Analysis of Incomplete Multivariate Data

J. L. Schafer
Department of Statistics
The Pennsylvania State University
USA

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress.

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

First edition 1997
First CRC Press reprint 1999
Originally published by Chapman & Hall
No claim to original U.S. Government works
International Standard Book Number 0-412-04061-1
Printed in the United States of America 3 4 5 6 7 8 9 0
Printed on acid-free paper

Contents

8 Loglinear models

9 Methods for mixed data

Appendices

Preface

The last quarter of a century has seen enormous developments in general statistical methods for incomplete data. The EM algorithm and its extensions, multiple imputation and Markov chain Monte Carlo provide a set of flexible and reliable tools for inference in large classes of missing-data problems. Yet, in practical terms, these developments have had surprisingly little impact on the way most data analysts handle missing values on a routine basis. My hope is that this book will help to bridge the gap between theory and practice, making a multipurpose kit of missing-data tools accessible to anyone who may need them.

This book is intended for applied statisticians, graduate students and methodologically-oriented researchers in search of practical tools to handle missing data. The focus is applied rather than theoretical, but technical details have been included where necessary to help readers thoroughly understand the statistical properties of these methods and the behavior of the accompanying algorithms.

The methods presented here rely on three fully parametric models for multivariate data: the unrestricted multivariate normal distribution, loglinear models for cross-classified categorical data and the general location model for mixed continuous and categorical variables. In addition, the missing data are assumed to be missing at random, in the sense defined by Rubin (1976). My reviewers have correctly pointed out that many other vitally important topics could (and perhaps should) have been addressed: non-normal models such as the contaminated normal and multivariate-t; repeated measures and restricted covariance structures; censored and coarsened data; models for nonignorable nonresponse; latent variables; and hierarchical or random-effects models. Imputation for complex surveys and censuses, a topic in which I am deeply interested, deserves much more attention than it received. For better or worse, I decided to limit the material to a few important subjects, but to treat these subjects thoroughly and illustrate them with non-trivial data examples.

This book would not have been possible without the generous support and encouragement of many friends, colleagues and agencies. Don Rubin, whose countless contributions to the area of missing data provided a springboard for this work, was the first to suggest publishing it as a book. The initial round of software development was sponsored by Frank Sulloway, whose wonderfully incomplete dataset provided the first and most colorful application of these methods. Additional support was provided by the Bureau of the Census, the United States Department of Agriculture and the National Center for Health Statistics. Many helpful comments and suggestions were given by John Barnard, Rose Brunner, Andrew Gelman, Bonnie Ghosh, Xiao-Li Meng, Susan Murphy, Maren Olsen, Ritz Scheuren, Stef van Buuren, Recai Yucel and Alan Zaslavsky, and the editorial and production staff at Chapman & Hall. Data on the Foreign Language Attitude Scale were contributed by Mark Raymond. My parents, Chester and Dolores Schafer, created a loving and stable childhood environment, and my wife Sharon did not fail to encourage and inspire. Prayer support was provided by Dr. Samuel C. Lee and members of University Bible Fellowship.

Finally, I must acknowledge my debt to the late Clifford C. Clogg, to whom this book is dedicated. Cliff's steady encouragement and careful review greatly improved the quality of the book, especially the first five chapters. His warmth, love for learning, hard work and faith continue to inspire the many who were close to him. Personally and professionally, it is most gratifying to know that Cliff regarded this book as 'good stuff'.

Joseph L. Schafer
University Park, Pennsylvania
October 1996

CHAPTER 1

Introduction

1.1 Purpose

This book presents methods of statistical inference from multivariate datasets with missing values where missingness may occur on any or all of the variables. Such datasets arise frequently in statistical practice, but the tools for effectively dealing with them are not readily available to data analysts. It is our goal to provide these tools, along with the knowledge of how to use them.

When faced with missing values, practitioners frequently resort to ad hoc methods of *case deletion* or *imputation* to force the incomplete dataset into a rectangular complete-data format. Many statistical software packages, for example, automatically omit from a linear regression analysis any case t hat has a missing value for any variable. Imputation is a generic term for filling in missing data with plausible values. In a multivariate dataset, each missing value may be replaced by the observed mean for that variable, or, in a slightly less naive approach, by some sort of predicted value from a regression model. Almost invariably, after the dataset has been altered by one of these methods no additional provision for missing data is made in the subsequent analysis. The research usually proceeds as if the omitted cases had never really been observed, or as if the imputed values were real data.

When the incomplete cases comprise only a small fraction of all cases (say, five percent or less) then case deletion may be a perfectly reasonable solution to the missing-data problem. In multivariate settings where missing values occur on more than one variable, however, the incomplete cases are often a substantial portion of the entire dataset. If so, deleting them may be inefficient, causing large amounts of information to be discarded. Moreover, omitting them from the analysis will tend to introduce bias, to the extent that the incompletely observed cases differ systematically from the completely observed ones. The completely observed cases that remain will be unrepresentative of the population for which the inference is usually intended: the population of *all* cases, rather than the population of cases with no missing data.

Ad hoc methods of imputation are no less problematic. Imputing averages on a variable-by-variable basis preserves the observed sample means, but it distorts the covariance structure, biasing estimated variances and covariances toward zero. Imputing predicted values from regression models, on the other hand, tends to inflate observed correlations, biasing them away from zero. When the pattern of missingness is complex, devising an ad hoc imputation scheme that preserves important aspects of the joint distribution of the variables can be a daunting task. Moreover, even if the joint distribution of all variables could be adequately preserved, it may be a serious mistake to treat the imputed data as if they were real. Standard errors, p-values and other measures of uncertainty calculated by standard complete-data methods could be misleading, because they fail to reflect any uncertainty due to missing data.

This book presents a unified approach to the analysis of incomplete multivariate data. We will consider datasets for which the variables are continuous, categorical, or both. This approach allows one to analyze the data by virtually any technique that would be appropriate if the data were complete. This is accomplished not by simply modifying the data in an ad hoc fashion to make them appear complete, but by principled methods that account for the missing values, and the uncertainty they introduce, at each step of the analysis in a formal way. These methods tend to be computationally intensive, requiring more computer time than ad hoc alternatives. However, they do not require a heavy investment of analyst time, and can be applied to a wide variety of problems more or less routinely without special efforts to develop new technology unique to each problem. This book is written from an applied perspective, attempting to bring together theory, computational methods, data examples and practical advice in a single source.

1.2 Background

The methods presented here have their origins in two distinct bodies of statistical literature. The first concerns likelihood-based inference with incomplete data and, in particular, the EM algorithm. The second concerns techniques of Markov chain Monte Carlo: Gibbs sampling, data augmentation, the Metropolis-Hastings algorithm, and related methods.

1.2.1 The EM algorithm

The EM algorithm is a general technique for finding maximum likelihood estimates for parametric models when the data are not fully observed. Although special cases of EM appear far back in the statistical literature, it was not until Dempster, Laird and Rubin (1977) coined the term *EM* and established its fundamental properties that the generality and usefulness of this algorithm were realized. EM spawned a revolution in the analysis of incomplete data, making it possible to compute efficient parameter estimates, and thus obviating the need for ad hoc methods like case deletion, in wide classes of statistical problems.

The influence of EM has been far reaching, not merely as a computational technique, but as a paradigm for approaching difficult statistical problems. There are many statistical problems which, at first glance, may not appear to involve missing data, but which can be reformulated as missing-data problems: mixture models, hierarchical or random effects models, experiments with unbalanced data and many more. In the last fifteen years, a surprisingly large number of applications for EM have been found in a wide variety of fields. Unfortunately, major producers of statistical software have been rather slow to incorporate general-purpose EM algorithms for incomplete data into their products. One notable exception is BMDP, which has EM algorithms for the multivariate normal model and for unbalanced repeated measures with structured covariance matrices (BMDP Statistical Software, Inc., 1992).

1.2.2 Markov chain Monte Carlo

Markov chain Monte Carlo is a body of methods for generating pseudorandom draws from probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one. As we proceed along the sequence, provided that certain regularity conditions axe met, the distributions of the elements stabilize to a common distribution known as the *stationary distribution*. In Markov chain Monte Carlo, one constructs a Markov chain whose stationary distribution is a distribution of interest. By repeatedly simulating steps of the chain, one is able eventually to simulate draws from the distribution of interest.

The two most popular methods of Markov chain Monte Carlo are Gibbs sampling and the Metropolis-Hastings algorithm. In Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990), one draws from the conditional distribution of each component of a multivariate random variable given the other components in a cyclic fashion. In Metropolis-Hastings (Metropolis *et al.*, 1953; Hastings, 1970), one draws from a probability distribution intended to approximate the distribution actually of interest, and then accepts or rejects the drawn value with a specified probability. Many variations of these are possible, for example, hybrid algorithms that perform steps of Metropolis-Hastings within iterations of Gibbs. These methods are related to more traditional Monte Carlo methods such as importance sampling (e.g. Kleijnan, 1974) and rejection sampling (e.g. Kennedy and Gentle, 1980).

As with EM, specific applications of Markov chain Monte Carlo have been in use for many years, notably in areas of statistical mechanics and image reconstruction. In the past decade, however, many new uses for these methods have been discovered and implemented that are of special interest to statisticians. In particular, Markov chain Monte Carlo has spawned a revolution of its own in the area of applied Bayesian inference.

In Bayesian inference, information about unknown parameters is expressed in the form of posterior probability distribution. Even with relatively simple probability models, the posterior distribution is often intractable: important summaries such as moments, marginal densities and quantiles are not readily available in closed form. Practitioners have typically resorted to asymptotic approximation, numerical integration and importance sampling to elicit meaningful summaries of intractable posteriors. Through Markov chain Monte Carlo, however, it is now possible in many cases to simulate the entire joint posterior distribution of the unknown quantities, and thereby obtain simulation-based estimates of virtually any features of the posterior that are of interest.

1.3 Why analysis by simulation?

Simulation of posterior distributions enjoys many advantages over more traditional methods of parametric inference. Some of these axe listed below.

1. In complex problems it may be easier to implement than other methods, both conceptually and computationally.

2. It may be the only method currently feasible when the unknown parameter is of high dimension.

3. It does not rely on asymptotic approximations. The algorithms converge scholastically to posterior distributions that are exact, regardless of sample size.

In an era when computing environments are becoming increasingly powerful and less expensive, simulation promises to be one of the mainstays of applied parametric modeling and data analysis in the years ahead.

Simulation is especially attractive at the present time as a general approach to the analysis of incomplete multivariate data. There are at least two major reasons for this. First, simulation by Markov chain Monte Carlo is a natural companion and complement to the current tools for handling missing data, and, in particular, the EM algorithm. Markov chain Monte Carlo can be applied to precisely the same types of problems as EM, and, computationally speaking, its implementation is often remarkably similar to that of EM. Whereas EM provides only point estimates of the unknown parameters, however, Markov chain Monte Carlo provides random draws from their joint posterior distribution. A point estimate, even if it is efficient, is not especially useful unless there is also some measure of uncertainty associated with it. With Markov chain Monte Carlo, Bayesian analogues of the standard tools of frequentist inference (standard errors, confidence intervals and p-values) are now readily simulated, providing these measures of uncertainty.

A second reason why simulation is a natural choice for missing-data problems is that it facilitates inference by *multiple imputation*. Multiple imputation (Rubin, 1987) is a technique in which each missing value is replaced by $m > 1$ simulated values. The m sets of imputations reflect uncertainty about the true values of the missing data. After the multiple imputations axe created, m, plausible versions of the complete data exist, each of which are analyzed by standard complete-data methods. The results of the m analyses are then combined to produce a single inferential statement (e.g. a confidence interval or a p-value) that includes uncertainty due to missing data.

Until now, the task of generating multiple imputations has been problematic except in some simple cases, such as univariate examples and datasets with only one variable subject to nonresponse. No straightforward, general-purpose algorithms have been available for generating proper multiple imputations in a multivariate setting. Using techniques of Markov chain Monte Carlo, however, it is now possible to do this quite easily.

Like other methods of inference, simulation based on Markov chains has certain disadvantages. Carrying out a simulation for a large dataset or a complicated model may require access to a fast computer with substantial memory. Monitoring the convergence of Markov chain Monte Carlo algorithms can be difficult. Moreover, the use of the Bayesian paradigm and the introduction of prior distributions for unknown parameters, even if the impact on conclusions is minimal, may be regarded by some as artificial or undesirably subjective. In the chapters ahead, we will try to address these issues carefully and thoughtfully as they arise.

1.4 Looking ahead

This book presents iterative algorithms for simulating multiple imputations of missing values in incomplete datasets under some important classes of multivariate models. The same algorithms may also be used to draw values of parameters from their posterior distributions. The algorithms are described in detail, focusing on practical issues of computation. The computational efficiency and low data-storage requirements of the algorithms make them suitable even for datasets that are quite large, and they have been applied routinely to datasets with over 10 000 observations and 30 variables. The use of these algorithms is demonstrated on a variety of real data examples, with accompanying discussion on issues of practical importance to data analysts.

Because of their good performance, we believe that these algorithms will find widespread use in a variety of applications. We expect that they will become standard supplements to the current tools of missing-data analysis. Perhaps the most important aspect of this work is that now, for the first time, multiple imputation and parameter simulation are made available to nonspecialists who know the importance of adjusting for missing data in their inference, but who lack the resources or special expertise needed to develop and implement these techniques on their own.

1.4.1 Scope of the rest of this book

Chapter 2 presents the key assumptions that will be made throughout this book, the parametric data model and the assumption of ignorable missingness, and discusses their relevance in various applied settings. Chapter 3 presents necessary background material on EM and Markov chain Monte Carlo. Chapter 4 discusses in practical terms the various methods of conducting inference by simulation. The remaining chapters describe algorithms for specific multivariate models and illustrate their use in a variety of examples. Chapters 5-6 discuss methods for the multivariate normal distribution; Chapters 7-8, models for cross-classified categorical data; and Chapter 9, multivariate models for datasets with both continuous and categorical variables.

Chapters 3 and 4 serve as a reference for the subsequent chapters. Readers interested primarily in applications may find it helpful to initially skim through Chapters 3 and 4 and then return to them as necessary while working through Chapters 5-9.

1.4.2 Knowledge assumed on the part of the reader

We assume that the reader is familiar with basic concepts of probability theory, inference based on the likelihood function, and multivariate distributions, especially the multivariate normal and the multinomial. Matrix notation will be used throughout. We assume that the reader is also comfortable with the basic concepts of Bayesian inference, although not necessarily having experience with applying Bayesian techniques in real examples. Some knowledge of standard categorical-data techniques, especially loglinear models, is also helpful but not absolutely necessary.

1.4.3 Software and computational details

The algorithms described in this book have been implemented by the author for general use as functions in the statistical language S (Becker, Chambers, and Wilks, 1988), using subroutines written in Fortran-77. The programs are available to anyone free of charge, and information on them is provided in Appendix C.

As you read this book, especially the later chapters, you may be surprised at the unusual amount of attention devoted to computational issues. Enough detail has been provided to enable a dedicated reader to reinvent the crucial portions of computer programs, if he or she chooses to do so. These details were provided for the following reasons:

1. to encourage others to implement the algorithms in other computer languages or software packages, if they are better served by these environments;

2. to encourage others to improve upon these algorithms, if they discover ways to make them more efficient; and

3. to foster development of similar algorithms for more general classes of models, perhaps using these routines as building blocks for larger and more complex programs.

1.5 Bibliographic notes

A general overview of techniques for missing data, with discussion of various ad hoc approaches as well as the EM algorithm, is given by Little and Rubin (1987). The original article on EM by Dempster, Laird, and Rubin (1977) with discussion, now almost twenty years old, still provides a helpful introduction to EM; its simple examples anticipate many of the major types of EM algorithms in use today. For a comprehensive bibliographic review of EM, see Meng and Pedlow (1992).

Excellent overviews of Markov chain Monte Carlo methods, including data augmentation, Gibbs, sampling, and the Metropolis-Hastings algorithm, appear in books by Tanner (1993) and Gilks, Richardson, and Spiegelhalter (1996). Tanner's book also contains an entire chapter on EM.

A classic introduction to Bayesian inference is given by Box and Tiao (1992), and Gelman *et al.* (1995) discuss practical Bayesian data analysis from a modern perspective.

Good reference material on cross-classified categorical data and loglinear models is given by Bishop, Fienberg and Holland (1975) and Agresti (1990).

CHAPTER 2

Assumptions

2.1 The complete-data model

We will consider rectangular datasets whose rows can be modeled as independent, identically distributed (iid) draws from some multivariate probability distribution. A schematic representation of such a dataset is shown in Figure 2.1. The n rows represent observational units and the p columns represent variables recorded for those units. Missing values, denoted by question marks, may occur in any pattern.

Let Y denote the $n \times p$ matrix of complete data, which is not fully observed, and let y_i denote the ith row of Y, $i = 1$.

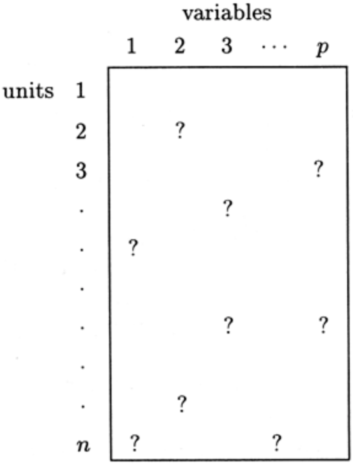


Figure 2.1. *Multivariate dataset with missing values.*

By the iid assumption, the probability density or probability function of the complete data may be written

$$P(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta), \tag{2.1}$$

where *f* is the density or probability function for a single row, and θ is a vector of unknown parameters. We will consider three classes of distributions *f*:

1. the multivariate normal distribution;
2. the multinomial model for cross-classified categorical data, including loglinear models; and
3. a class of models for mixed normal and categorical data (Krzanowski, 1980, 1982; Little and Schluchter, 1985).

On occasion, the two crucial modeling assumptions above, that the rows are iid, and that the we have correctly specified (up to the unknown θ) the full joint distribution of all *p* variables, will not be needed in their entirety and may be partially relaxed. We will sometimes be able to accommodate situations like regression, in which we seek to model the conditional distribution of one or more response variables given some predictor variables without specifying any probability model for the predictors. A discussion of this point will be given in Section 2.6. For now, we turn our attention to the mechanism of missingness.

2.2 Ignorability

2.2.1 Missing at random

Denote the observed part of *Y* by Y_{obs}, and the missing part by Y_{mis}, so that $Y = (Y_{obs}, Y_{mis})$. Throughout this book, we will assume that the missing data are *missing at random* (MAR) in the sense defined by Rubin (1976).

A precise definition for MAR will be given momentarily, but first we describe it in an informal way: the probability that an observation is missing may depend on Y_{obs}, but not on Y_{mis}, Another useful heuristic definition of MAR is the following. Let U and V be any two variables or non-overlapping groups of variables. Suppose that we restrict attention to units for which U is observed and equal to a specific value, say u. MAR means that among these units, the distribution of V is, apart from ordinary sampling variability, the same among the cases for which V is observed as it is among the cases for which V is missing.

Despite its name, then, MAR does not suggest that the missing data values are a simple random sample of all data values. The latter condition is known as *missing completely at random* (MCAR). MCAR is only a special case of MAR. MAR is less restrictive than MCAR because it requires only that the missing values behave like a random sample of all values within subclasses defined by observed data. In other words, MAR allows the probability that a datum is missing to depend on the datum itself, but only indirectly through quantities that are observed.

More formally, Rubin (1976) defines MAR in terms of a probability model for the missingness. Let R be an $n \times p$ matrix of indicator variables whose elements are zero or one depending on whether the corresponding elements of Y are missing or observed. We would not in general expect the distribution of R to be unrelated to Y, so we posit a probability model for R, $P(R|Y, \xi)$, which depends on Y as well as some unknown parameters ξ. The MAR assumption is that this distribution does not depend on Y_{mis},

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi) \qquad (2.2)$$

2.2.2 Distinctness of parameters

To proceed further, we also need to assume that the parameters θ of the data model and the parameters ξ of the missingness mechanism are *distinct*. From a frequentist perspective, this means that the joint parameter space of (θ,ξ) must be the Cartesian cross-product of the individual parameter spaces for θ and ξ. From a Bayesian perspective, this means that any joint prior distribution applied to (θ,ξ) must factor into independent marginal priors for θ and ξ. In many situations this is intuitively reasonable, as knowing θ will provide little information about ξ and vice-versa. If both MAR and distinctness hold, then the missing-data mechanism is said to be *ignorable* (Little and Rubin, 1987; Rubin, 1987).

2.3 The observed-data likelihood and posterior

2.3.1 Observed-data likelihood

Following arguments given by Rubin (1976) and Little and Rubin (1987), it can be shown that under ignorability, we do not need to consider the model for R nor the nuisance parameters ξ when making likelihood-based or Bayesian inferences about θ.

Because the observed data truly consist not only of Y_{obs}, but also of R, the probability distribution of the observed data is actually given by

$$P(R, Y_{obs} | \theta, \xi) = \int P(R, Y | \theta, \xi) dY_{mis}$$

$$= \int P(R | Y, \xi) P(Y | \theta) dY_{mis,} \qquad (2.3)$$

where the integral is understood to mean summation for distributions that are discrete. Under the MAR assumption, (2.3) becomes

$$P(R, Y_{obs} \Big| \theta, \xi) = P(R \Big| Y_{obs}, \xi) \int P(Y \big| \theta) dY_{mis}$$

$$= P(R|Y_{obs}, \xi) P(Y_{obs}|\theta). \qquad (2.4)$$

The likelihood of the observed data under MAR can thus be factored into two pieces, one pertaining to the parameter of interest θ and the other pertaining to the nuisance parameter ξ. When the two parameters are distinct, then likelihood-based inferences about θ will be unaffected by ξ or $P(R|Y_{obs},\xi)$. Maximum-likelihood estimation of θ, likelihood-ratio tests concerning θ, and so on can then be performed without regard for the missing-data mechanism; that is, the missing-data mechanism may be safely ignored.

The factor in (2.4) pertaining to θ (or, more precisely, any function proportional to this factor) is referred to by Little and Rubin (1987) as the likelihood ignoring the missing-data mechanism,

$$L(\theta|Y_{obs}) \propto P(Y_{obs}|\theta). \qquad (2.5)$$

For brevity, we will refer to (2.5) as the *observed-data likelihood*, although that name should, strictly speaking, be reserved for the complete function (2.4). Because we assume ignorability throughout, however, there is never a need to work with the complete function (2.4), and thus there will be no ambiguity.

Notice that at first glance, the factorization (2.4) seems to contain no implicit assumptions about the missingness mechanism. The joint distribution of any two random variables, say Z_1 and Z_2, can always be written as the marginal distribution of Z_1 multiplied by the conditional distribution of Z_2 given Z_1. A subtle but important difference exists between this basic rule of probability and the factorization (2.4), however, and the distinction lies in the definition of θ. In our framework, θ refers to the parameters of the model for the complete data Y = $(Y_{obs}\ Y_{mis})$, not the parameters for the distribution of Y_{obs}, alone. We assume that the ultimate goal of the analysis is to draw inferences about the parameters of the complete-data model, not the parameters governing the marginal distribution of Y_{obs}. If θ were redefined to pertain only to Y_{obs} then assumptions like MAR would not be necessary. This approach has some major conceptual difficulties, however, and may lead to results that are very hard to interpret, so we will not consider it further. For an interesting discussion related to this point, see the exchange between Efron (1994) and Rubin (1994).

2.3.2 Examples

Example 1: Incomplete univariate data. Suppose that a single variable is observed for units 1, 2,...$n_1 < n$ and missing for units $n_1 + 1,..., n$. Let $Y = (y_1, y_2,y_n)$ denote the complete data and $R = (r_1, r_2,...,r_n)$ the response indicators, where $r_i = 1$ if y_i is observed and $r_i = 0$ if y_i is missing. If the distribution of R does not depend on Y then the missingness mechanism is MAR. In fact, in this case it is MCAR. One such mechanism is simple Bernoulli selection in which each unit is observed with probability independently of all other units,

$$P(R|Y,\xi) = \prod_{i=1}^{n} \xi^{r_i}(1-\xi)^{1-r_i}.$$

Another MAR mechanism arises when the responding units are a simple random sample of all units, otherwise.

$$P(R|Y,\xi) = \begin{cases} \binom{n}{n_1}^{-1} & if \sum_{i=1}^{n} r_i = n_1, \\ 0 & \text{otherwise.} \end{cases}$$

The latter regards n_1 as fixed whereas the former regards n_1 as random. Under either of these mechanisms or any other mechanism that is MAR, it is appropriate to base inferences about parameters of the distribution of Y on the observed-data likelihood. This likelihood may be written

$$L(\theta|Y_{obs}) = \int P(Y|\theta)dY_{mis}$$

$$= \int \cdots \int \prod_{i=1}^{n_1} P(y_i \,|\, \theta) \prod_{i=n_1+1}^{n_1} P(y_i \,|\, \theta) dy_{n_1+1} \cdots dy_n \,.$$

The first product in the integrand does not involve Y_{mis} and can be brought out of the integral, and the second product integrates

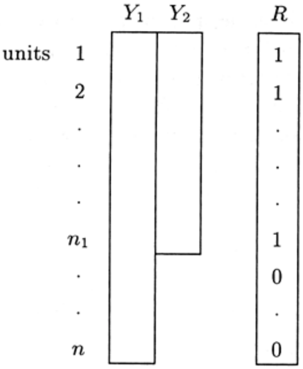


Figure 2.2. *Bivariate data with one variable subject to nonresponse.*

to one, yielding

$$L\left(\theta|Y_{obs}\right) = \prod_{i=1}^{n_1} P\left(y_i|\theta\right),$$

which is simply a complete-data likelihood based on the reduced sample $y_1, \ldots, y_{n_1}$.

