayesiana de Models Lineares Mistos Robustos Via Amostrador de Gibbs*. Ph. D. thesis, Escola Superior de Agricultura Luis de Queiroz, Piracicaba, Sao Paulo, Brazil.

Rosa, G. J. M., D. Gianola, and J. I. Urioste (2001). Assessing relationships between genetic evaluations using robust regression with an application to Holsteins in Uruguay. *Acta Agricultura Scandinavica Series A 51*, 21–34.

Ross, S. M. (1997). *Simulation*. Academic Press.

Royall, R. (1997). *Statistical Evidence*. Chapman and Hall.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1987b). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. Discussion of Tanner and Wong. *Journal of the American Statistical Association 82*, 543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford University Press.

Savage, L. J. (1972). *The Foundations of Statistics*. Wiley.

Schafer, J. L. (2000). *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley.

Searle, S. R. (1971). *Linear Models*. Wiley.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance Components*. Wiley.

Severini, T. A. (1998). Lilelihood functions for inference in the presence of a nuisance parameter. *Biometrika 85*, 507–522.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 623–656.

Sheehan, N. A. (2000). On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review 68*, 83–110.

Sheehan, N. A., B. Guldbrandtsen, M. S. Lund, and D. Sorensen (2002). Bayesian McMC mapping of quantitative trait loci in a half-sib design: A graphical model perspective. *International Statistical Review*. In press.

Sheehan, N. A. and A. Thomas (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics 49*, 163–175.

Sillanpää, M. J. and E. Arjas (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics 148*, 1373–1388.

Sillanpää, M. J. and E. Arjas (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics 151*, 1605–1619.

Silverman, B. (1992). *Density Estimation*. Chapman and Hall.

Sivia, D. S. (1996). *Data Analysis. A Bayesian Tutorial*. Oxford University Press.

Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician 46*, 84–88.

Smith, C. A. B. (1959). Some comments on the statistical methods used in linkage investigations. *American Journal of Human Genetics 11*, 289–304.

Smith, H. F. (1936). A discriminant function for plant selection. *Annals of Eugenics 7*, 240–250.

Sorensen, D. (1996). *Gibbs Sampling in Quantitative Genetics*. Danish Institute of Agricultural Sciences; Internal Report 82, 192 pp.

Sorensen, D., S. Andersen, D. Gianola, and I. R. Korsgaard (1995). Bayesian inference in threshold models using Gibbs sampling. *Genetics, Selection, Evolution 27*, 229–249.

Sorensen, D., R. L. Fernando, and D. Gianola (2001). Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research 77*, 83–94.

Sorensen, D., A. Vernersen, and S. Andersen (2000). Bayesian analysis of response to selection: A case study using litter size in Danish Yorkshire pigs. *Genetics 156*, 283–295.

Sorensen, D., C. S. Wang, J. Jensen, and D. Gianola (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics, Selection, Evolution 26*, 333–360.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*. In press.

Stein, S. K. (1977). *Calculus and Analytic Geometry*. McGraw-Hill.

Stephens, D. A. and R. D. Fisch (1998). Bayesian analysis of a quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics 54*, 1334–1347.

Stramer, O. and R. L. Tweedie (1998). Langevin-type models II: Self-targetting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability 1*, 307–328.

Strandén, I. (1996). *Robust Mixed Effects Linear Models with t-Distributions and Applications to Dairy Cattle Breeding*. Ph. D. thesis, University of Wisconsin-Madison.

Strandén, I. and D. Gianola (1998). Atenuating effects of preferential treatment with Student-t mixed linear models: A simulation study. *Genetics, Selection, Evolution 30*, 565–583.

Strandén, I. and D. Gianola (1999). Mixed effects linear models with t-distributions for quantitative genetic analysis: A Bayesian approach. *Genetics, Selection, Evolution 31*, 25–42.

Stuart, A. and J. K. Ord (1987). *Kendall's Advanced Theory of Statistics. Distribution Theory*. Edward Arnold.