Example 2: Bivariate data with one variable subject to nonresponse. Consider a dataset with variables Y_1 and Y_2 as shown in Figure 2.2, where Y_1 is observed for units 1, 2,..., n but Y_2 is observed only for units 1,2,...,$n_1 < n$ .The missing data will be MAR if the probability that Y_2 is missing does not depend on Y_2, although it may possibly depend on Y_1. Let y_{i1}, and y_{i2} denote the values of Y_1 and Y_2, respectively, for unit i. The observed-data likelihood may be written

$$\int \prod_{i=1}^{n_1} P(y_{i1}, y_{i2}|\theta) \prod_{i=n_1+1}^{n} P(y_{i1}|\theta) \prod_{i=n_1+1}^{n} P(y_{i2}|y_{i1}, \theta) dY_{mis}.$$

The first two products in the integrand do not involve Y_{mis} and the last product integrates to one, hence

$$L(\theta \mid Y_{obs}) = \prod_{i=1}^{n_1} P(y_{i1}, y_{i2} \mid \theta) \prod_{i=n_1+1}^{n} P(y_{i1} \mid \theta) \qquad (2.6)$$

$$= \prod_{i=1}^{n} P(y_{i1} \mid \theta) \prod_{i=1}^{n_1} P(y_{i2} \mid y_{i1}, \theta). \qquad (2.7)$$

For example, suppose that Y_1 and Y_2 have a bivariate normal distribution with parameter

$$\theta = \left(\mu_1, \sigma_{11}, \mu_2, \sigma_{22}, \sigma_{12}\right),$$

where $\mu_i = E\left(Y_i \mid \theta\right)$and $\sigma_{ij} = \mathrm{Cov}\left(Y_i Y_j \mid \theta\right), i, j = 1, 2$. The observed-data likelihood may be written as in (2.6),

$$L(\theta \mid Y_{obs}) \propto |\Sigma|^{-n_1/2} \left\{ -\frac{1}{2} \sum_{i=1}^{n_1} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\} \right\}$$

$$\times \sigma_{11}{}^{-(n-n_1)/2} \exp\left\{ -\frac{1}{2\sigma_{11}} \sum_{i=n_1+1}^{n} (y_{i1} - \mu_1)^2 \right\}, \qquad (2.8)$$

where $y = (y_{i1}, y_{i2})^T$, $\mu = \left(\mu_1, \mu_2\right)^T$and Σ is the 2 × 2 matrix with elements σ_{ij}. Alternatively, the parameter of the bivariate normal distribution may be expressed as

$$\phi = (\mu_1, \sigma_{11}, \beta_0, \beta_1, \sigma_{22\cdot 1}),$$

where $\beta_1 = \sigma_{12} / \sigma_{11}, \beta_0 = \mu_2 - \beta_1 \mu_1$ and $\sigma_{22\cdot 1} = \sigma_{22} - \sigma_{12}^2 / \sigma_{11},$ so that $E(Y_2 \mid Y_1, \varphi) = \beta_0 + \beta_1 Y_1$ and $V(Y_2 \mid Y_1, \varphi) = \sigma_{22\cdot 1}.$ The transformation $\varphi = \varphi(\theta)$ is one-to-one. Following (2.7), the observed data likelihood may be written in terms of φ as

$$L(\theta \mid Y_{obs}) \propto \sigma_{11}^{-n/2} \exp\left\{ -\frac{1}{2\sigma_{11}} \sum_{i=1}^{n_1} (y_{i1} - \mu_1)^2 \right\} \qquad (2.9)$$

$$\times \sigma_{22\cdot 1}^{-n/2} \exp\left\{ -\frac{1}{2\sigma_{22\cdot 1}} \sum_{i=1}^{n} (y_{i2} - \beta_0 - \beta_1 y_{i1})^2 \right\},$$

an expression first given by Anderson (1957).

Expressions (2.8) and (2.9) are equivalent, but the latter has A, convenient interpretation as the product of two complete-data likelihood functions: the univariate normal likelihood for Y_1 based on units 1, 2,..., n, and the likelihood for the normal linear regression of Y_2 on Y_1 based on units 1, 2,..., n_1. Because the parameters $\varphi_1 = (\mu_1, \sigma_{11})$ and $\varphi_2 = (\beta_0, \beta_1, \sigma_{22\cdot 1})$ corresponding to these two factors are distinct, inferences about them may proceed independently. For example, maximum-likelihood estimates may be obtained by independently maximizing the likelihoods for φ_1 and φ_2, each of which corresponds to a straightforward complete-data problem. Expression (2.8) also appears to be the product of two complete-data likelihoods, but the parameters in the two factors are not distinct because μ_1 and σ_{11} appear in both.

Example 3: Multivariate normal data with arbitrary patterns of missing values. Now consider a p-variate normal data matrix with missing values on any or all variables as in Figure 2.1. It is convenient to group the rows of the matrix according to their missingness patterns. A missingness pattern is a unique combination of response statuses (observed or missing) for Y_1, Y_2,..., Y_p. With p variables there are 2^p possible missingness patterns: It is usually the case, especially when p

is large, that not all possible patterns are represented in the sample. Index the unique missingness patterns that actually appear in the sample by s, where $s = 1, 2,\ldots, S$, and let $I(s)$ denote the subset of the rows $i = 1, 2,\ldots, n$ that exhibit pattern s. A generalization of the arguments leading to (2.6) and (2.8) allows us to write the observed-data likelihood as

$$\prod_{s=1}^{S} \prod_{i \in I(s)} \left|\Sigma_s^*\right|^{½} \exp\left\{-\tfrac{1}{2}(y_i^* - \mu_s^*)^T \Sigma_s^{*-1}(y_i^* - \mu_s^*)\right\}, \qquad (2.10)$$

where y_i^* denotes the observed part of row i of the data matrix, and μ_s^* and Σ_s^* denote the subvector of the mean vector and the square submatrix of the covariance matrix Σ, respectively, that pertain to the variables that are observed in pattern s. Notice that if any rows of the data matrix are completely missing, then those rows drop out of the observed-data likelihood; under the ignorability assumption, these rows contribute nothing to the inference and may be ignored.

Despite the concise appearance of (2.10), this likelihood tends to be a complicated function of the individual means μ_i and covariances σ_{ij}, $i ,j = 1, 2,\ldots,p$. Except in special cases, there is no way to express this likelihood as in (2.7) and (2.9), the product of simple complete-data likelihoods whose parameters are distinct (Rubin, 1974). Moreover, the first two derivatives of (2.10) or its logarithm with respect to the individual μ_i and σ_{ij} tend to be very complicated as well, making (2.10) awkward to maximize by gradient methods such as Newton-Raphson. A much more convenient method for maximizing this likelihood is provided by the EM algorithm (Beale and Little, 1975; Dempster, Laird, and Rubin, 1977), to be introduced in Chapter 3.

The complicated nature of (2.10) is typical of the observed-data likelihood functions one encounters with incomplete multivariate data. Except in special cases, meaningful summaries of these functions (e.g. modes) are not available in closed form, nor are they readily computable from classical numerical methods; we typically need to resort to special iterative techniques like EM.

2.3.3 Observed-data posterior

Definition

In the Bayesian framework all inferences are based on a posterior probability distribution for the unknown parameters that conditions on the quantities that are observed. Returning to the notation of Section 2.3.1, the unknown parameters are (θ,ξ) and the observed quantities are Y_{obs}, and R. By Bayes s Theorem, the posterior distribution may be written as

$$P(\theta,\xi \mid Y_{obs}, R) = k^{-1} P(R, Y_{obs} \mid \theta, \xi)\pi(\theta,\xi), \tag{2.11}$$

where $\pi(\cdot)$ denotes a prior distribution applied to (θ,ξ) and k is the normalizing constant

$$k = \int\int P(R, Y_{obs} \mid \theta\xi)\pi(\theta,\xi)d\theta\, d\xi.$$

Under the assumption of MAR, we may substitute (2.4) into (2.11) to obtain

$$P(\theta,\xi \mid Y_{obs}, R) \propto P(R \mid Y_{obs}, \xi)P(Y_{obs} \mid \theta)\pi(\theta,\xi).$$

Bayesian inferences about θ alone are based on the marginal posterior obtained by integrating this function over the nuisance parameter ξ. When θ and ξ are distinct according to the definition in Section 2.2.2, the prior distribution factors as

$$\pi(\theta,\xi) = \pi_\theta(\theta)\pi_\xi(\xi).$$

Hence the marginal posterior for θ is, under ignorability,

$$P(\theta \mid Y_{obs}, R) = \int P(\theta, \xi \mid Y_{obs}, R) d\xi$$

$$\propto P(Y_{obs} \mid \theta) \pi_{\theta}(\theta) \int P(R \mid Y_{obs}, \xi) \pi_{\xi}(\xi) d\xi$$

$$\propto L(\theta \mid Y_{obs}) \pi_{\theta}(\theta), \tag{2.12}$$

where the proportionality is up to a multiplicative factor that does not involve θ. Note that R does not appear on the right-hand side of (2.12) and therefore $P(\theta \mid Y_{obs}, R) = P(\theta \mid Y_{obs})$. We have thus shown that under ignorability all information about θ is summarized in the posterior that ignores the missing-data mechanism,

$$P(\theta \mid Y_{obs}) \propto L(\theta \mid Y_{obs}) \pi_{\theta}(\theta). \tag{2.13}$$

We shall refer to (2.13) as the *observed-data posterior*.

In most practical applications one would not be interested in the function (2.13) itself but in meaningful summaries of it: posterior moments, marginal posterior densities and quantiles of univariate components of θ, etc. Note that these summaries are typically integrals of the density (2.13) or functions involving it over the parameter space. In many commonly used probability models with complete data, computation of these integrals can be simplified by choosing a prior distribution for θ within a natural conjugate class (e.g. Box and Tiao, 1992). With incomplete data, however, the usual natural conjugate priors no longer lead to posteriors that are recognizable or easy to summarize.

A bivariate normal example

Let us return to Example 2 of Section 2.3.2 in which Y_1 is observed for all units but Y_2 is missing for some. Assuming that the complete data are bivariate normal with parameter $\theta = (\mu, \Sigma)$, the observed data likelihood is given by (2.8). In the absence of strong prior beliefs about θ, a prior distribution commonly used with complete data is

$$\pi_\theta(\theta) \propto |\Sigma|^{-3/2}, \tag{2.14}$$

which can be derived by applying the invariance principle of Jeffreys to μ and Σ (e.g. Box and Tiao, 1992). The prior (2.14) is said to be *improper* because it is not a true probability density function; its integral over the parameter space is not finite. With complete data this prior leads to a posterior distribution for θ that is the product of an inverted-Wishart distribution for Σ and a normal distribution for Σ given μ. The properties of this normal-inverted Wishart distribution are well known, and summaries (marginal densities, moments, etc.) are readily available in closed form. When some values of Y_2 are missing, however, the posterior under (2.14) is no longer normal-inverted Wishart.

One way to characterize this posterior is to express it in terms of the alternative parameterization $\varphi = \left(\mu_1, \sigma_{11}, \beta_0, \beta_1, \sigma_{22\cdot 1}\right)$, where $\beta_1 = \sigma_{12} / \sigma_{11}, \beta_0 = \mu_{2-}\beta_1\mu_1$ and $\sigma_{22\cdot 1} = \sigma_{22} - \sigma_{12}^2 / \sigma_{11.}$ As shown in (2.9), the likelihood for φ factors neatly into a complete-data likelihood for φ_1=(μ,σ_{11}) based on all the sample units and a complete-data likelihood for φ_2=(β_0, β_1,$\sigma_{22\cdot 1}$) based on the units for which Y_2 is observed. Moreover, the prior distribution (2.14) also factors into independent priors for φ_1 and φ_2. To see this, note that φ = $\varphi(\theta)$ is a one-tone transformation, and the density for φ implied by (2.14) can thus be written as

$$\pi_\phi(\phi) = \pi_\theta\left(\phi^{-1}(\phi)\right)\|J\|^{-1},$$

where $\theta = \varphi^{-1}(\varphi)$ denotes the transformation from φ back to θ, and $\|J\|$ denotes the absolute value of the determinant of the Jacobian (first-derivative) matrix for the transformation from

θ to φ. Notice that $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22\cdot1}$. Moreover, it will be shown in Section 5.2.4 that $\|J\| = \sigma_{11}^{-1}$, and thus

$$\pi_\phi(\phi) = \sigma_{11}^{-1/2}\sigma_{22\cdot1}^{-3/2}. \tag{2.15}$$

Combining (2.15) with the likelihood (2.9), the observed-data posterior can be written as

$$P(\phi \mid Y_{obs}) = P(\phi_1 \mid Y_{obs})P(\phi_2 \mid Y_{obs}), \tag{2.16}$$

where

$$P(\phi_1 \mid Y_{obs}) \propto \sigma_{11}^{-(n+1)/2} \exp\left\{-\frac{1}{2\sigma_{11}}\sum_{i=1}^{n}(y_{i1} - \mu_1)^2\right\}$$

and

$$P(\varphi_2 \mid Y_{obs}) \propto \sigma_{22\cdot1}^{-(n_1+3)/2} \exp\left\{-\frac{1}{2\sigma_{22\cdot1}}\sum_{i=1}^{n_1}(y_{i2} - \beta_0 - \beta_1 y_{i1})^2\right\}.$$

After some manipulation, it can be shown that $P(\varphi_1 \mid Y_{obs})$ is the product of a normal and a scaled-inverted chisquare density,

$$\mu_1 \mid \sigma_{11}, Y_{obs} \sim N\left(\bar{y}_1, n^{-1}\sigma_{11}\right),$$

$$\sigma_{11} \mid Y_{obs} \sim (n-1)S_{11}\chi_{n-2}^{-2}, \tag{2.17}$$

where $\bar{y}_1$ and S_{11} are the usual sample mean and variance of Y_1, based on all n units and χ_{n-2}^{-2} denotes an inverted chisquare variate n (i.e. the reciprocal of a chisquare variate) with n - 2 degrees of freedom. Moreover, $P(\varphi_2 \mid Y_{obs})$ can be shown to be the product of a bivariate normal and an inverted chisquare,

$$\beta \mid \sigma_{22\cdot 1}, Y_{obs} \sim N\left(\hat{\beta}, \sigma_{22\cdot 1}\left(X^T X\right)^{-1}\right),$$

$$\sigma_{22\cdot 1} \mid Y_{obs} \sim \hat{\varepsilon}^T \hat{\varepsilon} \chi^{-2}_{n_1 - 1}, \qquad (2.18)$$

where $\beta = (\beta_0, \beta_1)^T$,

$$X = \begin{bmatrix} 1 & y_{11} \\ 1 & y_{21} \\ \vdots & \vdots \\ 1 & y_{n_1,1} \end{bmatrix}, \quad y = \begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n_1,2} \end{bmatrix},$$

$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$ and $\hat{\varepsilon} = y - X\hat{\beta}$. Details of the calculations leading to these posteriors may be found in standard texts on Bayesian analysis and will be reviewed in Chapters 5-6.

In the above example, a particular factorization of the observed data likelihood enabled us to express the posterior in a tractable form. This will not always be the case. One cannot always factor the observed-data likelihood into complete-data likelihoods whose parameters are distinct. The techniques of Markov chain Monte Carlo to be introduced in Chapter 3 will free us from many of the constraints of mathematical tractability, allowing us to create random draws from the observed-data posterior whether or not it can be written in a tractable form.

2.4 Examining the ignorability assumption

The statistician unaccustomed to missing-data problems might be led to believe that the observed-data likelihood is *always* the relevant likelihood function for θ whenever data are not fully observed; it is, after all, simply the marginal distribution of Y_{obs} the observed part of Y. But as we have seen, the observed data consist of both Y_{obs}, and R, and one needs the special condition of ignorability to make the observed value of

R noninformative with respect to θ. Therefore it is crucial for the data analyst to understand the implications of the ignorability assumption, particularly MAR, and assess its appropriateness in any given problem.

2.4.1 Examples where ignorability is known to hold

On occasion, the MAR condition is known to hold exactly. Some examples of this are given below.

Double sampling. In sample surveys that employ double sampling (e.g. Cochran, 1977), some characteristics, say Y_1, Y_2,..., Y_k, are recorded for all units in the sample, and then additional characteristics $\mathrm{Y}_{k+1}, \ldots, \mathrm{Y}_p$ are recorded for a subsample of the original sample. If this subsample is selected by a mechanism that depends on Y_1, Y_2,..., Y_k alone, even in a systematic or deterministic way, then the missing values of $\mathrm{Y}_{k+1}, ..., \mathrm{Y}_p$ for those units not included in the subsample are MAR.

Sampling for nonresponse followup. In censuses and large surveys, initial attempts to collect data may fail for a substantial proportion of units. In a mail-based household survey, for example, some residents will inevitably fail to mail back their forms. In many cases, more intensive data-collection efforts (e.g. personal interviews) would be successful, but attempting to follow up every nonresponding unit may be economically infeasible. If the intensive followup effort is applied to a random sample of nonresponding units, then the missing data for the remaining nonrespondents is MAR. A famous early discussion of this method is given by Hansen and Hurwitz (1946).

Randomized experiments with unequal numbers of cases per treatment group. In many designed experiments, the researcher strives to assign an equal number of cases or subjects to each treatment, because data that are balanced in this fashion are typically easier to analyze than data that are unbalanced. Moreover, principles of efficiency often support the use of balanced designs. Sometimes balance is not feasible, however, and the data are unbalanced by design. The analysis of unbalanced data can often be simplified by imagining a number of additional cases which, if they were included in the experiment under the appropriate treatment groups, would result in a balanced experiment. Because the hypothetical missing data within each treatment group were missing with probability one, they are MAR.

Medical screening with multiple tests, where not all tests are administered to all subjects. In many medical studies, an inexpensive or easily administered test is given to a large number of subjects, and for purposes of calibration a second, more expensive, and more reliable test is administered to a subsample. The calibrating sample may be chosen completely at random, on the basis of subject specific covariates, or even on the basis of the outcome of the first test. As long as all the information used to choose the subsample is recorded and regarded as part of the observed data Y_{obs}, then the missing data will be MAR.

Matrix sampling for questionnaire or test items. In matrix sampling (e.g. Thayer, 1983) a test or questionnaire is divided into sections, and groups of sections are administered to subjects in a randomized fashion. The resulting data matrix will have rectangular patches of missing data corresponding to sections of the test or questionnaire that were not administered to subject groups. If all the variables used in the sampling plan are included in Y_{obs}, then the missing data will be MAR.

In most of the above examples, the missing data may be said to be *missing by design*, because it was never the intention of the investigator to record all potential variables for all subjects. When missing data are missing by design, they tend also to be MAR.

2.4.2 Examples where ignorability is not known to hold

In many other missing-data contexts, however, it is not known whether or not the MAR condition is satisfied. Examples include:

1. Sample surveys where some sampled individuals are not at home, unwilling to be interviewed, or do not otherwise provide useful responses to some or all of the questionnaire items. Notice, however, that if followup data can later be obtained for a probability sample of nonrespondents, the missing data can be converted to MAR.

2. Planned experiments where, for reasons unforeseen or unintended by the investigator, one or more outcomes of interest cannot be recorded: culture dishes break, production runs fail, subjects drop out of the study, etc.

3. Observational studies in which data of economic, historic or other scientific interest are collected for analysis, but for reasons beyond the control of the investigator some of the variables desired are simply not available for some cases.

Sometimes the fact that a numerical observation is not recorded is more like a response than a missing value. In a laboratory experiment, for example, an animal may die for reasons related to the treatment that was applied to it. If so, then the hypothetical missing data, the measurements that could not be recorded because the animal died, are counterfactual and poorly defined. In such instances, careful thought should be given to whether it is sensible to analyze such data by missing-data methods.

When data are missing for reasons beyond the investigator s control, one can never be certain whether MAR holds. The MAR hypothesis in such datasets cannot be formally tested unless the missing values, or at least a sample of them, are available from an external source. When such an external source is unavailable, deciding whether or not MAR is plausible will necessarily involve some guesswork, and will require careful consideration of conditions specific to the problem at hand.

2.4.3 Ignorability is relative

One final point that must be made is that MAR and ignorability are relative, defined with respect to a particular set of observed data Y_{obs}. In many situations, the status of the missing data (whether or not they are MAR) may change if the definition of Y_{obs} is changed. For example, consider a sample survey that involves probability sampling for nonresponse followup. Let Y_{mis} refer to the data for nonrespondents not included in the followup effort. If Y_{obs} is the data for nonrespondents who were included in the followup effort, then the missing data are MAR. If the definition of Y_{obs} is expanded to include the original respondents, however, and no information (e.g. dummy indicator) is retained to distinguish them from the nonrespondents who were subsequently followed up, then the missing data are no longer MAR.

In other situations, a nonresponse mechanism may not be exactly known to the analyst, but covariates are available that could plausibly explain or predict the missingness to a great extent. The plausibility of MAR would then depend on whether these covariates are included in the analysis. Further discussion of this point is given by Graham *et al.* (1994) with regard to attrition in a longitudinal study of adolescent drug use.

2.5 General ignorable procedures

Virtually all of the missing-data procedures used in statistical practice, both ad hoc approaches and principled ones, rely at least implicitly on an assumption of ignorability. Often the assumptions made are even stronger. For example, the case-deletion method used by many statistical packages (omitting all incomplete cases from the analysis) may introduce bias into inferences about θ unless the missing data are MCAR. Even if MCAR holds, case deletion may still be grossly inefficient.

We shall call a missing-data procedure a *general ignorable procedure* if it is based upon either an observed-data likelihood or an observed-data posterior. The EM algorithm, for example, will be seen to be a general ignorable procedure because it maximizes the observed-data likelihood. Omitting all incomplete cases from an analysis, however, is not a general ignorable procedure because it leads to a different likelihood or posterior, one based only on the complete cases. All of the methods developed in this book are general ignorable procedures.

A common feature of ad hoc missing-data treatments like case deletion is that they tend to discard information from certain units and/or variables in order to make the estimation problem more tractable. General ignorable procedures by nature, however, do not discard such information because the observed-data likelihood or posterior conditions fully on Y_{obs}. From standpoints of efficiency and bias, full conditioning on Y_{obs} is advantageous because it leads to inferences that are proper under any missing-data mechanism that is ignorable, whereas procedures that are not fully conditional may perform well in some but not all ignorable scenarios. This point can be illustrated with a simple hypothetical example.

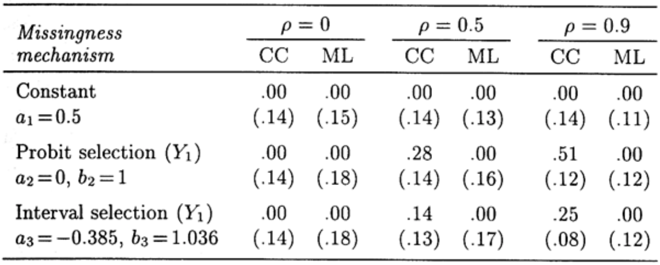
2.5.1 A simulated example

Consider again the bivariate data described in Example 2 of Section 2.3.2, in which Y_1 is always observed but Y_2 is sometimes missing. Under a bivariate normal data model with ignorable missingness, A-data we may easily find the value $\hat{\phi}$ that maximizes the observe likelihood function $L(\varphi \mid Y_{obs})$ in (2.9) and then apply the inverse transformation $\hat{\theta} = \varphi^{-1}(\hat{\varphi})$ to find the maximum-likelihood (ML) estimate for θ. Straightforward calculation shows that the ML estimate of μ_2 is $\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_{1,}$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are least-squares estimates from the regression of Y_2 on Y_1 based on units $1,2, ...,n_1$ and μ_1, is the average value of Y_1 among units $1,2,...,n$ (Anderson, 1957). Alternatively, one may compute $\hat{\mu}_2$ by first imputing regression-based predictions for the missing values of Y_2, i.e. letting $y_{i2} = \hat{\beta}_0 + \hat{\beta}_1 y_{i1}$ for $i = n_1 + 1,...,n$, and then computing the average of Y_2 among units $1,2,...,n$ in the imputed dataset.

Estimating μ_2 by $\hat{\mu}_2$ is a general ignorable procedure because $\hat{\mu}_2$ is obtained by maximizing the observed-data likelihood $L(\varphi \mid Y_{obs})$. Another plausible estimate is the complete-case (CC) estimate $\tilde{\mu}_2 = n_1^{-1}\Sigma_{i=1}^{n_1} y_{i2}$, the average of Y_2 among the completely observed cases. Estimating μ_2 by $\tilde{\mu}_2$ may certainly be regarded as an ignorable procedure because it is consistent with the belief that the missing data are MCAR, a special case of MAR. It is not, however, a general ignorable procedure because it does not condition fully on Y_{obs}; in particular, it discards the observed values of Y_1 for units n_1 + 1,..., n. Consequently, the ML estimate tends to be more efficient than the CC estimate, and it exhibits better performance over a variety of missingness mechanisms.