Stuart, A. and J. K. Ord (1991). *Kendall's Advanced Theory of Statistics. Classical Inference and Relationship*. Edward Arnold.

Swendsen, R. and J. Wang (1987). Non-universal critical dynamics in Monte Carlo simulations. *Physical Review Letters 58*, 86–88.

Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer–Verlag.

Tanner, M. A. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–550.

Thompson, E. A. (2001). Monte Carlo methods on genetic structures. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg (Eds.), *Complex Stochastic Systems*, pp. 175–218. Chapman and Hall.

Thompson, E. A. and S. C. Heath (2000). Estimation of conditional multilocus gene identity among relatives. In F. Seiller-Moiseiwitsch (Ed.), *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology*, pp. 95–113. Institute of Mathematical Statistics, Hayward, CA.: IMS Lecture Note-Monograph Series, Volume 33.

Thompson, R. (1973). The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics 29*, 527–550.

Thompson, R. (1976). Estimation of quantitative genetic parameters. In E. Pollak, O. Kempthorne, and T. B. Bailey (Eds.), *Proceedings of the International Conference on Quantitative Genetics*, pp. 639–657. Iowa State University.

Thompson, R. (1980). Maximum likelihood estimation of variance components. *Mathematische Operationsforschung und Statistik 11*, 545–561.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics 22*, 1701–1786.

Tierney, L. and J. B. Kadane (1989). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Toutenburg, H. (1982). *Prior Information in Linear Models*. Wiley.

Uimari, P. and I. Hoeschele (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics 146*, 735–743.

Van Tassell, C. P. and L. D. Van Vleck (1996). Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co)variance component inferences. *Journal of Animal Science 74*, 2586–2597.

Van Tassell, C. P., L. D. Van Vleck, and K. E. Gregory (1998). Bayesian analysis of twinning and ovulation rates using a multiple-trait threshold model and Gibbs sampling. *Journal of Animal Science 76*, 2048–2061.

Van Vleck, L. D. (1993). *Selection Index and Introduction to Mixed Model Methods*. CRC Press.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 307–333.

Waagepetersen, R. and D. Sorensen (2001). A tutorial on reversible jump MCMC with a view towards applications in QTL-mapping. *International Statistical Review 69*, 49–61.

Wang, C. S., D. Gianola, D. Sorensen, J. Jensen, A. Christensen, and J. J. Rutledge (1994). Response to selection in Danish Landrace pigs: A Bayesian analysis. *Theoretical and Applied Genetics 88*, 220–230.

Wang, C. S., R. L. Quaas, and E. J. Pollak (1997). Bayesian analysis of calving ease scores and birth weights. *Genetics, Selection, Evolution 29*, 117–143.

Wei, G. C. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association 85*, 699–704.

Weinstock, R. (1974). *Calculus of Variations with Applications to Physics and Engineering*. Dover.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates.

Wiggans, G. R. and M. E. Goddard (1997). A computationally feasible test-day model for genetic evaluation of yield traits in the United States. *Journal of Dairy Science 80*, 1795–1800.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*, 60–62.

Willham, R. L. (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics 19*, 18–27.

Williamson, R. E., R. H. Crowell, and H. F. Trotter (1972). *Calculus of Vector Functions*. Prentice–Hall.

Wilton, J. W., D. A. Evans, and L. D. Van Vleck (1968). Selection indices for quadratic models of total merit. *Biometrics 24*, 937–949.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika 80*, 791–795.

Wolfinger, R. and X. Lin (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics and Data Analysis 25*, 465–490.

Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics 19*, 506–536.

Wright, S. (1968). *Evolution and the Genetics of Populations. Genetic and Biometric Foundations*. University of Chicago.

Yi, N. and S. Xu (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics 156*, 411–422.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *Journal of the American Statistical Association 71*, 400–405.

Zellner, A. and R. Highfield (1988). Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *Journal of Econometrics 37*, 195–210.