We can easily compare the performance of the ML and CC estimates by simulation. As shown in Figure 2.2, let us define

Table 2.1. *Simulated means (standard deviations) of the CC and ML estimates under three MAR mechanisms*



Missingness mechanism	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
	CC	ML	CC	ML	CC	ML
Constant $a_1 = 0.5$	.00 (.14)	.00 (.15)	.00 (.14)	.00 (.13)	.00 (.14)	.00 (.11)
Probit selection (Y_1) $a_2 = 0$, $b_2 = 1$	.00 (.14)	.00 (.18)	.28 (.14)	.00 (.16)	.51 (.12)	.00 (.12)
Interval selection (Y_1) $a_3 = -0.385$, $b_3 = 1.036$	.00 (.14)	.00 (.18)	.14 (.13)	.00 (.17)	.25 (.08)	.00 (.12)

$R = (r_1,...,r_n)$ to be a vector of response indicators where $r_i = 1$ if Y_2 is observed and $r_i = 0$ if Y_2 is missing for unit i, i = 1, 2,...,n. In particular, consider the class of ignorable mechanisms in which $\mathrm{P}(\mathrm{r}_i = 1|\mathrm{Y}) = \mathrm{g}(\mathrm{y}_{i1})$ independently for units i = 1, 2,..., n, where g is some function that maps the real line into the unit interval [0, 1]. Three possible choices for g are

constant	$g_1(y_{i1}) = a_1, 0 \le a_1 \le 1,$
probit selection	$g_2(y_{i1}) = \Phi(a_2 = b_2 y_{i1}).$
interval selection	$g_3(y_{i1}) = 1$ if $a_3 \le y_{i1} \le b_3$, else 0,

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The constant function g_1 is MCAR, whereas g_2 and g_3 are MAR but not MCAR. A simulation was conducted in which samples of size n = 100 were drawn from bivariate normal populations with $\mu_1 = \mu_2 = 0$, $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \rho = 0$, 0.5 and 0.9, respectively. Random patterns of missingness were imposed on each sample according to g_1 with $a_1 = 0.5$, g_2 with $a_2 = 0$, $b_2 = 1$ and g_3 with $a_3 = \Phi^{-1}(0.35) = -0.385$, $b_3 = -\Phi^{-1}(0.85) = 1.036$. These constants were chosen to yield an expected response rate n_1/n of 50% under each mechanism, a level higher than is found in most typical applications.

The means and standard deviations of the CC and ML estimates over 5000 repetitions are shown in Table 2.1. Under the non-MCAR mechanisms, the CC estimate is biased whenever $\rho \neq 0$. Except under the unrealistic condition that $n_1 / n \rightarrow 1$, this bias does not vanish as $n \rightarrow \infty$, causing $\tilde{\mu}_2$ to be inconsistent. The ML estimate, however, is unbiased and consistent under the three mechanisms used here, as well as under essentially all other ignorable mechanisms. From considerations of bias and consistency, the ML estimate has a clear advantage over the CC estimate.

In fairness, one should note that it is possible to construct a missingness mechanism for which the ML estimate would be less biased than the CC estimate. Such a mechanism would be neither MAR nor MCAR, but a peculiar nonignorable mechanism in which Y_1 and Y_2 would have correlations of opposite sign among the respondent and nonrespondent groups. Such mechanisms, although mathematically possible, are somewhat atypical and should not be expected to occur often in practice with real data. Further discussion of this point in an applied setting is given by Schafer (1992).

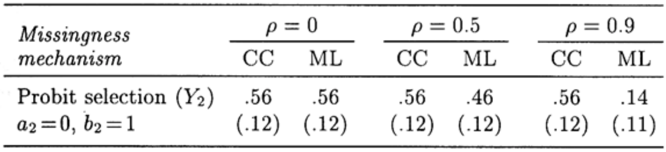
Under the more restrictive condition of MCAR, both the ML and CC estimates are unbiased, but ML still has an advantage over CC for $\rho = 0.5$ and $\rho = 0.9$ because its variance is lower. This reduction in variance occurs because Y_1 becomes an increasingly valuable predictor of the missing values of Y_2 as ρ increases. The only situations in Table 2.1 for which CC appears to dominate ML are when Y_1 and Y_2 are unrelated ($\rho = 0$), in which case $\hat{\mu}_2$ has more variability than $\tilde{\mu}_2$. Here CC enjoys an advantage because it correctly assumes that the correlation between Y_1 and Y_2 is zero, whereas ML uses an estimated regression line whose slope $\hat{\beta}_1$ randomly varies about zero. This advantage of CC over ML would be much less dramatic if the average missingness rate were lower. Moreover, the benefit of CC s lower variance in this special situation tends to be outweighed by the protection against bias afforded by ML when $\rho \neq 0$ and the mechanism is not MCAR.

2.5.2 Departures from ignorability

The above example illustrates the advantages of general ignorable procedures when missing data are MAR. Even when the missing data are not precisely MAR, however, general ignorable procedures still tend to be better than ad hoc procedures such as case deletion for the following reason: general ignorable procedures remove all of the nonresponse bias explainable by Y_{obs} whereas ad hoc procedures may not.

To demonstrate this point, let us modify the example of Section 2.5.1 to include a mechanism that is not MAR. Suppose that propensity to respond for Y_2 is no longer a function of Y_1 but is now directly related to Y_2. We will assume that $P(\mathrm{r}_i = 1|Y) = g(y_{i2})$ independently for $i = 1, 2, \ldots, n$. A simulation was conducted using probit selection, $g(y_{i2}) = \Phi(a_2 + b_2 y_{i2})$ with $a_2 = 0$, $b_2 = 1$ and all

Table 2.2. *Simulated means (standard deviations) of d2 and A2 under a non-MAR mechanism*



Missingness mechanism	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
	CC	ML	CC	ML	CC	ML
Probit selection (Y_2) $a_2 = 0$, $b_2 = 1$	.56 (.12)	.56 (.12)	.56 (.12)	.46 (.12)	.56 (.12)	.14 (.11)

other parameters as before. Results shown in Table 2.2 show that the CC estimate $\tilde{\mu}_2$ and the general ignorable ML estimate $\hat{\mu}_2$ are equally biased when $\rho = 0$ but that ML becomes substantially less biased for larger values of ρ. By making use of Y_1 to predict missing values of Y_2, the ML procedure removes the nonresponse bias in the observed values of Y_2 attributable to Y_1. In this example nonresponse is related to Y_2 rather than Y_1, but as ρ increases Y_1 becomes an increasingly useful proxy for Y_2.

Limited practical experience with real data also suggests that general ignorable procedures may tend to perform well even when the ignorability assumption is suspect, especially in multivariate settings. In surveys containing questions about income, for example, nonresponse rates on income-related questions tend to be high, and both experience and common sense suggest that the probability of response is likely to be related to level of income. In a study of missing-data methods for an income question in the Current Population Survey, Greenlees *et al.* (1982) established definite relationships between response and income itself. Upon further investigation, David *et al.* (1986) found little evidence of bias in ignorable procedures that imputed missing values of income on, the basis of other demographic and questionnaire items that were observed. This evidence came from knowledge of the missing values obtained from an external source, the actual wages and salary reported to the Internal Revenue Service on the individuals tax returns. David *et al.* (1986) also concluded that further improvements in the missing-data procedures would probably come from better modeling of the multivariate structure of the data, not from nonignorable modeling.

The crucial assumption made by ignorable methods is not that the propensity to respond is completely unrelated to the missing data, but that this relationship can be explained by data that are observed. Whether ignorability is plausible in a particular setting is therefore closely related to the richness of the observed data Y_{obs} and the complexity of the data model $P(Y|\theta)$. If Y_{obs} contains a lot of information relevant for predicting Y_{mis}, and if the data model is sufficiently complex to make use of this information, then we should expect the residual dependence of R upon Y_{mis} after conditioning on Y_{obs} to be relatively minor. Thus in multivariate datasets where both the observed data and the complete-data model are rich, general ignorable procedures may tend to perform well in practice. Even if they do not perform well, they still may provide an important and useful baseline for assessing and comparing any available alternatives.

2.5.3 Notes on nonignorable alternatives

Various approaches to incomplete data that do not assume ignorability have also appeared in the literature. A detailed review of nonignorable methods is beyond the scope of this text, but we note that these nonignorable methods tend to have a common approach. They generally involve joint probability modeling of both the data $P(Y|\theta)$ and the missingness mechanism $P(R|Y, \xi)$ and joint estimation of θ and ξ from Y_{obs} and R. These joint models for Y and R typically involve more parameters than can be estimated from Y_{obs} and R alone. In order to make them identifiable, one must either (a) impose a priori restrictions on the joint parameter space for θ and ξ or (b) impose an informative Bayesian prior distribution on (θ,ξ).

For continuous data, one group of nonignorable methods is based on models known in the econometric literature as stochastic censoring or selection models (Heckman, 1976; Amemiya, 1984). For categorical data, nonignorable contingency-table approaches are described by Fay (1986);

Baker and Laird (1988); Rubin, Schafer and Schenker (1988); Winship and Mare (1989); and Park and Brown (1994). A review of some nonignorable methods is given by Little and Rubin (1987, chap. 11). Glynn, Laird and Rubin (1993) describe nonignorable modeling based on followup data. Approaches to nonignorable modeling for longitudinal data are discussed by Conaway (1992, 1994); Diggle and Kenward (1994); and Baker (1994). Little (1993) discusses in general terms a class of nonignorable models called pattern-mixture models, in which the joint distribution of Y and R is specified in terms of the marginal distribution of R and the conditional distribution of Y given R.

2.6 The role of the complete-data model

Having discussed at length the assumption of ignorability, we now return to the role of the model for the complete data presented in Section 2.1. A good model should be plausible and have sufficient flexibility to preserve the subtle features of the data at hand, but in realistic settings one must also, consider issues of convenience and mathematical tractability. The classes of models considered in this book, the multivariate normal for continuous data, loglinear models for cross-classified categorical data and the class of models for mixed continuous and categorical data, are general-purpose multivariate models that are both mathematically tractable and appropriate in many but not all situations. With categorical and mixed data, the analyst has considerable freedom to tailor a model to the particular dataset; two-way, three-way or higher associations between variables may be included if they are thought to be important. The multivariate normal model for continuous data is less flexible, however, because it fits only pairwise associations. Sometimes the normality assumption may be made more plausible by applying suitable transformations to one or more variables (Box and Cox, 1964; Emerson, 1991).

2.6.1 Departures from the data model

When making inferences about a parameter θ, the assumed parametric form of the model and the iid assumption often play a crucial role. If these assumptions are seriously violated, then even under ignorability the likelihood (2.5) may be a poor summary of the data s evidence about θ. Indeed, if the data model does not hold, the interpretation of θ itself may be ambiguous. Selection of a data model should proceed with care, and diagnostics for assessing goodness of fit should be used whenever possible. Of course, in all analyses of real (not simulated) data a probability model is only an approximation to reality, and some departures from modeling assumptions are inevitable. In practice one must judge whether these departures are of a magnitude and nature that seriously impairs the quality of the inference about θ, or whether they are of only minor importance and may be safely ignored.

Complex sample surveys

The assumption (2.1) that the rows of the data matrix are independent and identically distributed (iid) can be problematic, especially when the rows do not correspond to observational units that are *exchangeable* (i.e. like a simple random sample). This assumption is commonly violated in large-scale surveys, which typically employ complex probability sampling methods. Sample surveys often have special features (unequal probabilities of selection, stratification, clustering, etc.) that cause the units to be non-exchangeable.

In some cases, it may be possible to apply the models of this book to survey data in a way that preserves the integrity of the complex design. Stratification and the general issue of design variables are discussed in Section 2.6.2 below. Data with clusters, naturally occurring groups of observations that are potentially intercorrelated, should not in general be handled by iid models. If the clusters are few and large, it may be feasible to fit separate models for each cluster that treat the units within clusters as exchangeable. Another possibility is to add a cluster identifier to the model as an additional categorical variable, and posit some simple associations between this variable and the other variables in the dataset; an example of this approach is described in Section 9.5.3.

When the number of clusters is large, and there are relatively few observations per cluster, the data are more appropriately described by hierarchical probability models with an explicit multilevel error structure. These models, sometimes called random-effects or mixed models, have been extensively applied in the univariate setting with continuous responses (Laird and Ware, 1982; Searle *et al.*, 1992). Recent advances in computational methods have allowed extensions to discrete (Zeger and Karim, 1991) and multivariate continuous (Everson and Morris, 1996) response models. Application to problems of missing data in surveys, however, will typically require hierarchical models for multivariate categorical responses. A proper treatment of these models is well beyond the scope of this book. For a detailed example of such a model in the context of the U.S. Decennial Census, see Schafer (1995). More discussion of missing-data problems in complex surveys will appear in Sections 4.3 and 4.5.

The role of imputation

It may be possible to mitigate some of the effects of model failure through multiple imputation (Section 4.3). Inference by multiple imputation proceeds in two stages. First, m simulated versions of the missing data are created under a data model. Second, the m versions of the complete data are then analyzed by complete-data statistical techniques, and the results are combined to produce one overall inference. Sometimes the complete-data statistical analyses of the second stage will involve different models than the one used to produce the multiple imputations in the first stage. When analyzing data from a sample survey, for example, one may impute the missing data on the basis of an elaborate multivariate model, but then proceed to analyze the data using classical nonparametric survey methods in which inferences are based entirely on the randomization used to draw the sample (e.g. Cochran, 1977). Even if the model used for imputation is somewhat restrictive or unrealistic, it will effectively be applied not to the entire dataset but only to its missing part. Multiple imputation thus has a natural advantage over some other methods of inference in that it may tend to be more robust to departures from the complete-data model, especially when the amounts of missing information are not large. Hence, even though the classes of models examined in this book may not realistically describe many of the multivariate datasets one encounters in the real world, we suspect that they will still prove useful in a wide variety of data analyses if applied within the framework of multiple imputation. The role of modeling assumptions in multiple imputation will be revisited in Section 4.5.4.

2.6.2 Inference treating certain variables as fixed

Sometimes the only relevant analysis of a multivariate dataset involves modeling the conditional distribution of certain variables given others. In regression analysis, for example, the goal is to model one or more response variables given one or more predictors. When analyzing data from surveys, it is common practice to estimate means, proportions, etc. not for the population as a whole but within subdomains (strata or poststrata) that are considerably smaller. The individual subdomain estimates are then combined, using population proportions derived from an external source (e.g. a census), to yield estimates for larger domains. The relative sizes of the subdomains in the population are assumed to be known and are not estimated from the sample. Similarly, with data from planned experiments, the relevant analysis is usually a comparison of mean responses across two or more treatment groups; the manner in which experimental units are allocated to treatments is determined by the experimenter and does not need to be modeled.

In discussing situations like these, we will refer to variables in a generic sense as either *response* variables or *design* variables, with the latter being those that the statistical analysis ultimately regards as fixed. Predictors in regression analyses, variables defining strata or poststrata in sample surveys and indicators of treatment groups in planned experiments are all examples of design variables. When a dataset contains one or more design variables, the iid assumption (2.1) is typically violated, as design variables are often under direct control of the investigator and are thus not random in the same sense as the response variables are random. It is usually not desirable to impose any probability model at all on the design variables, but to model only the conditional distribution of the response variables given the design variables.

Datasets with design variables can be accommodated in our framework with just a few additional assumptions. Suppose that each row y_i of the complete-data matrix can be partitioned into two parts, a vector u_i of design variables and a vector v_i of response variables. Furthermore, suppose that the density of y_i can be factored as

$$f(y_i \mid \theta) = f_u(u_i \mid \alpha) f_{v|u}(v_i \mid u_i, \beta), \tag{2.19}$$

where α and β are distinct parameters. Finally, suppose that the design variables are completely observed for all units. Under these assumptions, we can write the complete-data matrix as

$$Y = (Y_{obs}, Y_{mis}) = (U, V_{obs}, V_{mis}),$$

where U denotes the design variables, V_{obs} the observed response variables, and V_{mis} the missing response variables, so that Y_{obs} = (U, V_{obs}) and Y_{mis} =V_{mis}. The probability distribution of the observed quantities may then be written as

$$\int P(R \mid U, V_{obs}, V_{mis}, \xi) P(U \mid \alpha) P(V_{obs}, V_{mis} \mid U, \beta) dV_{mis}, \tag{2.20}$$

where ξ is a set of parameters governing the response mechanism. Under ignorability, this response mechanism does not depend on V_{mis},

$$P(R \mid U, V_{obs}, V_{mis}, \xi) = P(R \mid U, V_{obs}, \xi).$$

Thus the first two factors in the integrand of (2.20) do not involve V_{mis}, so (2.20) becomes

$$P(R, Y_{obs} \mid \theta, \xi) = P(R \mid U, V_{obs}, \xi) P(U \mid \alpha) P(V_{obs} \mid U, \beta). \quad (2.21)$$

The factorization in (2.21) implies that inferences about β, the parameters of the conditional distribution of V given U, may be based on the conditional observed-data likelihood function

$$L(\beta \mid U, V_{mis}) \propto P(V_{obs} \mid U, \beta),$$

or on the observed-data posterior

$$P(\beta \mid U, V_{mis}) \propto L(\beta \mid U, V_{mis}) \pi_{\beta}(\beta),$$

where π_β is a prior distribution applied to β independently of any prior on α or ξ. Notice that (2.22) and (2.23) are the same for all a and even for all $P(U|\alpha)$. In other words, we will obtain a correct inference about β even if the marginal model for the design variables is misspecified.

To summarize, when design variables are present we can often apply a joint probability model such as (2.1) to the complete data Y, even a model in which the distribution of the design variables is misspecified. We will obtain correct inferences about the parameters of interest provided that (a) the design variables U are observed for all units in the sample, and (b) the joint density of (U, V) factors into densities for U and for $V|U$, with the parameters of the two densities being distinct. Assumption (b) will be a characteristic of some, but not all, of the multivariate models used in this book. Assumption (a) often does hold in practice; it would be somewhat unusual, for example, for stratification variables in a sample survey or treatment indicators in a planned experiment to be missing. In regression analysis with incomplete predictors, the factorization in (2.21) will not precisely hold, but it may still be approximately true provided that the amount of information missing on the predictors is not large.

Example: a comparison of two sample means

The experimental data in Table 2.3 reported by Snedecor and Cochran (1989, Table 6.9.1) show the weight gains of two groups of female rats, one fed a low-protein diet and the other fed a high-protein diet. The low-protein group has 7 rats and the high-protein group has 12. Snedecor and Cochran perform a classical analysis assuming that the observations are independent and normally distributed and the within-group variances are equal. A pooled estimate of the common variance is

$$\frac{6(425)+11(457)}{17} = 445.7$$

Table 2.3. *Weight gains in grams of two groups of female rats (28-84 days old) under two diets*

Low protein	70	118	101	85	107	132
	94					
	mean = 101			variance = 425		
High protein	134	146	104	119	124	161
	107	83	113	129	97	123
	mean = 120			variance = 457		

Source: Snedecor and Cochran (1989, Table 6.9.1)

on 17 degrees of freedom, and a 95% confidence interval for the difference in mean weight gain between the two groups is

$$(120-101) \pm t_{17,\ .975}\sqrt{445.7\left(\frac{1}{7}+\frac{1}{12}\right)}, \qquad (2.24)$$

where $t_{v,p}$ denotes the pth quantile of the t distribution with v degrees of freedom. The interval (2.24) extends from -2.2 to 40.2, barely covering zero, so the difference in means is almost significant at the 0.05 level.

Let us now regard Table 2.3 as incomplete data from a balanced experiment; that is, we now suppose that the low-protein group had 12 potential observations, 5 of which are missing. This supposition is for illustrative purposes only, because the analysis of variance for an experiment with a single factor is no more difficult when the data are unbalanced than when they are balanced. The complete data could then be regarded as a 24 × 2 matrix in which the first variable Y_1 is a treatment indicator (0=low protein, 1=high protein) and the second variable Y_2 is weight gain.

Suppose that we modeled the joint distribution of Y_1 and Y_2 as bivariate normal. The implied marginal normal distribution of the design variable Y_1 would be clearly erroneous. But note the conditional distribution of Y_2 given Y_1 implied by this model, a normal linear regression with constant variance, is precisely the same model that underlies the classical analysis and the confidence interval (2.24). Because Y_1 is coded as 0 or 1, the slope of this regression is identical to the difference in mean weight gain between the two groups. Likelihood-based or Bayesian inferences about the regression slope would yield essentially the same result as the classical interval (2.24), perhaps with minor differences depending on how the observed-data likelihood is summarized or on what prior distribution is chosen.

One possible advantage of using the bivariate normal model here is that a general ignorable procedure devised for incomplete multivariate normal data could be applied to this dataset and the resulting inference about the regression slope would still be valid, provided, of course, that the conditional normal model for Y_2 given Y_1 was correct. When design variables are present it will often be convenient to apply model-fitting and simulation algorithms devised for iid probability models like the multivariate normal, even though parts of the model pertaining to the design variables may be incorrect, because developing a more specialized algorithm then becomes unnecessary.

This example also raises an unrelated but important issue regarding unbalanced experimental data. Classical methods of analysis such as the *t*-interval in (2.24), and other methods for unbalanced data arising from more complicated designs (e.g. Dodge, 1985), almost invariably contain an implicit assumption that the mechanism causing the imbalance is ignorable. If the data are unbalanced not by design but by accident, e.g. if responses for one or more units could not be recorded because of mishaps or other unforeseen occurrences, then these methods should not be applied without first considering the plausibility of MAR.

CHAPTER 3

EM and data augmentation

3.1 Introduction

Assuming that the complete-data model and ignorability assumptions are correct, all relevant statistical information about the parameters is contained in the observed-data likelihood $L(\theta|Y_{obs})$ or observed-data posterior $P(\theta|Y_{obs})$. Except in special cases, however, these tend to be complicated functions of θ, and extracting meaningful summaries such as parameter estimates and standard errors requires special computational tools. EM and data augmentation provide those tools. The key ideas behind EM and data augmentation are the same: to solve a difficult incomplete-data problem by repeatedly solving tractable complete-data problems. As a result, the two methods share many features in common, and their implementation in specific examples is often remarkably similar. In this chapter, EM and data augmentation are introduced together to highlight the similarities between them.

3.2 The EM algorithm

3.2.1 Definition

EM capitalizes on the interdependence between missing data Y_{mis} and parameters θ. The fact that Y_{mis} contains information relevant to estimating θ, and θ in turn helps us to find likely values of Y_{mis} suggests the following scheme for estimating θ in the presence of Y_{obs} alone: Fill in the missing data Y_{mis}

based on an initial estimate of θ, re-estimate θ based on Y_{obs} and the filled-in Y_{mis}, and iterate until the estimates converge. This idea is so intuitively appealing that specific applications of it have appeared in the statistical literature as far back as 1926 (Little and Rubin, 1987; Meng and Pedlow, 1992). Dempster, Laird and Rubin (1977) formalized the meaning of filling in the missing data at each step and presented the algorithm in its full generality, naming it Expectation-Maximization or EM.

In any incomplete-data problem, the distribution of the complete data Y can be factored as

$$P(Y \mid \theta) = P\left(Y_{obs} \mid \theta\right)P\left(Y_{mis} \mid Y_{obs}, \theta\right). \tag{3.1}$$

Viewing each term in (3.1) as a function of θ, it follows that

$$l(\theta \mid Y) = l\left(\theta \mid Y_{obs}\right) + \log P\left(Y_{mis} \mid Y_{obs}, \theta\right) + c \tag{3.2}$$

where $l(\theta/Y)$ = log$P(Y|\theta)$ denotes the complete-data loglikelihood, $l(\theta/Y_{obs})$ = log $P(Y|\theta)$ the observed-data loglikelihood, and c an arbitrary constant. The term $P(Y_{mis}|Y_{obs}, \theta)$, which we shall call the *predictive distribution of the missing data given* θ, plays a central role in EM because it captures the interdependence between Y_{mis}, and θ. When viewed as a probability distribution it summarizes knowledge about Y_{mis}, for any assumed value of θ, and when viewed as a function of θ it conveys the evidence about θ contained in Y_{mis} beyond that already provided by Y_{obs}.

Because Y_{mis} is unknown we cannot calculate the second term on the right-hand side of (3.2), so instead we take the average of (3.2) over the predictive distribution $P(Y_{mis}|Y_{obs},\theta^{(t)}$, where $\theta^{(t)}$ is a preliminary estimate of the unknown parameter. This averaging yields

$$Q\left(\theta \mid \theta^{(t)}\right) = l(\theta \mid Y_{obs}) + H\left(\theta \mid \theta^{(t)}\right) + c, \tag{3.3}$$

where

$$Q\left(\theta \mid \theta^{(t)}\right) = \int l(\theta \mid Y) P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right) dY_{mis}$$

and

$$H\left(\theta \mid \theta^{(t)}\right) = \int \log P(Y_{mis} \mid Y_{obs}, \theta) P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right) dY_{mis.}$$

A central result of Dempster, Laird, and Rubin (1977) is that if we let $\theta^{(t+1)}$ be the value of θ that maximizes $Q(\theta/\theta^{(t)})$, then $\theta^{(t+1)}$ is a better estimate than $\theta^{(t)}$ in the sense that its observed-data loglikelihood is at least as high as that of $\theta^{(t)}$,

$$l\left(\theta^{(t+1)} \mid Y_{obs}\right) \geq l\left(\theta^{(t)} \mid Y_{obs}\right). \qquad (3.4)$$

This can be seen by writing

$$l\left(\theta^{(t+1)} \mid Y_{obs}\right) \geq l\left(\theta^{(t)} \mid Y_{obs}\right) = Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)$$

$$+ H\left(\theta^{(t)} \mid \theta^{(t)}\right) - H\left(\theta^{(t+1)} \mid \theta^{(t)}\right).$$

The quantity $Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)$ is non-negative because $\theta^{(t+1)}$ has been chosen to satisfy

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q\left(\theta \mid \theta^{(t)}\right) \text{ for all } \theta. \qquad (3.5)$$

The remainder $H\left(\theta^{(t)} \mid \theta^{(t)}\right) - H\left(\theta^{(t+1)} \mid \theta^{(t)}\right)$, which can be written

$$\int \log\left[\frac{P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right)}{P\left(Y_{mis} \mid Y_{obs}, \theta^{(t+1)}\right)}\right] P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right) dY_{mis,}$$

is easily shown to be non-negative by Jensen s inequality and the convexity of the function x log x.

It is convenient to think of one iteration of EM, defined by (3.5), as consisting of two distinct steps:

1. the Expectation or E-step, in which the function $Q(\theta|\theta^{(t)})$ is calculated by averaging the complete-data loglikelihood $l(\theta/Y)$ over $P(Y_{mis}|Y_{obs}, \theta^{(t)})$; and

2. the Maximization or M-step, in which $\theta^{(t+1)}$ is found by maximizing $Q(\theta|\theta^{(t)})$.

Alternately performing the E- and M-steps beginning with a starting value $\theta^{(0)}$ defines a sequence of iterates $\{\theta^{(t)}:t = 0,1,2,...\}$. Dempster, Laird, and Rubin (1977) and Wu (1983) provide conditions under which this sequence converges reliably to a stationary point of the observed-data loglikelihood. In well-behaved problems this stationary point is a global maximum and EM yields the unique maximum-likelihood estimate (MLE) of θ, the maximizer of $l(\theta|Y_{obs})$. Not all problems are well-behaved, however, and sometimes EM does not converge to a unique global maximum; these situations are taken up in Section 3.3.1 below.

EM for regular exponential families

The E-step of EM clarifies the intuitive idea of filling in the missing data under an assumed value of θ. In some problems (e.g. incomplete data that are purely categorical), we shall see that the E-step actually does correspond to filling in the missing data in the sense that it replaces Y_{mis} with its average or expected value $E(Y_{mis}|Y_{obs}, \theta)$ under the assumption $\theta=\theta^{(t)}$. In other problems, however, it does not. In particular, when the complete-data probability model falls in a *regular exponential family*, the complete-data loglikelihood based on n iid (possibly multivariate) observations $Y = (y_1, y_2,..., y_n)$ may be written

$$l(\theta \mid Y) = \eta(\theta)^T T(Y) + ng(\theta) + c, \tag{3.6}$$

where

$$\eta(\theta) = \left(\eta_1(\theta), \eta_2(\theta), ..., \eta_s(\theta)\right)^T$$

is the canonical form of the parameter θ,

$$T(Y) = \left(T_1(Y), T_2(Y), ..., T_s(Y)\right)^T$$

is an s-dimensional vector of complete-data sufficient statistics and c is a constant term that does not involve θ. Moreover, each of the sufficient statistics has an additive form,

$$T_j(Y) = \sum_{i=1}^{n} h_j(y_i)$$

for some function h_j. Because $l(\theta|Y)$ is a linear function of the sufficient statistics, the E-step replaces $T_j(Y)$ by $E(T_j(Y)|Y_{obs}, \theta^{(t)})$ for j =1,2, . . . ,s; in other words, the E-step fills in not the missing elements of Y per se, but rather the missing portions of the complete-data sufficient statistics. In our regular exponential-family models, the expectations $E(T_j(Y)|Y_{obs},\theta^{(t)})$ will be available in closed form and thus the E-step will be computationally straightforward.

In many cases the M-step will also be straightforward; $Q(\theta|\theta^{(t)})$ will have the same functional form as a complete-data loglikelihood, so finding $\theta^{(t+1)}$ will be computationally no different from finding the MLE in the complete-data case. For regular exponential families, the complete-data MLE can be found as the solution to the moment equations

$$E\left(T(Y) \mid \theta\right) = t, \tag{3.7}$$

where t is the realized value of the vector $T(Y)$ and the expectation is taken with respect to $P(Y|\theta)$ (e.g. Cox and

Hinkley, 1974). If these equations can be solved for an arbitrary t, then they can just as easily be solved when t is replaced by the output of an E-step. In many of the models appearing in this book, the moment equations can be solved for θ algebraically, and thus the M-step will be available in closed form. When an algebraic solution is not available, one can still maximize the loglikelihood numerically for any given t using standard complete-data iterative techniques such as Newton-Raphson. In the latter situation, implementation of EM would require undesirable nested iterations because each M-step would itself be iterative. When this arises, however, we are often able to streamline the computation by applying a generalization of EM known as ECM, to be discussed in Section 3.2.5.

3.2.2 Examples

Example 1: Incomplete univariate normal data. Suppose that $Y = (y_1, y_2,..., y_n)$ represents n iid observations from a univariate normal distribution with mean μ and variance ψ, so that $\theta = (\mu, \psi)$ is the unknown parameter. The reader may easily verify that the complete-data loglikelihood $l(\theta|Y)$ can be written in exponential-family form (3.6) with sufficient statistics

$$T(Y) = \left(T_1, T_2\right)^T = \left(\Sigma_{i=1}^n y_i, \Sigma_{i=1}^n y_i^2\right)^T.$$

Letting t_1 and t_2 denote the realized values of T_1 and T_2 respectively, the moment equations

$$E\left(T_1\right) = n\mu = t_1,$$
$$E\left(T_2\right) = n\psi + n\mu^2 = t_2$$

lead immediately to the well-known MLEs $\hat{\mu} = \bar{y} = n^{-1}\Sigma_{i=1}^n y_i$ and $\hat{\psi} = n^{-1}\Sigma_{i=1}^n y_i^2 - (\bar{y})^2$.

Now suppose that only the first n_1 components of the data vector Y are observed, and the remaining $n_0 = n - n_1$ components are missing at random (which, in this simple example, is equivalent to MCAR). It follows from Example 1, Section 2.3.2 that the observed-data likelihood $L(\theta|Y_{obs})$ is just a complete data likelihood based only on $Y_{obs} = (y_1, y_2, \ldots, y_{n_1})$ and the observed-data MLEs for μ and ψ are thus $\bar{y}_{obs} = n_1^{-1}\Sigma_{i=1}^{n_1} y_i$ and $n_1^{-1}\Sigma_{i=1}^{n_1} y_i^2 - (\bar{y}_{obs})^2$, respectively. In this trivial example the observed-data MLEs exist in closed form, but one could also compute them using the EM algorithm. Because of the iid structure the predictive distribution of the missing data given θ does not depend on the observed data, and

$$P(Y_{mis} \mid Y_{obs}, \theta) = P(Y_{mis} \mid \theta) = \prod_{i=n_1+1}^{n} P(y_i \mid \theta). \tag{3.8}$$

The E-step replaces T_1 and T_2 by their expected values under $P(Y_{mis}|Y_{obs}, \theta)$,

$$\begin{aligned} E(T_1 \mid Y_{obs}, \theta) &= E\left[\Sigma_{i=1}^{n_1} y_i + \Sigma_{i=n_1+1}^{n} y_i\right] \\ &= \Sigma_{i=1}^{n_1} y_i + n_0 \mu, \\ E(T_2 \mid Y_{obs}, \theta) &= E\left[\Sigma_{i=1}^{n_1} y_i^2 + \Sigma_{i=n_1+1}^{n} y_i^2\right] \\ &= \Sigma_{i=1}^{n_1} y_i^2 + n_0\left(\psi + \mu^2\right). \end{aligned}$$

Inserting these expected sufficient statistics into the expressions for the complete-data MLEs yields a single iteration of EM,

$$\mu^{(t+1)} = n^{-1}\left[\sum_{i=1}^{n_1} y_i + n_0 \mu^{(t)}\right], \tag{3.9}$$

$$\psi^{(t+1)} = n^{-1}\left[\sum_{i=1}^{n_1} y_i^2 + n_0\psi^{(t)} + n_0\left(\mu^{(t)}\right)^2\right] \\ -n^{-2}\left[\sum_{i=1}^{n_1} y_i + n_0\mu^{(t)}\right]^2. \tag{3.10}$$

In this simple example, the fixed-point equations $\mu^{(t+1)} = \mu^{(t)}$ and $\psi^{(t+1)} = \psi^{(t)}$ can be solved explicitly to show that the iterations converge to the correct observed-data MLEs (Little and Rubin, 1987).

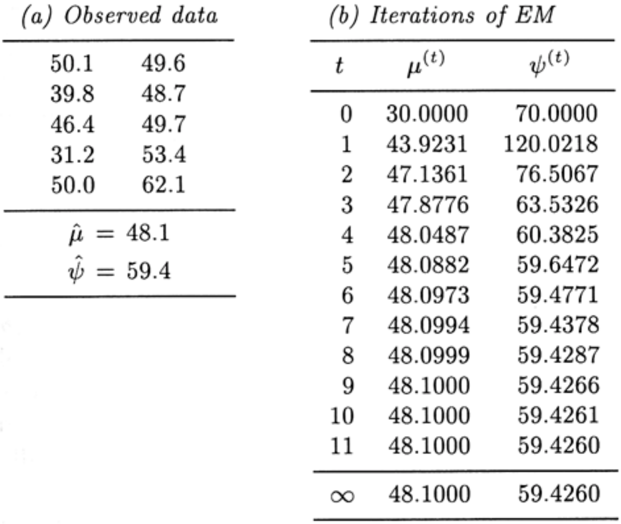
The behavior of this EM algorithm in a small numerical example is displayed in Table 3.1. A sample of $n_1 = 10$ observations is shown with mean $\hat{\mu} = 48.1$ and sample (maximum-likelihood) variance $\hat{\psi} = 59.4$. Arbitrarily choosing the number of missing observations to be $n_0 = 3$ and the starting values to be $\mu^{(0)}$, the algorithm converges to within four decimal places of the MLEs by the 11th iteration. It is apparent from (3.9)-(3.10) that convergence can be accelerated by taking no to be small, and with $n_0 = 0$ the MLE is achieved after just one iteration regardless of the starting values. We shall see in Section 3.3.3 that the rate of convergence in this example is determined by $n_0|(n_0+n_1)$, the proportion of observations that are missing. More generally, the convergence rate of EM is governed by the fractions of information about components of θ missing due to nonresponse.

Example 2: Two binary variables with missing data on both. Suppose that Y_1 and Y_2 are two potentially related dichotomous variables, each taking values 1 or 2. If the n units in a sample are iid, the complete data may, without loss of information, be reduced to an array of counts $x=(x_{11}, x_{12}, x_{21}, x_{22})$ having a multinomial distribution, where x_{ij} is the number of sample units having $Y_1 = i$ and $Y_2 = j$. Let $\theta = (\theta_{11}, \theta_{12},, \theta_{21}, \theta_{22})$, where θ_{ij} is the probability that a unit has Y_1 = i and $Y_2 = j$. We will use the notation $x \sim M(n, \theta)$ to indicate that x has a multinomial distribution with index n and parameter θ. Because the complete-data loglikelihood is linear in the cell counts x_{ij},

$$l(\theta \mid x) = x_{11} \log \theta_{11} + x_{12} \log \theta_{12} + x_{21} \log \theta_{21} + x_{22} \log \theta_{22},$$

the counts are the sufficient statistics and the MLEs are found by equating each x_{ij} with its expectation $n\theta_{ij}$; hence the complete data MLEs are simply the sample proportions $\hat{\theta}_{ij} = x_{ij} / n$ for i, j=1,2.

Table 3.1. *Example of EM for incomplete univariate normal data with n_1 = 10 values observed and n_0 = 3 values*



(a) Observed data

50.1	49.6
39.8	48.7
46.4	49.7
31.2	53.4
50.0	62.1
$\hat{\mu}$ = 48.1	
$\hat{\psi}$ = 59.4	

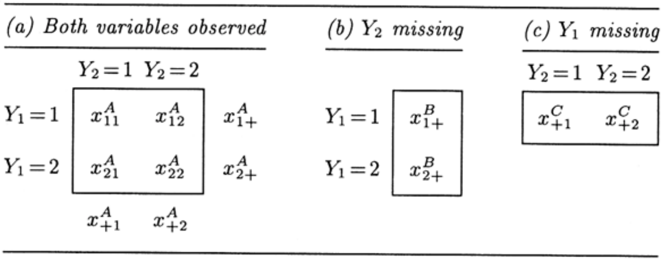
(b) Iterations of EM

t	$\mu^{(t)}$	$\psi^{(t)}$
0	30.0000	70.0000
1	43.9231	120.0218
2	47.1361	76.5067
3	47.8776	63.5326
4	48.0487	60.3825
5	48.0882	59.6472
6	48.0973	59.4771
7	48.0994	59.4378
8	48.0999	59.4287
9	48.1000	59.4266
10	48.1000	59.4261
11	48.1000	59.4260
∞	48.1000	59.4260

If missing values occur on both Y_1 and Y_2, we can partition the sample into three parts denoted by A, B and C, respectively, where A includes units having both variables observed, B includes those having only Y_1 observed and C includes those having only Y_2 observed. (Any units that have neither Y_1 nor Y_2 observed contribute nothing to the observed-data likelihood and may be excluded from the analysis under ignorability.) Each complete-data count x_{ij} can then be expressed as the sum of contributions from each of the three sample parts, $x_{ij} = x_{ij}^A + x_{ij}^B + x_{ij}^C$. Although x_{ij}^A is observed, x_{ij}^B and x_{ij}^C are not; for sample parts B and C, we observe only the marginal totals $x_{i+}^B = x_{i1}^B + x_{i2}^B$ and $x_{+j}^C = x_{1j}^C + x_{2j}^C$, respectively. The observed data $Y_{obs} = \left\{x_{ij}^A, x_{i+}^B, x_{+j}^C : i, j = 1, 2\right\}$ can be displayed as in Table 3.2, with a 2 × 2 table cross-classifying the units in A by Y_1 and Y_2, a 2 × 1 table classifying the units in B by Y_1 alone, and a 1 × 2 table classifying the units in C by Y_2 alone.

A convenient feature of the multinomial distribution is that if we regard the sum of any set of components of x as fixed, the conditional distribution of those components becomes another multinomial and is independent of the remaining components (e.g. Agresti, 1990). For example, the conditional distribution of (x_{11}, x_{12}) given

Table 3.2. *Classification of sample units by two incompletely observed binary variables*


(a) Both variables observed

	$Y_2=1$	$Y_2=2$	
$Y_1=1$	x_{11}^A	x_{12}^A	x_{1+}^A
$Y_1=2$	x_{21}^A	x_{22}^A	x_{2+}^A
	x_{+1}^A	x_{+2}^A	

(b) Y_2 missing

$Y_1=1$	x_{1+}^B
$Y_1=2$	x_{2+}^B

(c) Y_1 missing

$Y_2=1$	$Y_2=2$
x_{+1}^C	x_{+2}^C

$x_{1+}=x_{11}+x_{12}$ is multinomial with parameter $(\theta_{11}/\theta_{1+},\theta_{12}/\theta_{1+})$ where $\theta_1+ = \theta_{11}+\theta_{12}$; furthermore, (x_{11}, x_{12}) is conditionally independent of (x_{21}, x_{22}). Applying this property within parts B and C of the sample, the predictive distribution of the missing data given θ and the observed data becomes a set of independent multinomials or a *product multinomial*,

$$\left(x_{i1}^B, x_{i2}^B\right) Y_{obs}, \theta \sim M\left(x_{i+}^B, \left(\theta_{i1}/\theta_{i+}, \theta_{i2}/\theta_{i+}\right)\right), i=1,2.$$

$$\left(x_{i1}^C, x_{i2}^C\right) Y_{obs}, \theta \sim M\left(x_{+j}^C, \left(\theta_{1j}/\theta_{+j}, \theta_{2j}/\theta_{+j}\right)\right), j=1,2.$$

The E-step of EM replaces the unknown counts x_{ij}^B and x_{ij}^C in x_{ij} by their conditional expectations under an assumed value for θ,

$$E\left(x_{ij} \mid Y_{obs}, \theta\right) = E\left(x_{ij}^A + x_{ij}^B + x_{ij}^C \mid Y_{obs}, \theta\right)$$
$$= x_{ij}^A + x_{i+}^B \theta_{ij}/\theta_{i+} + x_{+j}^C \theta_{ij}/\theta_{+j}.$$

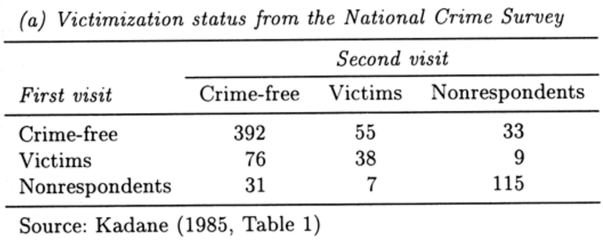
The M-step then estimates θ_{ij} by $E(x_{ij}|Y_{obs}, \theta)$. Combining the two steps yields a single iteration of EM,

$$\theta_{ij}^{(t+1)} = n^{-1}\left[x_{ij}^A + x_{i+}^B\left(\frac{\theta_{ij}^{(t)}}{\theta_{i+}^{(t)}}\right) + x_{+j}^C\left(\frac{\theta_{ij}^{(t)}}{\theta_{+j}^{(t)}}\right)\right],$$

i, j=1,2, an expression first given by Chen and Fienberg (1974).

The data in Table 3.3, previously analyzed by Kadane (1985), were obtained through the National Crime Survey conducted by the U.S. Bureau of the Census. Housing unit occupants were interviewed to determine whether they had been victimized by crime in the preceding six-month period. Six months later the units were visited again to determine whether the occupants had been victimized in the intervening months. Discarding the 115 households that

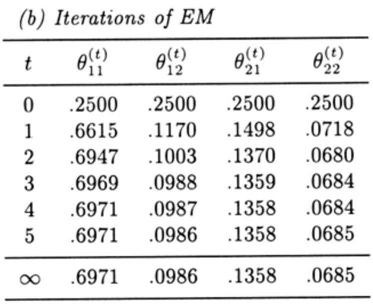
Table 3.3. *EM algorithm applied to victimization status of households on two occasions*



(a) Victimization status from the National Crime Survey

	Second visit		
First visit	Crime-free	Victims	Nonrespondents
Crime-free	392	55	33
Victims	76	38	9
Nonrespondents	31	7	115

Source: Kadane (1985, Table 1)



(b) Iterations of EM

t	$\theta_{11}^{(t)}$	$\theta_{12}^{(t)}$	$\theta_{21}^{(t)}$	$\theta_{22}^{(t)}$
0	.2500	.2500	.2500	.2500
1	.6615	.1170	.1498	.0718
2	.6947	.1003	.1370	.0680
3	.6969	.0988	.1359	.0684
4	.6971	.0987	.1358	.0684
5	.6971	.0986	.1358	.0685
∞	.6971	.0986	.1358	.0685

did not respond to the survey at either visit, we are left with a sample of n = 641 for which responses are available at one or both occasions. The EM algorithm for this example converges quite rapidly. Starting from a table of uniform probabilities (all θ_{ij} = 0.25), the estimated cell probabilities converge to four decimal places by the fifth iteration as shown in Table 3.3 (b).

One way to summarize the association between two binary variables is by the cross-product or odds ratio

$$\omega = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}},$$

with $\omega = 1$ under independence. The ML estimate of the odds ratio is $\left(\hat{\theta}_{11}\hat{\theta}_{22}\right)/\left(\hat{\theta}_{12}\hat{\theta}_{21}\right)$=3.57. Households that were victimized during the first period appear to be more than 3.5 times as likely, on the odds scale, to have been victimized in the second period than households that were crime-free in the first period. The question of whether this result is statistically significant will be addressed shortly.

3.2.3 EM for posterior modes

The EM algorithm is typically presented as a technique for finding MLEs. As pointed out by Dempster, Laird, and Rubin (1977), however, EM may also be used to compute posterior modes, values of θ for which the observed-data posterior density rather than the observed-data likelihood is highest.

Because the complete-data posterior density under the prior $\pi(\theta)$ is $P(\theta|Y) \propto P(\theta|Y)\,\pi(\theta)$, it follows from (3.2) that

$$\log P(\theta \mid Y) = l(\theta \mid Y_{obs}) + \log P(Y_{mis} \mid Y_{obs}, \theta) + \log \pi(\theta) + c.$$

Averaging this equation over the predictive distribution of Y_{mis} given $\theta = \theta^{(t)}$ gives

$$Q*\left(\theta \mid \theta^{(t)}\right) = \log P(\theta \mid Y_{obs}) + H\left(\theta \mid \theta^{(t)}\right) + \log \pi(\theta) + c,$$

where

$$Q*\left(\theta \mid \theta^{(t)}\right) = Q\left(\theta \mid \theta^{(t)}\right) + \log \pi\left(\theta^{(t)}\right),$$

and the functions $Q(\theta|\theta^{(t)})$ and $H(\theta|\theta^{(t)})$ are defined as before. If we choose the next iterate $\theta^{(t+1)}$ to maximize $Q^*(\theta|\theta^{(t)})$ i.e. to satisfy

$$Q*\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q*\left(\theta \mid \theta^{(t)}\right) \text{ for all } \theta,$$

then each iteration will increase $P(\theta|Y_{obs})$ and in a well-behaved problem the sequence of parameter estimates will converge to the mode of $P(\theta|Y_{obs})$.

It is evident that when the prior $\pi(\theta)$ is chosen to be a constant function over the parameter space, this algorithm reduces to the maximum-likelihood version of EM. If the prior is not constant the M-step will change, requiring maximization of $Q^*(\theta|\theta^{(t)})$ rather than $Q(\theta|\theta^{(t)})$. The E-step procedure will be the same as in the maximum-likelihood version, however, because the E-step is dependent upon a fixed value of θ and therefore does not involve the prior.

3.2.4 Restrictions on the parameter space

Thus far little has been said about the parameter space or domain of θ. In many problems it will be the natural parameter space, the set of all values of θ for which $P(Y|\theta)$ is a valid probability density or probability function. In Example 2 of Section 3.2.2, for instance, we assumed nothing about the multinomial parameter π except the minimal requirements $\pi_{ij} \geq 0$ for i,j = 1,2 and $\pi_{++} = \pi_{11} + \pi_{12} + \pi_{22} = 1$. In other situations, however, it is desirable to restrict θ to lie within some smaller set Θ_0, a subset of the natural parameter space which is typically of lower dimension. In the 2 × 2 contingency table, for example, we could require the cell probabilities to satisfy the condition of row-column independence, $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. Appropriate choices for Θ_0 generate useful classes of models in a variety of continuous and categorical-data contexts.

It often happens that we want to test a null hypothesis $\theta \in \Theta_0$ versus an alternative hypothesis $\theta \in v_1$, where Θ_0 is a lower-dimensional subset of Θ_1. If $\hat{\theta}_1$ is the maximizer of Θ_1 over Θ_0 and $\hat{\theta}_1$ the maximizer over Θ_1, then the well known large-sample approximation

$$2l\left(\hat{\theta}_1 \mid Y_{obs}\right) - 2l\left(\hat{\theta}_0 \mid Y_{obs}\right) \sim \chi_d^2 \qquad (3.12)$$

under the null hypothesis, where $d = \dim\Theta_1 - \dim\Theta_0$, forms the basis for a *likelihood-ratio test* (e.g. Cox and Hinkley, 1974). If the drop in $2l(\theta|Y_{obs})$ as we move from $\hat{\theta}_1$ to $\hat{\theta}_0$ is unusually large when compared to the χ_d^2 distribution, then the evidence against the null hypothesis in favor of the alternative is strong. In performing this test one needs to maximize the likelihood twice, once over Θ_0 and once over Θ_1.

Using the EM algorithm to maximize a likelihood or posterior over a restricted parameter space is conceptually no different from applying EM without such restrictions. The form of the E-step will not change, because taking the expectation of a quantity with respect to $P(Y_{mis}|Y_{obs},\theta)$ is computationally the same whether or not $\theta \in \Theta_0$. The M-step, however, will become a constrained maximization of the expected complete-data likelihood or posterior over Θ_0 and hence may require special, often iterative, optimization techniques. To avoid implementing an iterative M-step which would make the EM algorithm doubly iterative, it is often helpful to apply a recent extension of EM known as ECM, described below in Section 3.2.5.

Example: testing hypotheses for an incomplete 2 × 2 table

Returning to the National Crime Survey data in Table 3.3, we can apply the EM algorithm under the restriction that victimization status on the first occasion is independent of victimization status on the second occasion. The E-step is the same as in the unrestricted case, but the M-step is different. It is well known that with complete data, the ML estimates under independence are

$$\tilde{\theta}_{ij} = \frac{x_{i}+x_{+j}}{n^2} = \frac{(x_{i1}+x_{i2})(x_{1j}+x_{2j})}{n^2} \tag{3.13}$$

for i,j=1,2. The M-step of EM uses (3.13) with each count x_{ij} replaced by $E(x_{ij}|\theta,Y_{obs})$, the output of the E-step. Starting from a uniform table, EM converges after four iterations to the restricted ML estimate

$$\tilde{\theta} = \left(\tilde{\theta}_{11}, \tilde{\theta}_{12}, \theta_{21}, \tilde{\theta}_{22}\right)$$
$$= (0.6631, 0.1329, 0.1699, 0.0341).$$

Notice that the estimated cross-product ratio $\left(\tilde{\theta}_{11}\tilde{\theta}_{22}\right)/\left(\tilde{\theta}_{12}\tilde{\theta}_{21}\right)$ = 1.00 satisfies the independence condition as required.

To test whether row-column independence is plausible, we can perform a likelihood-ratio test. For the households that responded to the survey on both occasions, the loglikelihood contribution has the form of a complete-data multinomial loglikelihood,

$$l_A(\theta \mid Y_{obs}) = x_{11}^A \log\theta_{11} + x_{12}^A \log\theta_{12} + x_{21}^A \log\theta_{21} + x_{22}^A \log\theta_{22}.$$

The households that responded to the survey only on the first occasion provide information only about the marginal probabilities θ_{1+}=θ_{11}+θ_{12} and θ_{2+}=θ_{21}+θ_{22}; their loglikelihood contribution has the form of a binomial,

$$l_B(\theta \mid Y_{obs}) = x_{+1}^B \log\left(\theta_{11} + \theta_{12}\right) + x_{2+}^B \log\left(\theta_{21} + \theta_{22}\right).$$

Similarly, the households that responded only on the second occasion contribute

$$l_C(\theta \mid Y_{obs}) = x_{+1}^C \log\left(\theta_{11} + \theta_{21}\right) + x_{+2}^C \log\left(\theta_{12} + \theta_{22}\right).$$

The observed-data loglikelihood is the sum of the loglikelihood contributions from each missingness pattern,

$$l(\theta \mid Y_{obs}) = l_A(\theta \mid Y_{obs}) + l_B(\theta \mid Y_{obs}) + l_C(\theta \mid Y_{obs}).$$

Plugging in the observed data and the restricted ML estimate $\tilde{\theta}$ yields $l\left(\tilde{\theta} \mid Y_{obs}\right)$ = -575.19, the highest loglikelihood achievable under independence. Plugging in the unrestricted ML estimate $\hat{\theta}$ from Table 3.3 (b) yields $l\left(\hat{\theta} \mid Y_{obs}\right)$ = -562.50. The likelihood-ratio test statistic is thus 2(-562.50 + 575.19) =

25.38, which is well beyond the plausible range of a χ_1^2 random variate. We therefore conclude, not surprisingly, that victimization status on the two occasions is related.