List of Citations

Abramowitz and Stegun (1972), 86, 368
Agresti (1989), 626
Agresti (1990), 626
Agresti (1996), 626
Aitken (1934), 62
Akaike (1973), 421
Albert and Chib (1993), 435, 598, 606, 609
Albert and Chib (1995), 292, 435
Anderson and Aitkin (1985), 606
Anderson (1984), 43, 51, 56
Applebaum (1996), 334, 335, 338, 340, 342, 370, 371
Baldi and Brunak (1998), 338
Barndorff-Nielsen and Cox (1994), 166, 181
Barndorff-Nielsen (1983), 133
Barndorff-Nielsen (1986), 181
Barndorff-Nielsen (1991), 181
Barnett and Lewis (1995), 434, 591
Barnett (1999), 219
Bates and Watts (1988), 292
Bayarri (1981), 81, 104
Bayes (1763), 81
Becker (1984), 89
Berger and Bernardo (1992), 337, 356, 397
Berger and Pericchi (1996), 422, 423
Berger (1985), 406
Bernardo and Smith (1994), 4, 104, 120, 214, 219, 289, 330, 331, 333, 342, 356, 364, 379, 382, 393, 395, 397, 405, 406, 411, 412, 564, 607
Bernardo (1979), 337, 356, 379, 385, 397, 400
Bernardo (2001), 394, 395
Besag (1974), 511
Besag (1994), 517
Bibby and Toutenburg (1977), 301
Blasco and Varona (1999), 629, 633
Blasco (2001), 119, 121
Bliss (1935), 606

Box and Muller (1958), 105
Box and Tiao (1973), 17, 85, 104, 214, 227, 237, 244, 262, 278, 289, 337, 356, 360, 362, 366, 514, 545, 650, 687, 693
Box (1976), 671
Box (1980), 218
Brooks and Gelman (1998), 550
Brooks and Roberts (1998), 547
Brooks et al. (2001), 532
Brown et al. (2001), 13
Bulmer (1971), 44
Bulmer (1979), 14, 44, 46
Bulmer (1980), 26, 62, 65
Bunke (1975), 301
Carlin and Louis (1996), 543
Casella and Berger (1990), 13, 33, 138, 145, 148, 152
Casella and George (1992), 488
Casella and Robert (1996), 552
Casella et al. (2001), 556
Chen et al. (2000), 551, 554, 585
Chib and Greenberg (1995), 503, 504, 689
Chib and Jeliazkov (2001), 428
Chib (1995), 427, 428
Cohen (1986), 20
Collet (1994), 439
Congdon (2001), 421, 427, 437
Cowles and Carlin (1996), 547, 550
Cowles (1996), 612
Cox and Hinkley (1974), 157, 166
Cox and Miller (1965), 477, 479, 494
Cox and Reid (1987), 181
Cox and Snell (1989), 178, 188, 204, 607
Cox (1961), 174, 529
Cox (1962), 174, 529
Crow and Kimura (1970), 166, 175
Curnow and Smith (1975), 606
Curnow (1961), 62
Curnow (1972), 606
Dahlquist and Björck (1974), 162–164
Dempster and Lerner (1950), 89, 90, 92, 202, 605, 606
Dempster et al. (1977), 444, 584
Dempster (1974), 429
Dempster (1997), 429
Dennis and Schnabel (1983), 162
Devroye (1986), 39, 66
De Finetti (1975a), 4
De Finetti (1975b), 29
Dickey (1968), 278
Doerge and Churchill (1996), 685
Draper and Smith (1981), 292, 434
Ducrocq et al. (1988), 80
Durbin et al. (1998), 338
Earman (1992), 219
Edwards (1974), 119
Edwards (1992), 26, 119, 167, 181
Efron and Hinkley (1978), 131
Efron (1993), 181
Efron (1998), 133
Elston and Stewart (1971), 671
Fahrmeir and Tutz (2001), 626
Falconer and Mackay (1996), 287, 344
Falconer (1965), 202, 605, 606
Falconer (1967), 606
Famula (1981), 24
Fan et al. (2000), 174
Feller (1970), 477
Feng and McCulloch (1996), 147
Fernandez and Steel (1998), 595
Fisher (1918), 44, 102
Fisher (1920), 142
Fisher (1922), 119, 122, 142
Fisher (1925), 346
Fishman (1973), 19, 21
Flury and Zoppe (2000), 447
Foulley and Manfredi (1991), 606
Foulley et al. (1983), 606, 615
Foulley et al. (1987), 606
Fox (1987), 375, 377
Galton (1885), 287