Perhaps a more interesting and appropriate question for these data is not whether victimization during the two periods seems independent, but whether the victimization rate seems to have changed over time. That is, it may be of interest to test the hypothesis of marginal homogeneity, $\theta_{k+} = \theta_{+k}$, $k = 1,2$, which for a 2 × 2 table is equivalent to the hypothesis of off-diagonal symmetry, $\theta_{12} = \theta_{21}$. With complete data, a commonly used procedure for assessing marginal homogeneity/symmetry in a 2 × 2 table is *McNemar s test*, in which the statistic

$$M = \frac{x_{12} - x_{21}}{\sqrt{x_{12} + x_{21}}}$$

is compared to the standard normal distribution (McNemar, 1947; Agresti, 1990). With incomplete data, we can perform a likelihood-ratio test by maximizing the likelihood subject to $\theta_{12} = \theta_{21}$. Under, this restriction, we may collapse the off diagonal cells into a single cell and express the complete-data loglikelihood as that of a trinomial,

$$l(\theta \mid Y) = x_{11} \log \theta_{11} + (x_{12} + x_{21}) \log(2\theta_{12}) + x_{22} \log \theta_{22,}$$

with sufficient statistics x_{11}, x_{22} and $(x_{12} + x_{21})$; the moment equations (3.7) lead immediately to the ML estimates $\breve{\theta}_{11} = x_{11}/n$, and $\breve{\theta}_{12} = \breve{\theta}_{21} = (x_{12}+x_{21})/(2n)$. Revising the M-step to include the marginal homogeneity/symmetry restriction, EM quickly converges to

$$\breve{\theta} = (0.6970, 0.1173, 0.1173.0, 0685).$$

The loglikelihood at this estimate is $l(\breve{\theta} \mid Y_{obs}) = -564.25$, so the statistic for testing the null hypothesis of marginal homogeneity/symmetry is 2(-562.50 + 564.25) = 3.50 with a p-value of $P\left(\chi_1^2 \geq 3.50\right) = 0.06$. The evidence against marginal homogeneity/symmetry is thus fairly strong. Extensions of this procedure for testing marginal homogeneity in $r \times r$ tables for $r > 2$ are more complicated, because ML estimates do not exist in closed form; the M-step may be carried out using techniques of nonlinear programming, as discussed by Shih (1987).

3.2.5 The ECM algorithm

The ECM or Expectation-Conditional Maximization algorithm is a useful extension of EM for situations where the M-step cannot be carried out without iteration (Meng and Rubin, 1993). ECM replaces a complicated M-step with a sequence of simpler conditional or constrained maximizations known as a CM-step. ECM retains the reliable convergence properties of EM while simplifying, and often reducing, the required computations.

The CM step of ECM is comprised of S conditional maximizations in which the Q function is maximized not over the entire parameter space as in (3.5), but over a smaller set in which a vector-valued function $g_s(\theta)$ is fixed at its previous value for $s = 1,2, ,\quad S$. The set of functions $G = \{g_s(\theta) : s = 1, , S\ \}$ must be pre-selected and must satisfy precise conditions defined by Meng and Rubin (1993). Once a is specified, one iteration of ECM proceeds as follows. Given the current value of the parameter $\theta^{(t)}$, first perform an E-step to obtain $Q(\theta|\theta^{(t)})$ as in the EM algorithm. Then find $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$ subject to the constraint

$$g_s(\theta) = g_s\left(\theta^{(t+(s-1)/S)}\right)$$

for s=1,2,...,S. The resulting parameter value $\theta^{(t+S/S)} = \theta^{(t+1)}$ becomes the input to the next E-step. Clearly $Q(\theta^{(t+1)}|\theta^{(t)})$ must be at least as large as $Q(\theta^{(t)}|\theta^{(t)})$, so subsequent iterations will never decrease the observed-data loglikelihood. Moreover, it can be shown that a stationary point of ECM, i.e. a value $\theta^{(t)}$ such that $\theta^{(t+1)} = \theta^{(t)}$, is also a stationary point and typically a maximum of the observed-data loglikelihood.

The basic condition required of the set of constraints G in ECM is that repeated application of the CM-step in the absence of missing data would result in the likelihood being maximized over the whole parameter space. Many, but not all, iterative algorithms used for maximizing likelihoods with complete data can be interpreted as repeated application of a CM-step. One example is the method of cyclic ascent or iterated conditional modes (Besag, 1986), in which θ is partitioned into a set of S subvectors and the likelihood function is successively maximized with respect to each subvector holding the others constant. Another example is iterative proportional fitting of loglinear models for contingency tables (Bishop, Fienberg, and Holland, 1975), in which the estimated cell probabilities are proportionately adjusted at each step to match the observed cell proportions on sets of margins determined by the model. These algorithms have often been regarded as less desirable than gradient methods such as Newton-Raphson, because they tend to converge more slowly in complete-data problems. When paired with an E-step and applied to incomplete data, however, they tend to produce computationally stable and reliable ECM algorithms that are guaranteed to increase the likelihood at each step.

Our uses of the ECM algorithm will be confined to loglinear models for categorical data (Chapter 8) and models for mixed continuous and categorical data that employ loglinear constraints (Chapter 9). Other examples of ECM and further references are given by Meng and Rubin (1993).

3.3 Properties of EM

3.3.1 Stationary values

For incomplete-data problems, the most attractive features of EM relative to other optimization techniques are its simplicity and its stability. Rather than maximizing the potentially complicated function $l(\theta|Y_{obs})$ directly, we repeatedly maximize the Q function, which is typically much easier and often equivalent to finding MLEs with complete data. Moreover, successive iterations of EM are guaranteed never to decrease $l(\theta|Y_{obs})$ which is not generally true of gradient methods like Newton-Raphson. In practice, evaluating $l(\theta|Y_{obs})$ at each step to ensure that it is increasing is often helpful for assessing the progress of EM and for detecting and diagnosing programming errors.

As with any optimization technique, however, EM is not guaranteed to always converge to a unique global maximum. In well-behaved problems the function $l(\theta|Y_{obs})$ is unimodal and concave over the entire parameter space, in which case EM converges to the unique MLE $\hat{\theta}$ from any starting value. Exceptions occur both in theory and in practice, however, and one needs to be alert to detect these abnormalities when they arise.

Multiple modes

Table 3.4 shows a hypothetical bivariate dataset reported by Murray (1977) with twelve sample units and four missing values for each variable. Under the assumption that the complete data are iid observations of a bivariate normal vector (Y_1, Y_2), one may calculate the observed-data likelihood using (2.10). It is apparent that the likelihood function is symmetric with respect to Y_1 and Y_2, and that the marginal means (μ_1, μ_2) and variances $(\sigma_{11}, \sigma_{22})$ are relatively better estimated than the correlation coefficient $\rho = \sigma_{12} / \sqrt{\sigma_{11}\sigma_{22}}$. Using analytical methods or the EM algorithm for multivariate normal data to be presented in Section 5.3, one may

Table 3.4. Bivariate dataset with missing values

1	1	−1	−1	2	2	−2	−2	?	?	?	?
1	−1	1	−1	?	?	?	?	2	2	−2	−2

verify that $l(\theta/Y_{obs})$ has two modes, one at $\theta = (\mu_1,\mu_2,\sigma_{22,}\rho) =$ (0, 0, 2.67, 2.67, 0.5) and the other at $\theta =$ (0, 0, 2.67, 2.67, -0.5). EM will converge to the first mode if started with $\rho^{(0)} > 0$ and to the second mode if $\rho^{(0)} < 0$.

The bimodality in this example is due to the symmetry of the data and the unusual pattern of missingness in which the observations with high leverage, i.e. those that would contribute the most information about ρ, have one of the two components missing. In real multivariate datasets where the data are sparse and/or the missingness pattern is unusually pernicious, multiple modes do sometimes occur. Unlike this symmetric example in which the likelihood values at the two modes are equal, there will typically be one major and one or more minor modes, and the mode to which EM converges may change depending on the choice of $\theta^{(0)}$. To detect multiple modes, it is helpful to run EM from a variety of starting values to see whether it always converges to the same answer.

Saddlepoints

A saddlepoint is a value of θ at which the directional derivatives of $l(\theta|Y_{obs})$ are zero but which is neither a local maximum nor a local minimum. EM could possibly converge to a saddlepoint, but in practice this is quite rare. Convergence to a saddlepoint requires not only that the saddlepoint exist, but that the successive iterates of θ approach it only from certain directions. The loglikelihood for the data in Table 3.4 has a saddlepoint at $\theta =$ (0, 0, 2.5, 2.5, 0), and EM converges to it if started from $\rho^{(0)} = 0$. However, even a very slight perturbation from $\rho = 0$ will cause EM to leave the saddle and go to one of the two modes. In real data examples convergence to a saddlepoint is rarely encountered, and thus saddlepoints are not a cause for concern.

Likelihood ridges

It may happen that the maximum value of the likelihood function is achieved not at a single value of θ but at a whole continuum of values. This phenomenon occurs when one or more components or functions of θ are inestimable in the sense that they do not appear in the likelihood, and thus $l(\theta|Y_{obs})$ is the same for any value of those components; that is, $l(\theta|Y_{obs})$ is flat in certain directions. Consider the bivariate dataset shown below.

-2	-1	0	1	2	?	?	?	?	?
?	?	?	?	?	2	1	0	-1	-2

Under the bivariate normal model $l(\theta|Y_{mis})$ is the sum of two complete-data loglikelihoods, one pertaining to (μ_1, σ_{11}), and the other pertaining to (μ_2, σ_{22}), and the correlation ρ is inestimable. The maximum value of $l(\theta|Y_{mis})$ is achieved for $\hat{\mu}_1 = \hat{\mu}_2 = 0$, $\hat{\sigma}_{11} = \hat{\sigma}_{22} = 2$ and any value of ρ, so the likelihood is said to have a one-dimensional ridge. If EM were applied to this dataset from various starting values, it would converge to different points on the ridge. When the likelihood has a ridge, any value along the ridge is a stationary value of EM. The algorithm does not wander aimlessly on the ridge but stops once the ridge is reached.

When two normal variables are not observed together, the correlation between them will be inestimable and the likelihood will have a ridge. Similar results apply to incomplete datasets with three or more variables. Consider the trivariate case where Y_1 and Y_2 are sometimes observed together, Y_1 and Y_3 are sometimes observed together, but Y_2 and Y_3 are never jointly observed; in this case it can be shown that the partial correlation of Y_2 and Y_3 given Y_1 is inestimable. Estimability of parameters in multivariate normal datasets where not all variables are observed together is discussed by Rubin (1974) and Rubin and Thayer (1978). It should be noted that merely having joint observations of all variables is not sufficient to guarantee uniqueness of the MLE, as in the example below.

-2	-1	1	2	0	0	?	?	?	?
?	?	?	?	0	0	2	1	-1	-2

Here the two joint observations of Y_1. and Y_2 are identical and thus provide no information about ρ.

Boundary estimates

Yet another abnormality is an ML estimate on the boundary of the parameter space. Consider the dataset below for which the ML estimates are $\hat{\mu}_1 = \hat{\mu}_2 =$, $\hat{\sigma}_{11} = \hat{\sigma}_{22}$ = 1.80 and = $\hat{\rho}$ = -1.

-2	0	0	2	-1	1	?	?	?	?
?	?	?	?	1	-1	2	0	0	-2

If convergence is assessed by relative changes in the components of $\theta = (\mu_1,\mu_2,\sigma_{11},\sigma_{22},\rho)$ then EM will converge reliably in this example. If ρ is examined on some open-ended scale, however (for example, the familiar Fisher's *z*-transformation

$$z = \tfrac{1}{2} \log \frac{1+\rho}{1-\rho}, \tag{3.14}$$

which takes values on the whole real line) then the iterations of EM will appear to diverge as $z \to -\infty$.

General comments on the method of maximum likelihood

It is important to note that the abnormalities described above, multiple modes, saddlepoints, ridges and boundary solutions, are not shortcomings of the EM algorithm but inherent features of $l(\theta|Y_{obs})$ that would impact any optimization method. Indeed, when such features exist, EM is often remarkably well-behaved in comparison with other computational methods. With incomplete multivariate data, these abnormalities Are typically associated with small samples, high rates of missingness and models that are clearly overparameterized (i.e. having too many parameters) relative to the amount of information in Y_{obs}. In the data of Table 3.4, for example, an unusually large portion of the information in Y about ρ is concentrated in Y_{mis} and inferences about ρ will be highly sensitive to untestable assumptions about missing data and the missing-data mechanism.

From a theoretical point of view, the desirability of ML estimates stems primarily from their large-sample properties. Under suitable regularity conditions, MLEs are asymptotically unbiased, normal, and efficient with variance determined by the curvature of the loglikelihood near the mode (e.g. Cox and Hinkley, 1974). In large samples the loglikelihood function tends to be unimodal and approximately quadratic. In such cases an MLE and an estimate of its variance provide an excellent summary of the data s information about θ, and large-sample procedures such as asymptotic confidence intervals, likelihood-ratio tests, etc. will tend to be reliable. When this is not the case, however, when the loglikelihood is oddly-shaped with multiple modes, suprema on the boundary, etc., the behavior of large-sample procedures may be seriously impaired and attractiveness of the ML method is greatly diminished.

When abnormalities are found in $l(\theta|Y_{obs})$, the analyst is faced with several options. One option is to reduce the size of the model by eliminating parameters or imposing restrictions on the parameter space. Another possibility is to introduce additional information about θ through a prior distribution π and base inference on the observed-data posterior $P(\theta|Y_{obs})$ rather than the observed-data likelihood. A posterior mode will tend to be a better estimate of θ than the MLE when substantial prior knowledge about θ is available. Even when prior knowledge is scarce, however, we will find that adding small amounts of information through π may be a useful technique for ensuring that EM converges to a unique value of θ in the interior of the parameter space. When data are sparse or missing values occur in such a way that one or more components of θ are poorly estimated, a judiciously chosen prior may greatly improve the numerical stability of computations and perhaps even strengthen the estimate of θ from a statistical point of view.

3.3.2 Rate of convergence

Like any algorithm for successive approximation, EM implicitly defines a function that maps the parameter space to itself. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_{\mathrm{k}})^T$ be the k-dimensional parameter. Denote a single iteration of EM by

$$\theta^{(t+1)} = M\left(\theta^{(t)}\right) = \left(M_1\left(\theta^{(t)}\right), M_2\left(\theta^{(t)}\right), \ldots M_k\left(\theta^{(t)}\right)\right)^T$$

so that both the E and M-steps are incorporated into the vector function M. Expanding $M(\theta^{(t)})$ in a Taylor series about $\hat{\theta}$ gives a first-order approximation

$$M\left(\theta^{(t)}\right) - M\left(\hat{\theta}\right) \approx M'\left(\hat{\theta}\right)\left(\theta^{(t)} - \hat{\theta}\right) \tag{3.15}$$

in the neighborhood of $\hat{\theta}$, where $M'(\theta)$ is the $k \times k$ first-derivative or Jacobian matrix for $M(\theta)$ with typical element $\partial M_i(\theta)/\partial\theta_j$. If $\hat{\theta}$ is a stationary value of EM, then $M(\hat{\theta}) = \hat{\theta}$ and (3.15) becomes

$$\left(\theta^{(t+1)} - \hat{\theta}\right) \approx M'\left(\hat{\theta}\right)\left(\theta^{(t)} - \hat{\theta}\right) \tag{3.16}$$

or $\varepsilon^{(t+1)} = D\varepsilon^{(t)}$, where $\varepsilon^{(t)} \approx \theta^{(t)} - \hat{\theta}$ is the error in approximation at step t and D is shorthand for $M'\left(\hat{\theta}\right)$. EM s convergence is thus said to be linear, because $\varepsilon^{(t+1)}$ is approximately a linear transformation of $\varepsilon^{(t)}$ near the mode. Newton-Raphson and other superlinear methods have the property that $D = 0$ so the Taylor series approximation in (3.15) is dominated by the smaller second-order term.

The speed at which EM converges in any particular application is determined by the rate matrix D. In the case of a scalar parameter we have $\left|\varepsilon^{(t+1)}\right| \approx D\left|\varepsilon^{(t)}\right|$ where D is a single number between 0 and 1. The convergence will be rapid when D is near zero and slow when D is near one. The situation for $k \geq 2$ is more complicated, however, and depends on the eigenstructure of D.

Any vector v such that $Dv = \lambda v$ for some constant λ is said to be an eigenvector of D, and λ is its associated *eigenvalue*. The eigenvalues must also satisfy the equation $|D - \lambda I| = 0$, and because the determinant of a $k \times k$ matrix is a polynomial of order k this equation has at most k distinct roots. When the roots $\lambda_1, \lambda_2, \ldots, \lambda_k$ are distinct the corresponding eigenvectors $v_1, v_2, \ldots, v_k$ are linearly independent, and any k-dimensional vector can be written as a linear combination of the eigenvectors. In particular, we can write the error vector $\varepsilon^{(t)} = \theta^{(t)} - \hat{\theta}$ as

$$\varepsilon^{(t)} - c_1 v_1 + c_2 v_2 + \cdots c_k v_k .$$

Then the error at the next iteration becomes

$$\begin{aligned}\varepsilon^{(t+1)} &\approx D\left(c_1 v_1 + c_2 v_2 + \cdots c_k v_k\right)\\ &\approx c_1 \lambda_1 v_1 + c_2 \lambda_2 v_2 + \cdots + c_k \lambda_k v_k,\end{aligned}$$

and after r iterations,

$$\varepsilon^{(t+r)} \approx c_1 \lambda_1^r v_1 + c_2 \lambda_2^r v_2 + \cdots + c_k \lambda_k^r v_k.$$

In ordinary problems all the eigenvalues will satisfy $0 \leq \lambda_j < 1$ and successive iterations of EM beginning from any $\theta^{(t)}$ in a neighborhood of $\hat{\theta}$ will shrink the error toward zero. If $\hat{\theta}$ is a saddlepoint then one or more eigenvalues could exceed one. If $\varepsilon^{(t)}$ happens to be precisely orthogonal to the eigenvectors corresponding to those eigenvalues, then EM will converge to the saddlepoint. For a randomly chosen $\theta^{(t)}$ in the neighborhood of $\hat{\theta}$, however, this will happen with negligible probability, so in most cases the iterates will diverge from a saddlepoint. One or more eigenvalues equal to one indicates that the likelihood is flat in certain directions and is maximized along a ridge (Dempster, Laird and Rubin, 1977).

The missing information principle

It is well known that in regular problems the large-sample precision of the MLE is determined by the curvature of the loglikelihood function. With complete data, the Fisher information is defined to be

$$I^*(\theta \mid Y) = -\int \left[\frac{\partial^2}{\partial \theta^2} l(\theta \mid Y) \right] P(\theta \mid Y) dY, \tag{3.17}$$

where $\partial^2 l(\theta|Y)/\partial\theta^2$ is the $k \times k$ matrix with typical element $\partial^2 l(\theta|Y)/\partial\theta_i\partial\theta_j$. One estimate of the covariance matrix of $\hat{\theta}$ in large samples is $\left[I*\left(\hat{\theta}\mid Y\right)\right]^{-1}$ the inverse of the Fisher information matrix evaluated at the complete-data MLE. Another, asymptotically equivalent, estimate is the inverse of

$$I\left(\hat{\theta}\mid Y\right)=-\frac{\partial^2}{\partial\theta^2}l(\theta\mid Y)\Big|_{\theta=\hat{\theta}}, \tag{3.18}$$

which fixes *Y* In the loglikelihood function at its realized value rather than averaging over its distribution.

With incomplete data, differentiating (3.2) twice yields

$$-\frac{\partial^2}{\partial\theta^2}l(\theta\mid Y)=-\frac{\partial^2}{\partial\theta^2}l\left(\theta\mid Y_{obs}\right)-\frac{\partial^2}{\partial\theta^2}\log P\left(Y_{mis}\mid Y_{obs},\theta\right).$$

Taking the expectation of this over $P(Y_{mis}|Y_{obs},\theta)$, we obtain a fundamental relationship: the complete information is equal to the observed information plus the missing information. This relationship, called the *missing information principle* by Orchard and Woodbury (1972), was also investigated by Dempster, Laird and Rubin (1977), Louis (1982) and Meng and Rubin (1991a). Assuming sufficient regularity to interchange the order of differentiation and integration, we can write the complete information as

$$I_c(\theta)=-\frac{\partial^2}{\partial\theta^2}Q(\theta\mid\theta),$$

the observed information as

$$I_o(\theta)=-\frac{\partial^2}{\partial\theta^2}l\left(\theta\mid Y_{obs}\right),$$

and the missing information as

$$I_m(\theta)=-\frac{\partial^2}{\partial\theta^2}H(\theta\mid\theta),$$

so that

$$I_c(\theta) = I_o(\theta) + I_m(\theta).$$

Note that each quantity in (3.19) is a function of Y_{obs} although this fact has been suppressed in the notation. Also note that the concept of information used in (3.19) is more consistent with (3.18) than with (3.17), because we have fixed Y_{obs} at its realized value and averaged only over the distribution of the unknown Y_{mis}. A natural large-sample estimate of the covariance matrix with incomplete data is

$$I_0^{-1}\left(\hat{\theta}\right) = \left[-\frac{\partial^2}{\partial\theta^2} l(\theta \mid Y_{obs}) \right]^{-1}_{\theta=\hat{\theta}}, \tag{3.20}$$

where $\hat{\theta}$ is now the observed-data MLE, the maximizer of $l(\theta|Y_{obs})$.

Missing information and convergence

Dempster, Laird and Rubin (1977) established an important connection between these information quantities and D = $M'\left(\hat{\theta}\right)$, the asymptotic rate matrix of EM. In regular problems where $\theta^{(t+1)}$ is obtained as a solution to $\partial Q(\theta|\theta^{(t)})/\partial\theta=0$, they showed that

$$D = I_c^{-1}\left(\hat{\theta}\right) I_m\left(\hat{\theta}\right). \tag{3.21}$$

In the extreme case where Y_{mis} provides no additional information about θ not already contained in Y_{obs}, then $I_m\left(\hat{\theta}\right)$ = 0 and (3.21) implies that EM essentially converges in a single iteration. More generally, for a scalar parameter (3.21) implies that D is the ratio of the missing information to the complete information. For brevity we will call this ratio the *fraction of missing information*, although a more precise term would be the fraction of information missing due to nonresponse. If we denote the fraction of missing information in the scalar case by $D = \lambda$, then each iteration approximately multiplies the error by λ,

$$\left(\theta^{(t+1)} - \hat{\theta}\right) \approx \lambda\left(\theta^{(t)} - \hat{\theta}\right), \tag{3.22}$$

which demonstrates one of the fundamental properties of the EM algorithm: the rate of convergence of EM is determined by the fraction of missing information.

When θ is a vector of length $k > 1$ the fraction of missing information is no longer a number but a matrix; yet a result similar to (3.22) holds for multiparameter problems as well. Suppose we order the eigenvalues of D so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, and let $v_1, v_2, \ldots, v_k$ be eigenvectors of D corresponding to these ordered eigenvalues. As before, we can write the error vector as

$$\varepsilon^{(t)} = \theta^{(t)} - \hat{\theta} = c_1 v_1 + c_2 v_2 + c_k v_k$$

for some $c_1, c_2, \ldots, c_k$, so that after r iterations

$$\varepsilon^{(t+r)} \approx c_1 \lambda_1^r v_1 + c_2 \lambda_2^r v_2 + \cdots + c_k \lambda_k^r v_k.$$

By analogy with (3.22) we may regard $\lambda_1, \lambda_2, \ldots, \lambda_k$ as fractions of missing information corresponding to the particular directions v_1, $v_2, \ldots, v_k$. Moreover, if λ_2 is strictly less than λ_1, we have

$$\varepsilon^{(t+r)} \approx \lambda_1^r \left(c_1 v_1 + R\right), \tag{3.23}$$

where the remainder term

$$R = c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^r v_2 + \cdots + c_k \left(\frac{\lambda_k}{\lambda_1}\right)^r v_k$$

approaches zero at a rate determined by λ_2/λ_1. Thus in the vicinity of the mode

$$\left(\theta^{(t+1)} - \hat{\theta}\right) \approx \lambda_1\left(\theta^{(t)} - \hat{\theta}\right) \tag{3.24}$$

where λ_1, is the largest eigenvalue of D. If the s largest eigenvalues happen to be equal, then (3.23) becomes

$$\varepsilon^{(t+r)} \approx \lambda_1^r\left(c_1 v_1 + \cdots + c_s v_s + R\right)$$

where R approaches zero at a rate determined by λ_{s+1}/λ_1, and (3.24) still holds. In the multiparameter case, we can thus say that *EM s rate of convergence is governed by the largest fraction of missing information*. Exceptions to this rule are possible; for example, if $\varepsilon^{(t)}$ happens to be precisely orthogonal to the eigenvector(s) corresponding to the largest eigenvalue, then convergence will be dominated by the next largest eigenvalue. For most real-data problems, however, this basic result does hold. Further results and discussion on the convergence of EM are given by Meng (1990).

3.3.3 Example

In Example 1 of Section 3.2.2, we derived the EM algorithm for incomplete univariate normal data and applied it to the n_1 = 10 observations displayed in Table 3.1 (a) by assuming that an additional n_0 = 3 observations were missing. By varying the choice of no we can make the rate of convergence in this example arbitrarily large or small. Figure 3.1 displays the iterations of EM in this two-parameter problem from a variety of starting values under n_0 = 10 and n_0 = 90, corresponding to missingness rates of 50% and 90%, respectively. For visual clarity the variance ψ is shown on the log scale. With n_0 = 10 the convergence is quite rapid, whereas for n_0 = 90 it is much slower. Unlike gradient methods EM does not necessarily follow the path of steepest ascent, but often climbs the loglikelihood surface by a more circuitous route.

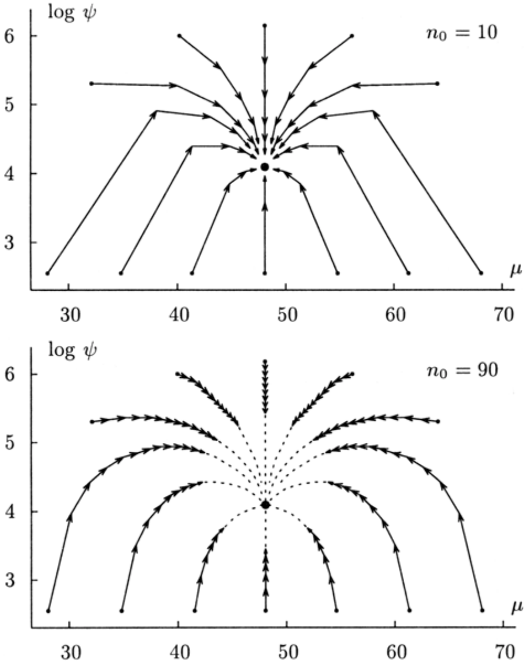


Figure 3.1. *Iterations of EM from various starting values for univariate normal data with* $n_\text{o} = 10$ *and* $n_\text{o} = 90$.

In simple examples like this one, it is feasible to investigate the convergence properties analytically. The elements of the matrix $M'\left(\theta^{(t)}\right)$, obtained by differentiating (3.9)-(3.10), are

$$\frac{\partial\mu^{(t+1)}}{\partial\mu^{(t)}} = \frac{n_0}{n},$$

$$\frac{\partial\mu^{(t+1)}}{\partial\psi^{(t)}} = 0,$$

$$\frac{\partial\psi^{(t+1)}}{\partial\mu^{(t)}} = 2\left(\frac{n_0 n_1}{n^2}\right)\left(\mu^{(t)} - \bar{y}_{obs}\right),$$

$$\frac{\partial\psi^{(t+1)}}{\partial\psi^{(t)}} = \frac{n_0}{n},$$

where $y_{obs} = n_1^{-1}\sum_{i=1}^{n_1} y_i$ is the observed-data MLE for μ. Evaluating these derivatives at the mode (i.e. taking $\theta^{(t)} = \hat{\theta}$) gives

$$D = M'\left(\hat{\theta}\right) = \begin{bmatrix} n_0 / n & 0 \\ 0 & n_0 / n \end{bmatrix}.$$

The eigenvalues of this matrix are $\lambda_1 = \lambda_2 = \lambda = n_0/n$. Whenever eigenvalues are repeated, the corresponding eigenvectors are not uniquely defined. In fact, because D is proportional to the identity matrix in this example, any error vector $\varepsilon^{(t)}$ is an eigenvector. Yet the overall convergence rate is still governed by A, and a single iteration of EM can be expressed as $\varepsilon^{(t+1)} \approx \lambda\varepsilon^{(t)}$.

In this and other univariate examples, the fraction of missing information λ is also the rate of missing observations. In multi-variate applications, however, the fractions of missing information corresponding to various components of θ will typically differ from the rates of missing observations, both overall and on a variable-by-variable basis, depending on the pattern of missingness and the observed interrelationships among variables.

3.3.4 Further comments on convergence

Monitoring and detecting convergence

Two basic methods for monitoring the convergence of EM involve (a) successive parameter values $\theta^{(t)}$, and (b) successive values of the observed-data loglikelihood $l(\theta^{(t)}|Y_{obs})$. In practice both are quite useful. Because each iteration of EM is guaranteed never to decrease the likelihood, evaluating $l(\theta|Y_{obs})$ at each step is helpful for detecting programming errors as well as for monitoring the progress of EM in specific examples. Convergence is typically judged by examining changes in individual components of $\theta = (\theta_1,\theta_2,\ldots,\theta_k)$ from one iteration to the next. If the changes are all relatively small, for example, if

$$\left|\theta_j^{(t)} - \theta_j^{(t-1)}| \leq \in |\theta_j^{(t)}\right|$$

for $j = 1, 2,..., k$ and a suitably small $\in$ (say, 0.0001), then we may say that EM has converged by iteration t.

If the elements of θ continue to change for many iterations with very little increase in $l(\theta|Y_{obs})$, then it should be taken as a sign that the loglikelihood is nearly flat in certain directions and that one or more functions of θ are very poorly estimated. Often in these problems there is little to be gained from additional iterations, because the value of θ to which EM would ultimately converge has loglikelihood only slightly higher than the current value; the observed data do little to distinguish between these values of θ. Slow convergence is also a sign that the fractions of missing information corresponding to certain aspects of θ are close to one, and that most of the information about them is being contributed by $P(Y_{mis}|Y_{obs},\theta)$, the model for the missing data, rather than by Y_{obs}. Because the correctness of $P(Y_{mis}|Y_{obs},\theta)$ rests on both the complete-data model and the ignorability assumption, slow convergence warns us that inferences about certain aspects of θ are highly model-dependent. This does not automatically mean that all inferences about θ are suspect, because in some problems (e.g. when the missing data are missing by design and therefore known to be MAR) the model may be quite trustworthy. In other problems, however, particularly when it is not known whether the data are MAR, slow convergence is a useful warning that the estimate of θ may rest largely on our assumptions about the unknown Y_{mis}, rather than on the known Y_{obs}. If so, then a simplification of the model, perhaps by imposing restrictions on the parameter or by eliminating variables with high rates of missingness, may be a sensible strategy.

Asymptotic covariance matrices from EM

One drawback of EM relative to other optimization techniques is that it does not automatically provide standard errors associated with the parameter estimates. The asymptotic covariance matrix $I_0^{-1}\left(\hat{\theta}\right)$ defined in (3.20) is not readily available because implementation of EM does not require calculation of the derivatives of $l(\theta|Y_{obs})$, which are often complicated and tedious to derive.

In a comment on the original EM paper, Smith (1977) noted that for a scalar parameter the iterations of EM provide a good estimate of λ. From (3.22) it is apparent that after a large number of iterations λ will be well approximated by $\varepsilon^{(t+1)}/\varepsilon^{(t)}$, or equivalently, by $\hat{\lambda}^{(t)} = \left(\theta^{(t+1)} - \theta^{(t)}\right)/\left(\theta^{(t)} - \theta^{(t-1)}\right)$, because

$$\hat{\lambda}^{(t)} = \frac{\theta^{(t+1)} - \theta^{(t)}}{\theta^{(t)} - \theta^{(t-1)}} = \frac{\varepsilon^{(t+1)} - \varepsilon^{(t)}}{\varepsilon^{(t)} - \varepsilon^{(t-1)}} \approx \frac{\lambda\varepsilon^{(t)} - \varepsilon^{(t)}}{\varepsilon^{(t)} - \lambda^{-1}\varepsilon^{(t)}} = \lambda .$$

This estimate of λ may be used to obtain a large-sample standard error for $\hat{\theta}$, because (3.19) and (3.21) imply that the asymptotic variance of $\hat{\theta}$ is

$$I_0^{-1}\left(\hat{\theta}\right) = I_c^{-1}\left(\hat{\theta}\right)/(1-\lambda).$$

For exponential families, $I_c^{-1}\left(\hat{\theta}\right)$ is the complete-data asymptotic variance calculated from the expected sufficient statistics obtained at the last E-step. In other words, a standard error for $\hat{\theta}$ can be obtained simply by inflating the complete-data standard error by $\sqrt{1-\lambda}$, where λ is estimated from the steps of EM.

In most real-data applications, of course, θ is multidimensional, and obtaining a covariance matrix for $\hat{\theta}$ is less straightforward. A general numerical procedure for approximating given by Meng and Rubin (1991a), is called Supplemented EM or SEM. From (3.19) and (3.21) it can be shown that

$$I_0^{-1}\left(\hat{\theta}\right) = I_c^{-1}\left(\hat{\theta}\right) + I_c^{-1}\left(\hat{\theta}\right)(I-D)^{-1}D. \tag{3.25}$$

Meng and Rubin show how the elements of $D = M'\left(\hat{\theta}\right)$ can be estimated by repeated runs of a forced EM in which all but one of the individual elements of θ are fixed at their MLEs. The procedure corresponds to numerical differentiation with step sizes determined by EM. For a *k* dimensional parameter one needs to perform *k* runs of forced EM, one to estimate each row of *D*. Using (3.25), the numerical estimate of *D* is then combined with the complete-data asymptotic covariance matrix $I_c^{-1}\left(\hat{\theta}\right)$ to produce $I_0^{-1}\left(\hat{\theta}\right)$. Implementation of SEM thus requires only the code for computing an asymptotic covariance matrix from complete data and the code for the EM algorithm itself.

An asymptotic covariance matrix for $\hat{\theta}$ is typically used in conjunction with the large-sample normal approximation

$$\left(\hat{\theta} - \theta\right) \sim N\left(0, I_0^{-1}\left(\hat{\theta}\right)\right), \qquad (3.26)$$

which can be justified from either a frequentist or a Bayesian perspective (e.g. Cox and Hinkley, 1974). This approximation is expected to work well when the sample size is sufficiently large that $l(\theta|Y_{obs})$ resembles a quadratic function in the vicinity of d. In practice the validity of (3.26) often depends on the scale of the parameter, and transformations may need to be applied to one or more components of θ to make the approximation more accurate. Transformations to improve normality will alter the form of the complete-data covariance matrix $I_C^{-1}\left(\hat{\theta}\right)$. When $l(\theta|Y_{obs})$ has unusual features such as multiple modes, ridges or suprema on a boundary, then the value of an asymptotic covariance matrix is dubious at best, and (3.26) should not be used for making inferential statements about θ.

For many of the multivariate models and data examples in this book, the potentially large number of parameters makes the implementation of SEM computationally prohibitive. In some cases the validity of the normal approximation (3.26) will be suspect as well, even on a carefully chosen scale for θ. For this reason, we will adopt simulation rather than asymptotic approximation as the primary method of inference. In multiparameter problems, simulation is often feasible even when the dimension of θ is very large. Moreover, simulation-based inferences can be made about any transformation or function of θ with no special analytic work involved.

Elementwise rates of convergence

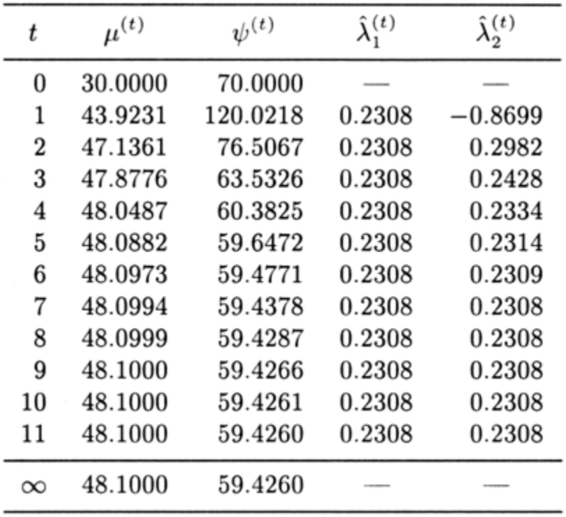
Apart from obtaining asymptotic standard errors, it may still be useful to examine rates of convergence corresponding to the individual elements of $\theta = (\theta_1, \theta_2, , \theta_k)$. These rates may be estimated from the iterations of EM by

$$\hat{\lambda}_j^{(t)} = \frac{\theta_j^{(t+1)} - \theta_j^{(t)}}{\theta_j^{(t)} - \theta_j^{(t-1)}} \tag{3.27}$$

for j-1,2,...,k at suitably large values of t. Unlike the eigenvalues of D, which are the fractions of missing information corresponding to the eigenvectors, these elementwise rates pertain to $u_1,u_2,...,u_k$, where u_j is a unit vector with a one in position j and zeroes elsewhere. As noted by Meng and Rubin (1991a), in most cases (3.27) will estimate the largest eigenvalue of D, because uj will have a nonzero component corresponding to the first eigenvector. If u_j happens to be precisely orthogonal to the first s eigenvectors, then (3.27) will converge to the (s+1) st largest eigenvalue of D. Consequently, the elementwise rates of convergence typically provide the first and perhaps a few additional eigenvalues of D, which can be a useful diagnostic for assessing how much information about θ is contained in $P(Y_{mis}|Y_{obs},\theta)$ relative to Y_{obs}.

For the EM example in Table 3.1, estimates $\hat{\lambda}_1^{(t)}$ and $\hat{\lambda}_2^{(t)}$ of the elementwise rates of convergence corresponding to the mean μ and the variance ψ, respectively, are displayed in Table 3.2. As previously shown, the eigenvalues of D are both equal to n_0/n=3/13=0.2308, and the elementwise rates converge to this number quite rapidly. Very close to the mode, successive values of θ are nearly identical and computation of (3.27) becomes numerically unstable. It is generally wise to compute (3.27) using double precision arithmetic and to estimate the rates from the last few iterations before numerical instability becomes evident. Note that in a multiparameter problem these elementwise rates alone are not sufficient

Table 3.5. *Iterations of EM for incomplete univariate normal data with estimated elementwise rates of convergence*



t	$\mu^{(t)}$	$\psi^{(t)}$	$\hat{\lambda}_1^{(t)}$	$\hat{\lambda}_2^{(t)}$
0	30.0000	70.0000	—	—
1	43.9231	120.0218	0.2308	−0.8699
2	47.1361	76.5067	0.2308	0.2982
3	47.8776	63.5326	0.2308	0.2428
4	48.0487	60.3825	0.2308	0.2334
5	48.0882	59.6472	0.2308	0.2314
6	48.0973	59.4771	0.2308	0.2309
7	48.0994	59.4378	0.2308	0.2308
8	48.0999	59.4287	0.2308	0.2308
9	48.1000	59.4266	0.2308	0.2308
10	48.1000	59.4261	0.2308	0.2308
11	48.1000	59.4260	0.2308	0.2308
∞	48.1000	59.4260	—	—

to obtain standard errors for the individual elements of $\hat{\theta}$. As seen from (3.25), the variance of a single element of $\hat{\theta}$ generally depends on the entire D matrix, whereas the elementwise rates provide at most only a few eigenvalues of D.

In some problems one or more components of θ may have no missing information at all. In the bivariate normal data depicted in Figure 2.2, for example, there are no missing observations of Y_1 and hence μ_1 and σ_{11} all have no missing information. An EM algorithm for these data would converge to the ML estimates for μ_1 and σ_{11} all in a single step from any starting value. When one or more components of θ converge immediately, the elementwise rates of convergence from the remaining components still estimate the largest fractions of missing information.

Accelerating convergence

The linear behavior of EM near the mode suggests some potentially useful methods for accelerating convergence. Rearranging (3.16), we can obtain an estimate of 0 in terms of two successive iterates $\theta^{(t)}$ and $\theta^{(t+1)}$,

$$\tilde{\theta}^{(t+1)} = \theta^{(t)} + (I - D)^{-1}\left(\theta^{(t-1)} - \theta^{(t)}\right), \tag{3.28}$$

which is typically closer to the mode than $\theta^{(t+1)}$. This technique, commonly known as *Aitken acceleration*, can make a linearly convergent algorithm like EM almost superlinear. When the individual components of θ appear to be converging at the same elementwise rate, (3.24) suggests that

$$\tilde{\theta}^{(t+1)} = \theta^{(t)} + \left(1 - \lambda_1\right)^{-1}\left(\theta^{(t-1)} - \theta^{(t)}\right) \tag{3.29}$$

may also work well, where λ_1 is the largest eigenvalue of D. These acceleration techniques require an estimate of D or at least its largest eigenvalue, which can be obtained by analytic methods or from the iterations of EM. The use of Aitken-type acceleration methods for EM have been investigated by Louis (1982); Laird, Lange and Stram (1987); and Lansky and Casella (1990). Another technique, proposed by Belin and Diffendal (1991), is to estimate the jth component of θ by

$$\tilde{\theta}_j^{(t+1)} = \theta_j^{(t)} + \left(1 - \hat{\lambda}_j\right)^{-1}\left(\theta_j^{(t-1)} - \theta_j^{(t)}\right) \tag{3.30}$$

for $j = 1,2,\ldots,k$, where $\hat{\lambda}_j$ is the estimated elementwise rate of convergence for θ_j given by (3.27). This third method, which may be regarded as intermediate between (3.28) and (3.29), is easier to compute than (3.28) and more appropriate than (3.29) in situations where the elements of θ appear to be converging at different rates.

Care should be taken in the use of these accelerated versions of EM, as they are not guaranteed to increase the loglikelihood at each step. Acceleration should not be employed until $\theta^{(t)}$ is close enough to the mode for (a) the steps of EM to be approximately linear, and (b) the estimated fractions of missing information to be stable. As previously mentioned, slow convergence of EM in an incomplete-data problem should be taken as a warning that certain aspects of θ are being estimated primarily from $P(Y_{mis}|Y_{obs},\theta)$ rather than Y_{obs}. In such problems it is sometimes more reasonable to bail out of the current analysis and fit a simpler model, rather than to continue iterating with a model whose parameters are poorly estimated.

Convergence and prior information

When EM is being used to find a posterior mode rather than an ML estimate, the missing information principle described in Section 3.3.2 applies but in a slightly modified form. The decomposition of (3.19) becomes

$$I_c(\theta) = I_o(\theta) + I_m(\theta) + I_\pi(\theta),$$

where $I_c(\theta)$, $I_o(\theta)$ and $I_m(\theta)$ are defined as above, and the additional term $I_\pi(\theta)$ is the information contained in the prior distribution,

$$I_\pi(\theta) = -\frac{\partial^2}{\partial\theta^2}\pi(\theta).$$

This term will be small when π is relatively flat and large when π is sharply peaked. The basic relationship (3.21) between these information quantities and the rate matrix $D = M'\left(\hat{\theta}\right)$ still applies,

$$D = I_c^{-1}\left(\hat{\theta}\right)I_m\left(\hat{\theta}\right),$$

but the complete information matrix $I_c\left(\hat{\theta}\right)$ now includes prior information. The introduction of a prior may thus be expected to reduce the magnitude of D and accelerate the convergence of EM in most cases. In particular, this will be true when the prior introduces substantial information about those aspects of θ that are most poorly estimated, those that influence the largest eigenvalue of D.

Convergence properties of ECM

The ECM algorithm introduced in Section 3.2.5 shares many of the convergence properties of EM. Like EM, it increases the loglikelihood at each step and converges reliably to a local maximum or (rarely) a saddlepoint of the loglikelihood. Like EM, it also exhibits linear convergence in the vicinity of the mode. ECM can be thought of as a combination of two linearly convergent algorithms: an EM algorithm, which pertains to the incomplete-data aspects of the problem, and a CM or conditional maximization algorithm, which pertains to the maximization of the likelihood in the complete-data case. As pointed out by Meng and Rubin (1992a), there seems to be little advantage to replacing the linearly convergent CM step with one or more steps of a superlinear technique such as Newton-Raphson, because the overall convergence of the combined algorithm will still be linear. Moreover, unless the superlinear algorithm is run to full convergence at each M-step, the loglikelihood would not be guaranteed to increase at each iteration.

With ECM, the global and elementwise rates of convergence cannot immediately be interpreted as fractions of missing information, because the simple identity $D=I_c^{-1}\left(\hat{\theta}\right)I_m\left(\hat{\theta}\right)$ does not generally hold. Basic results on ECM s rate of convergence, including relationships between the D matrix and information quantities, have been established by Meng (1994). Some of these results are counterintuitive; for instance, examples can be constructed where ECM converges more quickly than EM. A numerical method for obtaining large-sample covariance matrices from ECM, called

Supplemented ECM or SECM, is described by Meng and Rubin (1992a).

3.4 Markov chain Monte Carlo

Markov chain Monte Carlo is a collection of techniques for creating pseudorandom draws from probability distributions. In recent years it has been a subject of intense interest among statisticians, spawning a wide range of applications as well as a great deal of innovative theoretical work. In a broad sense, the goal of Markov chain Monte Carlo is to generate one or more values of a random variable Z, which is typically multidimensional. Let $P(Z)=f(Z)$ denote the density of Z, which we call the target distribution. Rather than attempting to draw from f directly, we generate a sequence $\{Z^{(1)}, Z^{(2)},\ldots,Z^{(t)},\ldots\}$ where each variate in the sequence depends in some fashion on the preceding ones, and where the stationary distribution (i.e. the limiting marginal distribution of $Z^{(t)}$ as $t \rightarrow \infty$) is the target f. For a t sufficiently large, $Z^{(t)}$ is approximately a random draw from f. Markov chain Monte Carlo is attractive when f is difficult to draw from directly, but drawing each variate in the sequence is straightforward.

Markov chain Monte Carlo methods have often been classified under Bayesian computation or Bayesian posterior simulation because many of the best known current applications have a strong Bayesian flavor; when viewed strictly as simulation methods, however, there is nothing inherently Bayesian about them. Also, despite the popularity of the term Markov chain Monte Carlo, depending on how the methods are viewed some of them are not strictly Markovian. In this new and rapidly evolving field, the lack of well defined and broadly accepted terminology has sometimes been a source of confusion. The reader should understand that names given to the methods below, and the definitions of these methods, are not universally accepted and may differ somewhat from what other authors have written.

This list of Markov chain Monte Carlo methods is not meant to be exhaustive, but concentrates on some that have proven most useful in the analysis of incomplete multivariate data. Presentations in a more general setting and additional references are given by Gelfand and Smith (1990); the articles by Gelman and Rubin (1992a), Geyer (1992) and Smith and Roberts (1993) with accompanying discussions; and Tierney (1994). Applications of Markov chain Monte Carlo are discussed by Gelfand *et al.* (1990); Casella and George (1992); Smith and Roberts (1993); and Gilks *et al.* (1993), among others. A comprehensive overview including theory and applications appears in the books by Tanner (1993) and Gilks, Richardson, and Spiegelhalter (1996).

3.4.1 Gibbs sampling

Gibbs sampling is the most popular and well known form of Markov chain Monte Carlo. Suppose that a random vector *Z* is partitioned into *J* subvectors,

$$Z^{(t)} = \left(Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_J^{(t)}\right),$$

Let *P*(*Z*) denote the joint distribution of *Z*, which is also the target distribution to be simulated. In Gibbs sampling, we iteratively draw from the conditional distribution of each subvector given all the others. Given the value of *Z* at step *t*, say

$$Z^{(t)} = Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_J^{(t)},$$

the value of *Z* at step *t*+1,

$$Z^{(t+1)} = \left(Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_J^{(t+1)}\right),$$

is obtained by successively drawing from the distributions

$$P\left(Z_J \mid Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_{J-1}^{(t+1)}\right)$$

$$\begin{array}{lcl} Z_1^{(t+1)} & \sim & P\left(Z_1 \mid Z_2^{(t)}, Z_3^{(t)}, \ldots, Z_J^{(t)}\right) \\ Z_2^{(t+1)} & \sim & P\left(Z_2 \mid Z_1^{(t+1)}, Z_3^{(t)}, \ldots, Z_J^{(t)}\right) \\ & \vdots & \\ Z_J^{(t+1)} & \sim & P\left(Z_J \mid Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_{J-1}^{(t+1)}\right) \end{array} \tag{3.31}$$

in a slight abuse of notation. In other words, we draw from the conditional distributions of Z$_1$, Z$_2$, up to Z$_J$, conditioning each time on the most recently drawn values of all other subvectors.

After the full set of subvectors has been drawn, we repeat the whole process to obtain $Z^{(t+2)}$, $Z^{(t+3)}$ and so on. The sequence $\{Z^{(t)}: t=0,1,2,...\}$ forms a Markov chain which, under mild regularity conditions, has a stationary distribution equal to *P*(*Z*); that is, $Z^{(t)} \rightarrow Z$ in distribution as $t \rightarrow \infty$.

The name *Gibbs* is not at all descriptive of this method, but actually refers to a class of probability distributions on lattice systems that have been used in problems of spatial analysis and statistical image reconstruction (Besag, 1974). The first use of Gibbs sampling in this context was made by Geman and Geman (1984), who provided a proof of convergence for a discrete *Z* with finite state space. The method was independently derived for *J*=2 by Li (1988), who presented an argument for convergence in the continuous case. Other convergence proofs under various conditions are given by Schervish and Carlin (1992); Liu, Wong, and Kong (1994, 1995); and Tierney (1994).

The regularity conditions necessary to establish convergence of the Gibbs sampler in a general setting are somewhat technical, but they do tend to be satisfied in most problems of practical interest. Informally, one can say that sufficient conditions Are (a) that the target distribution *P*(*Z*) must be a genuine probability distribution, and the sequence (3.31) must be the actual conditional distributions corresponding to this target; and (b) that the sample space of *Z* must be connected in the sense that it must be possible to reach any point in the sample space from any other point by repeated sampling from the conditionals in the manner of (3.31); periodicity and absorbing states are not allowed. For more discussion, see Roberts (1996) and Tierney (1996). For some examples of nonconvergence, see Casella and George (1992); Arnold (1993); and Section 3.5.2 below.

As pointed out by Liu, Wong and Kong (1995), the conditional distributions (3.31) in the Gibbs sampler need not be drawn from in any particular order in each iteration, nor do they need to be drawn from equally often. As long as each conditional distribution is visited infinitely often, the stationary distribution will be $P(Z)$. As a practical matter, of course, different visitation schemes will have different properties when only a finite number of iterations are performed. The distributions (3.31) are sometimes called the full conditionals because each one is the distribution of a subvector given all the other subvectors. Other sets of conditional distributions may also be grouped together to form sampling schemes that will converge to $P(Z)$, as described by Gelfand and Smith (1990).