García-Cortés and Sorensen (1996), 585
Geisser and Eddy (1979), 438
Geisser (1993), 438
Gelfand and Dey (1994), 429, 438
Gelfand and Smith (1990), 85, 540, 556
Gelfand et al. (1990), 547, 552
Gelfand et al. (1992), 436, 437
Gelfand et al. (1995), 602, 604
Gelfand et al. (1996), 602–604
Gelfand (1996), 436, 437
Gelman and Rubin (1992), 540, 548–550
Gelman et al. (1995), 10, 40, 57, 61, 214, 220, 577
Gelman et al. (1996), 437, 438
Geman and Geman (1984), 497
George et al. (2000), 694
Geweke (1989), 556
Geweke (1993), 598, 599
Geyer (1992), 540, 554, 555
Gianola and Fernando (1986), 52, 66, 313, 650
Gianola and Foulley (1983), 83, 606
Gianola et al. (1989), 26, 62
Gianola et al. (1990), 313
Gianola (1982), 202, 606
Gilks and Roberts (1996), 547
Gilks and Wild (1992), 658
Gilmour et al. (1985), 606
Go et al. (1978), 672
Goodman and Hartley (1958), 155
Good (1952), 441
Good (1958), 401
Green (1995), 497, 518
Grimmet and Stirzaker (1992), 482
Gross and Clark (1975), 439
Grossman and Turner (1974), 478
Guo and Thompson (1994), 672, 675, 676
Hacking (1965), 167
Hager (1988), 162
Haldane (1919), 680
Haldane (1948), 359, 360
Hammersley and Handscomb (1964), 556
Hampel et al. (1986), 589
Han and Carlin (2001), 428
Harville and Mee (1984), 606
Harville (1974), 472, 651
Harville (1977), 266
Hastings (1970), 497, 555
Hazel (1943), 576
Heath (1997), 679, 694
Heisenberg (1958), 219
Henderson and Meyer (2001), 13
Henderson et al. (1959), 62, 318
Henderson (1953), 26, 313
Henderson (1963), 26, 51, 125
Henderson (1973), 26, 50, 51, 125, 212, 279, 282, 318
Henderson (1975), 26, 62, 64
Henderson (1984), 102
Heringstad et al. (2001), 606
Hills and Smith (1992), 584, 604
Hills and Smith (1993), 604
Hobert and Casella (1996), 545, 566
Hoel et al. (1971), 79
Hoerl and Kennard (1970), 300
Hoeschele and Tier (1995), 609
Hoeting et al. (1999), 439, 441, 442
Hogg and Craig (1995), 3
Howson and Urbach (1989), 219
Jamrozik and Schaeffer (1997), 628
Janss et al. (1995), 672, 679
Jaynes (1957), 341
Jaynes (1994), 354, 367
Jeffreys (1961), 289, 356, 360, 361, 367, 401
Jensen et al. (1994), 571
Jensen et al. (1995), 679
Jensen (1994), 606, 615
Johnson and Kotz (1969), 4
Johnson and Kotz (1970a), 4
Johnson and Kotz (1970b), 4, 83
Johnson and Kotz (1972), 4