3.4.2 Data augmentation

Closely related to Gibbs sampling is the *data augmentation* algorithm of Tanner and Wong (1987). Suppose that a random vector z is partitioned into two subvectors, $z = (u, v)$, where the joint distribution $P(z)$ is not easily simulated but the conditional distributions $P(u|v)=g(u|v)$ and $P(v|u)=h(v|u)$ are. At iteration t, let

$$\begin{aligned} Z^{(t)} &= \left(z_1^{(t)}, z_2^{(t)}, \ldots, z_m^{(t)}\right) \\ &= \left(\left(u_1^{(t)}, v_1^{(t)}\right), \left(u_2^{(t)}, v_2^{(t)}\right), \ldots, \left(u_m^{(t)}, v_m^{(t)}\right)\right) \end{aligned}$$

be a sample of size m from a distribution that approximates the target distribution $P(z)$. This sample is updated in two steps.

First,

$$U^{(t+1)} = \left(u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_m^{(t+1)}\right)$$

is created by drawing

$$u_i^{(t+1)} \sim g\left(u \mid v_i^{(t)}\right)$$

independently for $i = 1, 2, \ldots, m$. Next,

$$V^{(t+1)} = \left(v_1^{(t+1)}, v_2^{(t+1)}, \ldots, v_m^{(t+1)}\right)$$

is drawn as an iid sample from the equally weighted mixture of the conditionals $h\left(v \mid u_i^{(t+1)}\right)$,

$$\bar{h}\left(v \mid U_i^{(t+1)}\right) = \frac{1}{m}\sum_{i=1}^{m} h\left(v \mid u_i^{(t+1)}\right), \qquad (3.32)$$

which completes the new sample

$$Z^{(t+1)} = \left(\left(u_1^{(t+1)}, v_1^{(t+1)}\right), \ldots, \left(u_m^{(t+1)}, v_m^{(t+1)}\right)\right)$$

Using functional analysis, Tanner and Wong (1987) show that the distribution of $Z(t)$ converges to $P(z)$ as $t \to \infty$. This result does not require a large value of m; in particular, with $m = 1$ data augmentation reduces to a special case of the Gibbs sampler (3.31) with the random quantities $z = (u, v)$ partitioned into two subvectors, u and v. More generally, if we modify the second step of each iteration by sampling

$$v_i^{(t+1)} \sim h\left(v \mid u_i^{(t+1)}\right)$$

independently for $i = 1, 2, \ldots, m$ rather than drawing them from the mixture (3.32), then the algorithm becomes m independent, parallel runs of a Gibbs sampler. The mixing of the conditionals $h(v|u)$ at each iteration may not provide much practical benefit in speeding the convergence of the $Z^{(t)}$ but when m is large (3.32) provides a good analytic approximation to the marginal density $P(v)=\int P(u,v)\,du$ if such an approximation is desired.

Application to missing-data problems

The name *data augmentation* arose from applications of this algorithm to Bayesian inference with missing data. In many incomplete-data problems, the observed-data posterior $P(\theta|Y_{obs})$ is intractable and cannot easily be summarized or simulated; when Y_{obs} is augmented by an assumed value of the Y_{mis}, however, the resulting complete-data posterior $P(\theta|Y_{obs},Y_{mis})$ becomes much easier to handle. Consider the following iterative sampling scheme: given a current guess $\theta^{(t)}$ of the parameter, first draw a value of the missing data from the conditional predictive distribution of Y_{mis},

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right). \tag{3.33}$$

Then, conditioning on $Y_{mis}^{(t+1)}$, draw a new value of θ from its complete-data posterior,

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right). \tag{3.34}$$

Repeating (3.33)-(3.34) from a starting value $\theta^{(0)}$ yields a stochastic sequence $\left\{\left(\theta^{(t)}, Y_{mis}^{(t)}\right) : t = 1,2,\ldots\right\}$ whose stationary distribution is $P(\theta|Y_{mis}|Y_{obs})$, and the subsequences $\{\theta^{(t)} : t = 1,2,\ldots\}$ and $\left\{ Y_{mis}^{(t)}: t = 1,2,\ldots\right\}$ have $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$ as their misrespective stationary distributions. Following the terminology of Tanner and Wong (1987), we will refer to (3.33) as the Imputation or I-step and (3.34) as the Posterior or P-step, because (3-33) corresponds to imputing a value of the missing data Y_{mis} and (3.34) corresponds to drawing a value of θ from a complete-data posterior. For a value of t that is suitably large, we can regard $\theta^{(t)}$ as an approximate draw from $P(\theta|Y_{obs})$; alternatively, we can regard $Y_{mis}^{(t)}$ as an approximate draw from $P(Y_{mis}|Y_{obs})$.

Many particular examples of the algorithm (3.33)-(3-34) will appear throughout the remainder of this book. The first use of this algorithm seems to have been made by Li (1988) who presented an argument for convergence and used it to create imputations of Y_{mis}. in incomplete-data problems. The algorithm can be regarded either as a special case of data augmentation with m=1 or as a special case of Gibbs sampling with (Y_{mis},θ) partitioned into Y_{mis} and θ. Because the former name is more descriptive for incomplete-data problems, we will refer to it as data augmentation rather than Gibbs sampling. For the most part, however, we will use only the special case of data augmentation with m=1. On occasion we will perform m > 1 parallel runs of data augmentation, but we will keep the runs independent; that is, we will not employ mixing (3.32) at each iteration.

Data augmentation bears a strong resemblance to the EM algorithm. The E-step of EM calculates the expected complete-data sufficient statistics, whereas the I-step of data augmentation simulates a random draw of the complete-data sufficient statistics. The implementation of an I-step is typically very similar to that of an E-step, usually requiring only minor modifications of the computer code. The M-step of EM is a maximization of a complete-data likelihood, while the P-step of data augmentation is a random draw from a complete-data posterior. The computational requirements of EM and data augmentation are therefore quite similar, as both involve repeated application of complete-data methods to solve an incomplete-data problem.

3.4.3 Examples of data augmentation

Example 1: Incomplete univariate normal data. Suppose that Y=$(y_1, y_2,...,y_n)$ is an iid sample from a normal distribution with mean μ and variance ψ which, for the moment, is assumed to be known. If we apply a normal prior distribution to μ with mean μ_0 and variance τ, it follows that the posterior distribution of μ given Y is also normal with mean

$$E(\mu \mid Y) = \left(\frac{n\psi^{-1}}{n\psi^{-1} + \tau^{-1}} \right) \bar{y} + \left(\frac{\tau^{-1}}{n\psi^{-1} + \tau^{-1}} \right) \mu_0 \qquad (3.35)$$

and variance $V(\mu|Y) = (n\,\psi^{-1} + \tau^{--1})^{-1}$, where $\bar{y}$ is the sample mean of $y_1, y_2, \ldots, y_n$. Letting $\tau \to \infty$, the posterior becomes normal with mean $\bar{y}$ and variance $n^{-1}\psi$, which may also be obtained by applying Bayes' s formula with the improper diffuse prior $\pi(\mu) \propto$ c where c is a constant (e.g. Box and Tiao, 1992).

Now suppose that only the first n_1 elements of Y are observed and the remaining $n_0 = n - n_1$ are missing. Under ignorability and the diffuse prior $\pi(\mu) \propto c$ the observed-data posterior $P(\mu|Y_{obs})$ becomes normal with mean $\bar{y}_{obs} = n_1^{-1} \sum_{i=1}^{n_1} y_i$ and variance $n_1^{-1}\psi$. In this trivial example, values of μ from $P(\mu|Y_{obs})$ can be simulated directly using standard routines for generating normal random variates. We can also simulate them iteratively, however, using the data augmentation routine of (3.33)-(3.34). Given a current parameter value $\mu^{(t)}$, the I-step simulates $Y_{mis}^{(t+1)}$ by drawing

$$y_i^{(t+1)} \mid \mu^{(t)}, Y_{obs} \sim N\left(\mu^{(t)}, \psi\right) \qquad (3.36)$$

independently for $i = n_1 + 1, \ldots, n$. The P-step then proceeds to draw $\mu^{(t+1)}$ from the complete-data posterior $P\left(\mu \mid Y_{obs}, Y_{mis}^{(t+1)}\right)$, anormal distribution with mean

$$\bar{y}^{(t+1)} n^{-1} \left[\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n} y_i^{(t+1)} \right]$$

and variance $n^{-1}\psi$.

In this simple example of data augmentation, one may analytically verify that the distribution of $\mu^{(t)}$ approaches the correct observed-data posterior $N\left(\bar{y}_{obs}, n_1^{-1}\psi\right)$ as t → ∞. This is possible because of the following well known property of the normal distribution: If $U|V \sim N(V,a)$ and $V \sim N(b,c)$ then $U \sim N(b, a+c)$. Applying this property, the conditional distribution of $\mu^{(t)}$ given Y_{obs} and the previous iterate $\mu^{(t\text{-}1)}$ is easily seen to be

$$\mu^{(t)} \mid \mu^{(t-1)} \sim N\left(\bar{y}_{obs} + \lambda\left(\mu^{(t-1)} - \bar{y}_{obs}\right), n_1^{-1}\psi\left(1 - \lambda^2\right)\right),$$

where $\lambda = n_0/n$ and conditioning on Y_{obs} has been suppressed in the notation. Similarly, the conditional distribution of $\mu^{(t)}$ given Y_{obs} and $\mu^{(t\text{-}2)}$ is also normal, with mean

$$\begin{aligned} E\left(\mu^{(t)} \mid \mu^{(t-2)}\right) &= E\left(E\left(\mu^{(t)} \mid \mu^{(t-1)}\right) \mid \mu^{(t-2)}\right) \\ &= E\left(\bar{y}_{obs} + \lambda\left(\mu^{(t-1)} - \bar{y}_{obs}\right) \mid \mu^{(t-2)}\right) \\ &= \bar{y}_{obs} + \lambda^2\left(\mu^{(t-2)} - \bar{y}_{obs}\right) \end{aligned}$$

and variance

$$\begin{aligned}
V\left(\mu^{(t)} \mid \mu^{(t-2)}\right) &= E\left(V\left(\mu^{(t)} \mid \mu^{(t-1)}\right) \mid \mu^{(t-2)}\right) \\
&\quad + V\left(E\left(\mu^{(t)} \mid u^{(t-1)}\right) \mid \mu^{(t-2)}\right) \\
&= n_1^{-1}\psi\left(1-\lambda^2\right) \\
&\quad + V\left(\bar{y}_{obs} + \lambda\left(\mu^{(t-1)} - \bar{y}_{obs}\right) \mid \mu^{(t-2)}\right) \\
&= n_1^{-1}\psi\left(i-\lambda^4\right).
\end{aligned}$$

Repeating this argument t times gives the marginal distribution of $\mu^{(t)}$ in terms of the starting value $\mu^{(0)}$,

$$\mu^{(t)} \mid \mu^{(0)} N\left(\bar{y}_{obs} + \lambda^t\left(\mu^{(0)} - \bar{y}_{obs}\right), n_1^{-1}\psi\left(1-\lambda^{2t}\right)\right), \qquad (3.37)$$

which approaches $N\left(\bar{y}_{obs}, n_1^{-1}\psi\right)$ as $t \to \infty$ for any fixed $\mu^{(0)}$ as long as $\lambda < 1$. If the starting value $\mu^{(0)}$ is not fixed but drawn from a probability distribution, then we can also investigate the unconditional distribution of $\mu^{(t)}$. In particular, if $\mu^{(0)}$ is drawn from the correct posterior $N\left(\bar{y}_{obs}, n_1^{-1}\psi\right)$ then (3.37) implies that $\mu^{(t)}$ will be normal with mean

$$\begin{aligned}
E\left(\mu^{(t)}\right) &= E\left(E\left(\mu^{(t)} \mid \mu^{(0)}\right)\right) \\
&= E\left(\bar{y}_{obs} + \lambda^t\left(\mu^{(0)} - \bar{y}_{obs}\right)\right) \qquad (3.38) \\
&= \bar{y}_{obs}
\end{aligned}$$

and variance

$$
\begin{aligned}
V\left(\mu^{(t)}\right) &= E\left(V\left(\mu^{(t)} \mid \mu^{(0)}\right)\right) + V\left(E\left(\mu^{(t)}\mu^{(0)}\right)\right) \\
&= n_1^{-1}\psi\left(1-\lambda^{2t}\right) + V\left(\bar{y}_{obs} + \lambda^{t}\left(\mu^{(0)} - \bar{y}_{obs}\right)\right) \\
&= n_1^{-1}\psi,
\end{aligned}
$$

and stationarity is achieved immediately.

We can also perform data augmentation in this example when the variance ψ is unknown. Under the diffuse prior $\pi(\mu,\psi) \propto \psi$-^{1} the complete-data posterior is

$$
\begin{aligned}
\mu \mid \psi, Y &\sim N\left(\bar{y}, n^{-1}\psi\right) \\
\psi \mid Y &\sim \left(n-1\right)S^{2}\chi_{n-1}^{-2}
\end{aligned}
\tag{3.40}
$$

where S^2 is the sample variance of y_1, y_2,...,y_n, and the observed-data posterior is

$$
\begin{aligned}
\mu \mid \psi, Y_{obs} &\sim N\left(\bar{y}_{obs}, n^{-1}\psi\right) \\
\psi \mid Y_{obs} &\sim \left(n_1-1\right)S_{obs}^{2}\chi_{n_1-1}^{-2}
\end{aligned}
\tag{3.41}
$$

where S_{obs}^2 is the sample variance of y_1, $y2$,...,y_{n1}. The I-step of data augmentation simulates $Y_{mis}^{(t+1)}$ by drawing

$$
y_i^{(t+1)} \mid \mu^{(t)}, \psi^{(t)}, Y_{obs} \sim N\left(\mu^{(t)}, \psi^{(t)}\right)
$$

independently for $i = n_1+1,\ldots,n$, and the P-step simulates $\mu^{(t+1)}$ and $\psi^{(t+1)}$ from (3.40) with $Y_{mis}^{(t+1)}$ substituted for Y_{mis}.

In this two-parameter problem, writing down the marginal distribution of $\theta^{(t)} = (\mu^{(t)},\psi^{(t)})$ at any step t is no longer a simple matter. We can, however, demonstrate empirically that the marginal distribution of $\theta^{(t)}$ approaches the observed-data posterior (3.41) in numerical examples. The algorithm was applied to the univariate sample of size $n_1 = 10$ in Table 3.1 (a), arbitrarily taking $n_0 = 3$ and starting values $\psi^{(0)} = 70$. Simulated marginal densities of $\mu^{(t)}$ and $\psi^{(t)}$ for $t = 1,2,3$ are displayed in Figure 3.2. Based on the marginals, it appears that convergence to the observed-data posterior is quite rapid. The densities in Figure 3.2 were simulated by $m = 500$ parallel chains of data augmentation, each starting from $\mu^{(t)}$ and $\psi^{(t)}$. The chains were run independently; no mixing as in (3.32) was used. For plotting purposes, however, the marginal densities were estimated by the Rao-Blackwell method, averaging

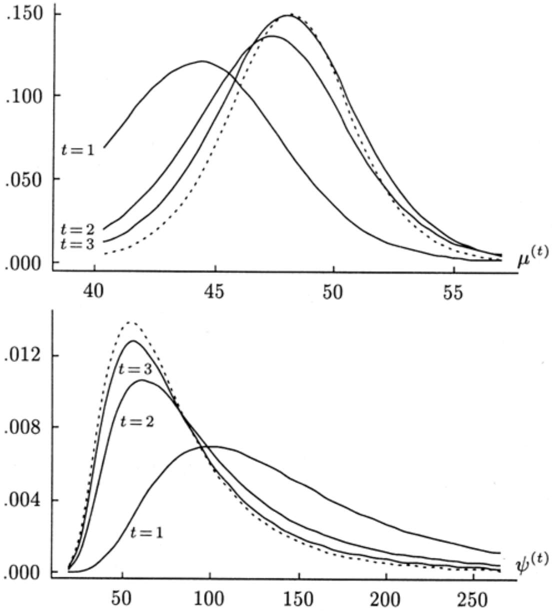


Figure 3.2. *Simulated marginal densities of $\mu^{(t)}$ and $\psi^{(t)}$ for t=1,2,3, with dotted lines showing the exact observed-data posteriors.*

formulas for the complete-data marginal posteriors over the 500 iterates of Y_{mis}. This and other techniques for extracting meaningful summaries from Markov chain Monte Carlo runs will be discussed in Chapter 4.

Example 2: Incomplete binary data. Let $Y=(y_1,y_2,...,y_n)$ represent the outcomes of n independent Bernoulli trials where $y_i = 1$ with probability θ and $y_i = 0$ with probability $1 - \theta$ for $i = 1, 2,..., n$. The beta prior distribution

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $\alpha > 0$, $\beta > 0$, denoted by $\theta \sim Beta\ (\alpha, \beta)$, leads to the complete-data posterior

$$\theta \mid Y \sim Beta\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n - \sum_{i=1}^{n} y_i\right).$$

For simplicity let us use the limiting form of this prior as $\alpha \rightarrow 0$ and $\beta \rightarrow 0$, so that the posterior becomes $\theta|Y \sim Beta\ (a,b)$ where $a = \sum_{i=1}^{n} y_i$ and $b = n-a$; this posterior is proper as long as Y contains at least one success and one failure. Now suppose that the first n_1 elements of Y are observed and the remaining $n_0 = n-n_1$ elements are missing. The observed-data posterior becomes $\theta|Y_{obs} \sim Beta\ (a_1, b_1)$ where $a_1 = \sum_{i=1}^{n_1} y_i$ and $b_1 = n_1 - a_1$, which is proper provided that $1 \leq a_1 \leq n_1 - 1$.

Applying data augmentation to this example, the I-step fills in the missing trial outcomes by letting $y_i^{(t+1)}=1$ with probability $\theta^{(t)}$ and 0 otherwise for $i = n_1+1,...,n$. The P-step then samples

$$\theta^{(t+1)} \mid Y_{obs}, Y_{mis}^{(t+1)} \sim Beta\left(a^{(t+1)}, b^{(t+1)}\right),$$

where

$$a^{(t+1)} = a_1 + a_0^{(t+1)} = a_1 + \sum\nolimits_{i=n_1}^{n} +1 y_i^{(t+1)}$$

$$b^{(t+1)} = b_1 + b_0^{(t+1)} = b_1 + n_0 - a_0^{(t+1)}.$$

If data augmentation works properly, then the marginal distribution of $\theta^{(t)}$ as t → ∞ should approach Beta (a_1, b_1) which has mean a_1/n_1. Also, for large t, the distribution of any imputed trial should be Bernoulli with marginal probability of success

$$E\left(y_i^{(t+1)}\right) = E\left(\theta^{(t)}\right) = a_1 / n_1,$$

where conditioning on Y_{obs} is assumed and has been suppressed in the notation. We can algebraically verify that $E\left(y_i^{(t)}\right)$ does in fact approach a_1/n_1 as t → ∞, because

$$\begin{aligned} E\left(y_i^{(t+1)} \mid Y_{mis}^{(t)}\right) &= E\left(\theta^{(t)} \mid Y_{mis}^{(t)}\right) \\ &= \left(a_1 + a_0^{(t)}\right) / n \\ &= a_1 / n_1 - \lambda\left(a_0^{(t)} / n_0 - a_1 / n_1\right) \end{aligned}$$

where $\lambda = n_0/n$. Similarly,

$$E\left(y_i^{(t+1)} \mid Y_{mis}^{(1)}\right) = a_1 / n_1 - \lambda^t\left(a_0^{(1)} / n_0 - a_1 / n_1\right)$$

which implies that

$$E\left(y_i^{(t+1)} \mid \theta^{(0)}\right) = a_1 / n_1 - \lambda^t\left(\mu^{(0)} - a_1 / n_1\right).$$

Clearly, (3.42) approaches a_1/n_1 from any starting value $\mu^{(o)}$ as long as $\lambda < 1$. If $\mu^{(o)}$ happens to be equal to a_1/n_1, or is drawn from a probability distribution with mean a_1/n_1, then convergence is immediate. This example was first used by Li (1988).

3.4.4 The Metropolis-Hastings algorithm

An older method of Markov chain Monte Carlo is the algorithm of Metropolis *et al.* (1953) and its generalization by Hastings (1970). The Metropolis-Hastings algorithm will not be needed in the remainder of this book; we briefly mention it, however, because of its usefulness in extending the basic algorithms of Chapters 5-9 to more complicated modeling situations.

In the Hastings version, a Markov chain $\{Z^{(t)}:t=0,1,2,...\}$ with stationary distribution $P(Z)=f(Z)$ is constructed as follows. Given $Z^{(t)}$, a candidate value $\tilde{Z}$ is drawn from a transition distribution $g(Z|Z^{(t)})$. Then the ratio

$$R^{(t+1)} = \frac{g\left(Z^{(t)}\tilde{Z}\right)}{g\left(Z \mid Z^{(t)}\right)}\frac{f\left(\tilde{Z}\right)}{f\left(Z^{(t)}\right)}.$$

is calculated. It $R^{(t+1)}$ is greater-than 1, we accept the value of the candidate variable and set $Z^{(t+1)}=\tilde{Z}$. If $R^{(t+1)} < 1$, we randomly accept the value of $\tilde{Z}$ as our next iterate $Z^{(t+1)}$ with probability $R^{(t+1)}$, and otherwise keep the current value, $Z^{(t+1)}=Z^{(t)}$. If the transition distribution g allows the process to eventually reach any state in the support of Z, then $Z^{(t)} \rightarrow Z$ in distribution as $t \rightarrow \infty$.

Metropolis-Hastings is useful when a transition distribution g can be found that (a) is easy to simulate, and (b) leads to acceptance ratios (3.43) that are easy to calculate. Because the target density f enters into the algorithm only through the acceptance ratio, we need only to be able to evaluate f up to a constant of proportionality. This makes Metropolis-Hastings attractive for simulation of Bayesian posterior distributions, for which the densities are typically known only up to a normalizing constant.

From a standpoint of efficiency, it is advantageous for the acceptance ratios to be close to one over the region where $f(Z)$ is appreciable, which occurs when $g(Z|Z^{(t)})$ is a good approximation to $f(Z)$. When $g(Z|Z^{(t)})=f(Z)$, then the acceptance ratio is always one and convergence is immediate. In practical applications, it is wise to choose a g that is somewhat more diffuse (i.e. having heavier tails) than f; otherwise, there may be little opportunity for a candidate value $\tilde{Z}$ to fall in some regions of the sample space where $f(Z)$ is appreciable, and convergence to the target distribution may be too slow for practical use. If g is too diffuse, however, than many of the candidate values will fall outside the range where $f(Z)$ is appreciable, in which case the rejection rate will be very high and convergence will again be slow.

3.4.5 Generalizations and hybrid algorithms

As noted by several authors (Gelman, 1992; Smith and Roberts, 1993; Middleton, 1993; Tierney, 1994), both the Gibbs sampler and the Metropolis-Hastings algorithm may be generalized in a variety of ways. Consider the Gibbs sampler for a random vector $Z=(Z_1,Z_2,...,Z_J)$. In one iteration of ordinary Gibbs, we sample from the full conditionals

$$Z_j^{(t+1)} \sim P\left(Z_j \mid Z_1^{(t+1)},...,Z_{j-1}^{(t+1)},...,Z_J^{(t)}\right) \tag{3.44}$$

for $j=1,2,...,J$. In practice, however, it is not necessary to generate $Z_j^{(t+1)}$ directly from (3.44), but only from the transition distribution of a Markov chain that has (3.44) as its stationary distribution. In other words, if a Markov chain Monte Carlo scheme can be found that would eventually converge to the local distribution (3.44), we need only to perform one or more cycles of this local algorithm instead of (3.44), and the stationary distribution of the Gibbs sampler will be preserved. In specific examples, it sometimes happens that one (or more) of the full conditional distributions is difficult to simulate directly, but another Gibbs sampler or Metropolis-Hastings algorithm can be found that converges to the desired conditional. By replacing the difficult conditional with one or more iterations of this sampling scheme, we obtain a hybrid algorithm that still converges to the proper target.

Another potential use of these generalized algorithms is in data augmentation with an inconvenient prior. Suppose that a prior distribution $\pi^*(\theta)$ exists that leads to a tractable complete-data posterior $P^*(\theta|Y_{obs},Y_{mis})$, but the prior that we would like to use for inference is $\pi(\theta)$ which leads to a posterior $P(\theta|Y_{obs},Y_{mis})$ that is intractable. In this situation, we can replace the P-step under $\pi(\theta)$ with one or more steps of a Metropolis-Hastings algorithm that draws a candidate value $\tilde{\theta}$ from $P^*(\theta|Y_{obs},Y_{mis})$. The acceptance ratio becomes

$$\begin{aligned} R^{(t+1)} &= \frac{P^*\left(\theta^{(t)} \mid Y_{obs}, Y_{mis}\right)}{P^*\left(\tilde{\theta} \mid Y_{obs}, Y_{mis}\right)} \frac{P^*\left(\tilde{\theta} \mid Y_{obs}, Y_{mis}\right)}{P^*\left(\theta^{(t)} \mid Y_{obs}, Y_{mis}\right)} \\ &= \frac{\pi\left(\tilde{\theta}\right)/\pi^*\left(\tilde{\theta}\right)}{\pi\left(\theta^{(t)}\right)/\pi^*\left(\theta^{(t)}\right)}, \end{aligned} \qquad (3.45)$$

which does not depend on Y_{obs}, or Y_{mis}. When Metropolis-Hastings is used within data augmentation in this manner, the result is a hybrid algorithm with stationary distribution equal to the correct observed-data posterior under the desired prior. Note that (3.45) requires the evaluation of π and π^* only up to a constant of proportionality, so the acceptance ratios are typically easy to calculate.

3.5 Properties of Markov chain Monte Carlo

3.5.1 The meaning of convergence

Below we discuss some basic properties of Markov chain Monte Carlo, with special emphasis on the data augmentation scheme of (3.33)-(3.34). Unlike optimization methods like EM, which are deterministic and converge to a point in the parameter space, Markov chain Monte Carlo algorithms are stochastic and converge to probability distributions. Yet certain important similarities exist between the convergence behavior of EM and data augmentation.