Norris (1997), 477
O'Hagan (1994), 214, 217, 219, 253, 264, 298, 354, 356, 364, 411, 414–416, 420–424, 526, 541
Oakes (1999), 453, 457, 458, 465
Odell and Feiveson (1966), 59
Ott (1999), 680, 685
Patterson and Thompson (1971), 26, 128, 183, 186
Pauler et al. (1999), 418
Pawitan (2000), 180
Pearson (1900), 606
Pearson (1903), 26, 62, 64
Peskun (1973), 507, 508, 523
Popper (1972), 219
Popper (1982), 219
Priestley (1981), 555
Propp and Wilson (1996), 556
Raftery et al. (1997), 439, 442
Raj (1968), 155
Rao (1947), 180
Rao (1973), 26, 43, 87, 115, 145, 200
Reid (1995), 181
Reid (2000), 181
Richardson and Green (1997), 518
Ripley (1987), 19, 554, 657
Robert and Casella (1999), 477, 498, 547, 551
Roberts and Sahu (1997), 584, 585, 602
Roberts et al. (1997), 504
Robertson and Lerner (1949), 90, 606
Robertson (1977), 62
Robert (1994), 20
Robert (1998), 547
Rodriguez-Zas (1998), 599, 670
Roff (1997), 56
Rogers and Tukey (1972), 589, 664
Rosa et al. (2001), 670
Rosa (1998), 589
Ross (1997), 19
Royall (1997), 167, 168
Rubin (1976), 576
Rubin (1987a), 577
Rubin (1987b), 556
Rubin (1988), 661
Savage (1972), 218
Schafer (2000), 577
Schwarz (1978), 420, 421
Scott (1992), 553
Searle et al. (1992), 26, 51, 125, 184, 234, 279, 343
Searle (1971), 32, 39, 43, 53, 67, 85, 92, 115, 147, 169, 183, 254, 284, 313, 318, 564
Searle (1982), 51, 53, 183, 603
Severini (1998), 181
Severini (2000), 166, 181
Shannon (1948), 337, 341, 346
Sheehan and Thomas (1993), 678
Sheehan et al. (2002), 679
Sheehan (2000), 675, 678, 679
Sillanpää and Arjas (1998), 691, 694
Sillanpää and Arjas (1999), 694
Silverman (1992), 553
Sivia (1996), 367, 370, 371
Smith and Gelfand (1992), 556
Smith (1936), 576
Smith (1959), 408
Sorensen et al. (1994), 26
Sorensen et al. (1995), 548, 555, 606, 607, 619
Sorensen et al. (2000), 438
Sorensen et al. (2001), 66
Sorensen (1996), 606, 615
Spiegelhalter et al. (2002), 429–431
Stein (1977), 132
Stephens and Fisch (1998), 694
Stramer and Tweedie (1998), 517
Strandén and Gianola (1998), 594
Strandén and Gianola (1999), 589, 599, 664
Strandén (1996), 589, 598, 670
Stuart and Ord (1987), 53

Stuart and Ord (1991), 39, 139, 166, 167
Swendsen and Wang (1987), 532
Tanner and Wong (1987), 241, 293, 473, 532, 556
Tanner (1996), 28, 454, 658
Thompson and Heath (2000), 679
Thompson (1973), 62
Thompson (1976), 62
Thompson (1980), 266
Thompson (2001), 672, 679
Tierney and Kadane (1989), 420
Tierney (1994), 497, 498, 501
Toutenburg (1982), 301
Uimari and Hoeschele (1997), 694
Van Tassell and Van Vleck (1996), 581
Van Tassell et al. (1998), 615
Van Vleck (1993), 571
Vuong (1989), 174
Waagepetersen and Sorensen (2001), 497, 518, 694
Wang et al. (1994), 24
Wang et al. (1997), 581, 606, 615, 616
Wei and Tanner (1990), 447, 462
Weinstock (1974), 375, 377
Weir (1996), 175, 195
Wiggans and Goddard (1997), 628
Wilks (1938), 166
Willham (1963), 208, 570
Williamson et al. (1972), 97
Wilton et al. (1968), 283
Wolfinger and Lin (1997), 653
Wolfinger (1993), 653
Wright (1934), 90, 202, 606
Wright (1968), 23, 69
Yi and Xu (2000), 694
Zellner and Highfield (1988), 370
Zellner (1971), 214, 263, 296, 337, 356, 360, 361, 403, 635
Zellner (1976), 589, 592, 599, 664

Subject Index