Assuming the conditions needed for convergence are satisfied, the output of data augmentation is a sequence $\left\{\left(\theta^{(t)}, Y_{mis}^{(t)}\right): t = 0, 1, 2, \ldots\right\}$ with stationary distribution $P(\theta, Y_{obs}, Y_{mis})$ For the sequence to have converged, it is sufficient for the distribution of $\theta^{(t)}$ to have converged to $P(\theta|Y_{obs})$, because $\theta^{(t)} \sim P(\theta|Y_{obs})$ implies that $\left(\theta^{(t+s)}, Y_{mis}^{(t+s)}\right) \sim P(\theta, Y_{obs}|Y_{mis})$ for all $s > 0$. Equivalently, it is sufficient for the distribution of $Y_{mis}^{(t)}$ to have converged to $P(Y_{mis}|Y_{obs})$. Also, convergence by t iterations means that $\theta^{(s)}$ and $Y_{mis}^{(s)}$ are independent of $\theta^{(s+t)}$ and $Y_{mis}^{(s+t)}$. In applications it is typically more convenient to monitor convergence through the behavior of successive values of θ than successive values of Y_{mis} because the latter is usually of higher dimension. Except in trivial examples like those in Section 3.4.3 for which data augmentation is not needed, summaries of $P(\theta|Y_{obs})$ are not available in closed form, making it difficult to know precisely when convergence has occurred. Techniques for assessing convergence are described in Chapter 4. For now we will discuss only in broad terms some issues surrounding convergence.

3.5.2 Examples of nonconvergence

Nonexistence of a stationary distribution

As mentioned above, convergence of a Gibbs sampler requires that the full conditionals (3.31) are the conditionals of a genuine joint probability distribution. It is possible to construct simple examples in which a set of proper conditional distributions does not define a proper joint distribution (Casella and George, 1992). For data augmentation, Bayes s Theorem guarantees a proper limiting distribution as long as the prior $\pi(\theta)$ is proper. In many real data applications of Bayesian analysis, however, it is convenient to use so-called noninformative priors that are actually improper but lead to proper posteriors when Bayes s formula is applied. Even when an improper π is known to yield a proper posterior in the case of complete data, it may not necessarily do so when some data are missing.

For a very simple example, let $Y = (y_1, y_2)$ represent two independent observations from $N(\mu, \psi)$ with μ and ψ both unknown. Under the standard noninformative prior $\pi(\mu, \psi) \propto \psi^{-1}$, the posterior distribution is given by (3.40) with $n = 2, \bar{y} = (y_1 + y_2)/2$ and $S^2 = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2$. Now suppose that only $y_1 = Y_{obs}$ is observed and $y_2 = Y_{mis}$ is missing. Applying Bayes ' s formula, the observed-data posterior becomes

$$P(\mu, \psi \mid Y_{obs}) \propto L(\mu, \psi \mid Y_{obs}) \pi(\mu, \psi)$$
$$\propto \psi^{-3/2} \exp\left\{ -\frac{(y_1 - \mu)^2}{2\psi}, \right\}$$

which is not a proper probability distribution because the integral

$$\int_0^\infty \int_{-\infty}^\infty \psi^{-3/2} \exp\left\{ -\frac{(y_1 - \mu)^2}{2\psi} \right\} d\mu d\psi = \int_0^\infty (2\pi)^{-1/2} \psi^{-2} d\psi$$

does not exist. Yet, one could naively apply data augmentation to this example under the improper prior. The I-step would be

$$y_2^{(t+1)} \sim N\left(\mu^{(t)}, \psi^{(t)}\right), \tag{3.46}$$

and the P-step would be

$$\psi^{(t+1)} \mid y_1, y_2^{(t+1)} \sim \left(S^2\right)^{(t+1)} \chi_1^{-2} \tag{3.47}$$

$$\mu^{(t+1)} \mid \psi^{(t+1)}, y_1 y_2^{(t+1)} \sim N\left(\bar{y}^{(t+1)}, \psi^{(t+1)} / 2\right),$$

where

$$\bar{y}^{(t)} = \left(y_1 + y_2^{(t)}\right) / 2$$

$$\left(S^2\right)^{(t)} = \left(y_1 - \bar{y}^{(t)}\right)^2 + \left(y_2^{(t)} - \bar{y}^{(t)}\right)^2.$$

Notice that even though the joint posterior does not exist, both the I-step and the P-step are defined at every iteration. One could naively alternate between (3.46) and (3.47) indefinitely without any clue that the algorithm is not converging.

The fundamental reason why data augmentation fails here is that the mean and variance cannot be jointly estimated on the basis of a single observation y_1. The observed data provide no information on one aspect of $\theta = (\mu, \psi)$, namely ψ, so that unless a proper prior distribution is applied to ψ there is no basis for inference. Although there is no compelling reason to use data augmentation in this trivial example, it is not difficult to construct more realistic multivariate problems where both the I- and P-steps of data augmentation are defined but the stationary distribution does not exist. These would be similar to the examples of Section 3.3.1 where the ML estimate of θ is not unique because the likelihood is maximized over a ridge. Whenever the likelihood has a ridge, certain aspects of the parameter are inestimable, and unless a proper prior is applied to those aspects of θ the posterior will not be proper. Sparse datasets with few observations or high rates of missingness may be prone to these problems. If we suspect that data augmentation may not be converging to a proper posterior, we can switch to a proper prior, thereby guaranteeing that the posterior will be proper as well.

Boundary values and absorbing states

Another basic requirement for the convergence of a Markov chain Monte Carlo algorithm is that the support of the target distribution must be connected in the sense that it must be possible to eventually reach any state from any other state. There must be no periodic states and no absorbing states, i.e. regions where the algorithm could become trapped with zero probability of escape. In the stochastic processes literature, this property is known as ergodicity.

Consider again the normal sample $Y = (y_1, y_2)$ where y_1 is observed and y_2 is missing, but now let us suppose that the population mean μ is known, and without loss of generality take $\mu = 0$. Suppose that we apply an improper prior distribution to the variance, $\pi(\psi) \propto \psi^{-(v+2)/2}$ where v is a fixed constant. Given $Y = (y_1, y_2)$, the complete-data posterior is

$$P(\psi \mid Y) \propto \psi^{-1} \exp\left\{ -\frac{\left(y_1^2 + y_2^2\right)}{2\psi} \right\} \psi^{-(v+2)/2}$$

or $\psi|Y \sim \left(y_1^2 + y_2^2\right)\chi^{-2}_{v+2}$, which is proper provided that v > −2. Given only $Y_{obs} = y_1$, the observed-data posterior is

$$P\left(\psi \mid Y_{obs}\right) \propto \psi^{-1/2} \exp\left\{ -\frac{y_1^2}{2\psi} \right\} \psi^{-(v+2)/2} \tag{3.48}$$

or $\psi|Y_{obs} \sim y_1^2 \chi^{-2}_{v+1}$, which is proper for any v >– 1 If data augmentation were applied to this example, the I-step would be

$$y_2^{(t+1)} \mid \psi^{(t)}, y_1 \sim N\left(0, \psi^{(t)}\right),$$

and the P-step would be

$$\psi^{(t+1)} \mid y_1 y_2^{(t+1)} \sim \left(y_1^2 + \left(y_2^{(t+1)} \right)^2 \right) \chi_{\nu+2}^{-2}.$$

The algorithm would proceed normally except in the unlikely event that $y_1 = 0$ happened to become zero at some iteration. If that were to occur, we would obtain $y_2^{(t)} = 0$ and $\psi^{(t)}$ for every iteration thereafter. In other words, $\psi = 0$ would be an absorbing state.

Even if y_1 happened to be zero, absorption would be unlikely because unless we start on the boundary ($\psi(0) = 0$) the event $\psi^{(t)}=0$ occurs in theory with probability zero. Depending on the computer and random variate generator used, there could be a small chance of falling within machine precision of the boundary, especially if the starting value $\psi^{(0)}$ is very close to zero. The presence of an absorbing state is not the only difficulty in this example, because the observed-data posterior (3.48) is not proper for $y_1 = 0$. In Chapter 8, however, we will see that in some real categorical-data problems absorption onto a boundary becomes a distinct possibility even when the posterior is technically proper, and we will need to handle such situations with care.

3.5.3 Rates of convergence

Assuming that a Markov chain Monte Carlo algorithm does converge to a proper stationary distribution, it is important to consider how quickly this convergence occurs. Convergence rates are typically defined in terms of a distance measure between the marginal distribution of the iterates at any given time and the target distribution. Some interesting theoretical results on convergence rates and further references are given by Schervish and Carlin (1992); Smith and Roberts (1993); Tierney (1994); and Liu, Wong and Kong (1995). This work, although reassuring, does not easily translate into practical guidelines for knowing when convergence has occurred in specific examples. Ongoing research regarding convergence behavior will undoubtedly lead to greater understanding in the future; for now, however, we can informally state a few general principles that apply to incomplete-data problems.

Convergence and missing information

Consider simple data augmentation in which we alternately perform an I-step

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right)$$

and a P-step

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right).$$

Intuitively, the rate of convergence should depend on how much information about the parameter is contained in missing data relative to the observed data and the prior. The complete-data posterior may be written

$$P(\theta \mid Y_{obs}, Y_{mis}) \propto P(\theta \mid Y_{obs}) P(Y_{mis} \mid Y_{obs}, \theta). \qquad (3.49)$$

In the extreme case where Y_{mis} provides no information about θ beyond that already contained in Y_{obs}, Y_{mis} and θ would be conditionally independent given Y_{obs} the last term in (3.49) would then be constant with respect to θ, and convergence to the target distribution would be immediate. More generally, if $P(Y_{mis}|Y_{obs},\theta)$ as a function of θ is relatively flat over the region of high posterior density (which is typically equivalent to the missing information, as defined in Section 3.3.2, being near zero), then each P-step will be nearly a draw from $P(\theta|Y_{obs})$ and the algorithm will converge rapidly. On the other hand, if the missing information is a large portion of the total information, then θ will depend heavily on Y_{mis} at each P-step, which will in turn depend on the value of θ used in the previous I-step; successive iterates of θ will tend to be highly correlated and convergence will be slow. Just as with EM, the rate of convergence of data augmentation and the fractions of missing information are fundamentally related.

This relationship between missing information and rate of convergence is difficult to formalize in a general way, but it can be easily demonstrated in simple examples. Consider again the univariate normal data $Y = (y_1, y_2,\ldots,y_n)$ with known variance ψ and unknown mean μ. In Example 1 of Section 3.4.3 we investigated data augmentation under the noninformative prior $\pi(\mu) \propto c$ (a constant), where the first n_1 elements of Y are observed and the remaining $n_0 = n - n_1$ are missing. We found that the stationary distribution of μ is $\mu | Y_{obs} \sim N\left(\bar{y}_{obs}, n_1^{-1}\psi\right)$ and the marginal distribution of $\mu^{(t)}$ is

$$\mu^{(t)} \mid \mu^{(0)}, Y_{obs} \sim N\left(\bar{y}_{obs} + \lambda^t\left(\mu^{(0)} - \bar{y}_{obs}\right), n_1^{-1}\psi\left(1 - \lambda^{2t}\right)\right) \quad (3.50)$$

where $\lambda = n_0/n$ is the fraction of missing information. Clearly, the algorithm will approach stationarity rapidly for λ near zero and slowly for λ near one. This example can easily be generalized to an informative prior $\mu \sim N(\mu_0,\tau)$, in which case we would obtain an expression like (3.50) but with the following changes: $\bar{y}_{obs}$ and $n_1^{-1}\psi$ would be replaced by the new observed-data posterior mean and variance, respectively, and λ would be replaced by the new fraction of missing information

$$\lambda^* = \frac{n_0\psi^{-1}}{n\psi^{-1} + \tau^{-1}},$$

where the prior information τ^{--1} now appears in the denominator as a part of the total information.

When the mean and variance are both unknown, the joint distribution of $\mu^{(t)}$ and $\psi^{(t)}$ is intractable, but we can still demonstrate empirically that the rate of convergence depends on n_0/n. Using the $n_1 = 10$ observations in Table 3.1 (a) and the noninformative prior $\pi(\mu, \psi) \propto \psi^{-1}$, we performed data augmentation under various choices of n_0. Independent sample paths for $(\mu^{(t)}, \psi^{(t)})$ beginning from four different starting positions are displayed in Figure 3.3, first for $n_0 = 10$ and again for n_0=90, with ψ shown on a log scale. The starting values were all chosen to be in the tails of the observed-data posterior, so that the iterates would exhibit an initial trend as they wander into the region of high posterior density. For $n_0 = 10$ the sample paths become heavily intertwined by $t = 8$, suggesting that for most practical purposes the algorithm has probably converged by eight or ten iterations. For $n_0 = 90$, however, the sample paths have still not crossed one another by $t = 25$; the algorithm takes smaller steps and the successive iterates are-more highly correlated.

Starting values and starting distributions

In the univariate normal example with known mean, (3.50) reveals that convergence behavior depends not only on the fraction of missing information but also on the choice of a starting value. If we happened to take $\mu^{(0)} = \bar{y}_{obs}$, then the distribution of $\mu^{(t)}$ would have the correct mean (i.e. the same mean as the stationary distribution) for every t. Even though the variance of $\mu^{(t)}$ is always less than the stationary variance, choosing a starting value near the center of the observed-data posterior makes the first moment more nearly correct.

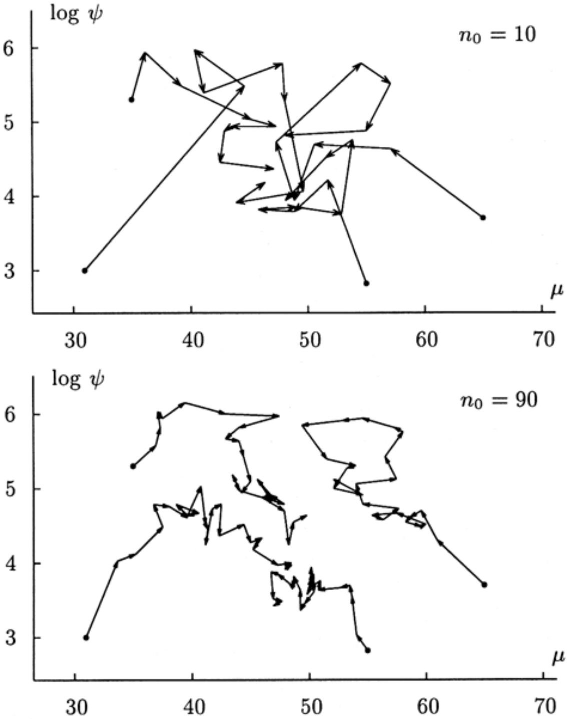


Figure 3.3. *Iterations of data augmentation form various starting values for univariate normal data with* n_0=10 *and* n_0=90.

It is also evident from (3.50) that the variance of $\mu^{(t)}$ does not depend on the starting value $\mu^{(0)}$ as long as $\mu^{(0)}$ is fixed. If we do not use a fixed $\mu^{(0)}$ but draw it from a probability distribution, however, then we can alter the second moment of $\mu^{(t)}$ as well as the first moment. Suppose that $\mu^{(0)}$ is drawn at random from a probability distribution with variance κ. Then (3.39) implies that the unconditional variance of $\mu^{(t)}$ is

$$V\left(\mu^{(t)}\right) = n_1^{-1}\psi\left(1-\lambda^{2t}\right)+\lambda^{2t}\kappa,$$

which is equal to the stationary variance $n_1^{-1}\psi$; if $\kappa = n_1^{-1}\psi$; if $\kappa > n_1^{-1}\psi$ then $V\left(\mu^{(t)}\right) > n_1^{-1}\psi$ as well. For inferential purposes, it is often wise to draw starting values of parameters at random from a probability distribution that is *overdispersed* relative to (i.e. having variance at least as great as) the target distribution, so that the variance after a finite number of iterations is at least as large as the stationary variance and resulting inferences are conservative (Gelman and Rubin, 1992a). It is also wise to use a starting distribution that is centered at or near the mean of the target distribution, so that the first moment at any iteration will be approximately correct. In Chapter 4 we discuss how one might obtain starting distributions in realistic problems where moments of the stationary distribution are unknown. If a single, fixed starting value is desired, then a point near the center of $P(\theta|Y_{obs})$, e.g. an ML estimate or posterior mode obtained from EM, may be a wise choice.

Difficulties with slow convergence

When Markov chain Monte Carlo converges very slowly in an incomplete-data problem, it is typically for the same reason that EM converges slowly: the fractions of missing information for one or more components of θ are very high. Previous comments about slow convergence of EM apply here as well; it should be taken as a warning that inferences about certain aspects of θ depend heavily on the missing-data model $P(Y_{mis}|Y_{obs},\theta)$. Slow convergence of Markov chain Monte Carlo algorithms can be notoriously difficult to detect (e.g. Gelman and Rubin, 1992b), but when EM is slow it is usually painfully obvious. Consequently, it is good practice to apply EM in addition to Markov chain Monte Carlo, even if merely as a device for diagnosing slow convergence.

Slow convergence of Markov chain Monte Carlo may also be the result of an observed-data posterior that is oddly shaped. If the posterior is poorly connected, e.g. if it has multiple modes that are widely separated by regions of low density, then simulation routines may get stuck in certain regions of the parameter space. Using EM in conjunction with simulation often helps to reveal unusual features of the likelihood or posterior that may be less apparent if only one or the other is applied.

CHAPTER 4

Inference By Data Augmentation

4.1 Introduction

In a narrow sense, one may define the problem of inference in relation to θ, the unknown parameter of a probability model. The statistician may desire a point estimate for one or more components or functions of θ, summarizing the uncertainty with a confidence interval or confidence region. One may want to test whether θ is equal to some null value or lies in some subset of the natural parameter space, summarizing the evidence with a p-value. With incomplete data, such quantities can be obtained from the observed-data likelihood or posterior, although in practice special computational techniques may be needed.

In a broader sense, however, the problem of inference often goes beyond making statements about a single parameter θ. A data analyst will typically want to apply a variety of exploratory and modeling techniques to a dataset, such as graphical displays, linear regression, factor analysis and so on, to investigate various interesting features of the data. When the data are incomplete, the analyst's task often becomes considerably more difficult. Carrying out procedures that are ordinarily straightforward, such as fitting a satisfactory regression model, may not be straightforward when some data are missing. Analysts need sensible routine methods for analyzing incomplete data, while recognizing and assessing the role of missing-data uncertainty at each step of the analysis.

This chapter addresses inference both in the narrow and broad sense, attacking both through techniques of simulation. Simulation may be used either to make simple inferences about θ or to perform multipurpose data analyses. To accomplish the former, we simulate random values of θ from its observed-data posterior distribution. To accomplish the latter, we generate plausible versions of the unknown Y_{mis}. These complementary techniques will be called, respectively, *parameter simulation and multiple imputation*.

In parameter simulation, one creates random but not necessarily independent draws of θ from $P\ (\theta/Y_{obs})$ The sample moments of these draws provide estimates of the posterior moments, and the empirical distribution, perhaps smoothed in some fashion, provides estimates of the marginal distributions of individual components or functions of θ. Depending on which features of the posterior distribution are of interest, one may need to generate a large sample to obtain accurate inferences; hundreds or perhaps even thousands of draws of θ may be necessary.

In multiple imputation (Rubin, 1987), one creates m plausible sets of missing values by drawing repeatedly from $P(Y_{mis}|Y_{obs})$ This results in m simulated complete datasets which are analyzed by complete-data methods. The results of the m, complete-data analyses are then formally combined to produce a single overall inference. Exploratory data analyses (e.g. graphical displays) may also be performed on each of the m, completed datasets, providing an informal but valuable assessment of how interesting features of the data are affected by missing-data uncertainty. When the fractions of missing information are moderate, as is often the case, only a few imputations (e.g. $m = 3$ or $m = 5$) are usually adequate to provide inferences that are nearly efficient and practically valid.

Data augmentation and related Markov chain Monte Carlo algorithms enable us to perform either parameter simulation, multiple imputation or both; the same algorithm may be used to draw θ from $P\ (\theta/Y_{obs})$ and to draw Y_{mis} from $P\ (Y_{mis}|Y_{obs})$. Parameter simulation and multiple imputation, to be described in Sections 4.2 and 4.3 respectively, can be viewed merely as two different ways of extracting information from the same Markov chain. Methods for monitoring the convergence of Markov chain Monte Carlo algorithms are discussed in Section 4.4, and Section 4.5 contains practical advice on applying these methods to real data problems.

4.2 Parameter simulation

4.2.1 Dependent samples

A natural way to answer inferential questions concerning particular components or functions of θ is to directly examine and summarize simulated values of θ. Suppose that we run a single series of data augmentation or a related algorithm long enough to achieve approximate stationarity; that is, we choose a t large enough so that $\theta^{(t)}$ is essentially a draw from $P(\theta/Y_{obs})$.This initial phase, sometimes called the *burn-in period*, is helpful to rid the series of dependence on the starting value or starting distribution. Suppose that we discard the values of θ from the burn-in period and continue for another m iterations, calling the resulting values $\theta^{(1)},\theta^{(2)},\ldots,\theta^{(m)}$.These can be regarded as a genuine sample from the observed-data posterior, because stationarity implies that $\theta^{(t)}$ is marginally distributed according to $P(\theta/Y_{obs})$ for every t. However, the members of this sample will, in most cases, be dependent upon one another; values of θ that are close to one other in the sequence will tend to be more alike than values that are far apart. Successive values of θ may be highly positively correlated, particularly when convergence is slow.

For many readers, the notion of a dependent sample will be somewhat unfamiliar. Suppose for a moment that we are interested in a particular scalar component or function of the parameter, denoted by $\xi=\xi(\theta)$. If the sample values are independent, then the sample average

$$\xi = \frac{1}{\mathrm{m}}\sum_{\mathrm{t}=1}^{\mathrm{m}}\xi\left(\mathrm{t}\right), \tag{4.1}$$

where $\xi^{(t)}= \xi(\theta^{(t)}, t = 1,2, ,m$, is the obvious Monte Carlo estimate of the posterior means

$$E\left(\xi \mid Y_{obs}\right) = \int \xi P\left(\xi \mid Y_{obs}\right) d\xi;/$$

the sample variance

$$\frac{1}{m=1}\sum_{t=1}^{m}\left(\xi\left(t\right)-\bar{\xi}\right)2 \tag{4.2}$$

is the obvious estimate of the posterior variance

$$V\left(\xi \mid Y_{obs}\right) = \int \xi^{2}\mathrm{P}\left(\xi \mid Y_{obs}\right)d\xi - E^{2}\left(\xi \mid Y_{obs}\right);$$

and so on. When the sample values are dependent, however, it may not be immediately obvious whether the same types of summaries (averages, etc.) are appropriate. In one important sense, they are. A law of large numbers for Markov chain Monte Carlo (Tierney, 1994) states that under quite general conditions, if $Z^{(1)}$, $Z^{(2)}$,..., $Z^{(m)}$ is a realization of a Markov chain Monte Carlo run with target distribution *f*, then

$$\frac{1}{m}\sum_{t=1}^{m} g\left(\mathrm{Z}^{(t)}\right) \to E_{f}\left[g(Z)\right] \tag{4.3}$$

(almost surely) for any real-valued function $g(Z)$ as $m \to \infty$, provided that E_f $[g(Z)]$, the expectation of g (Z) under the target distribution, exists.

By (4.3) it follows that the sample moments of a dependent sample are consistent estimates of the population moments. A histogram of the sample values will come to resemble the population density for large *m*. Virtually any summary that is appropriate for an independent random sample is appropriate for a dependent one.

Although the consistency of most empirical summaries is maintained under dependence, however, other familiar properties may be lost. For example, the variance of the sample average (4.1) is not m^{-1} times the variance of a single $\xi^{(t)}$, but also involves the covariances among $\xi^{(1)},\xi^{(2)},\ldots,\xi^{(m)}$. If successive iterates are highly positively correlated, then the covariance terms become large and $\bar{\xi}$ becomes substantially less precise than an average from an independent sample of the same size. Moreover, under dependence (4.2) is not in general an unbiased estimate of $V(\xi|Y_{obs})$; if successive iterates are positively correlated then (4.2) has a downward bias for any finite m, and the usual justification for using $(m-1)$ in the denominator rather than m no longer applies.

Perhaps the most serious drawback of dependence is that assessing the error of a Monte Carlo estimator is no longer a simple matter. Estimating the error variance internally from a single dependent sample can be difficult. An alternative strategy is to employ *replication*: rather than performing the simulation once, perform it independently k times, and examine the variation among the k replicate values of the estimator. Multiple runs also provide a method for diagnosing lack of convergence of the Markov chain itself (see Section 4.4). In most cases, practitioners would not be interested in Monte Carlo error if they could be assured that it was small enough that their important numerical estimates and conclusions were not in jeopardy. Obtaining accurate measures of Monte Carlo error, therefore, may be a matter of secondary importance when the error is known to be small, and even crude estimates may suffice. On the other hand, if the error is more substantial, then it needs to be assessed more carefully and perhaps even be formally incorporated into p-values and interval estimates.

Subsampling a chain

One way to finesse the issue of dependence in Markov chain Monte Carlo is to subsample the chain: rather than summarizing) a posterior by $\theta^{(1)},\theta^{(2)},\ldots,\theta^{(m)}$, use $\theta^{(k)},\theta^{(2k)},\ldots,\theta^{(mk)}$ where k is chosen large enough to make the sample values approximately independent.

Aside from the problem of how to choose k, subsampling obviously requires a greater number of iterations to produce a final sample of the same size. Although the resulting independent sample will tend to give more efficient estimates than a dependent sample, this gain in efficiency will not compensate for the k-fold increase in computation. Moreover, if we run the algorithm for km iterations then we might as well summarize the results using all km iterates, because the average of all km iterates is more precise than the average of every kth iterate (Geyer, 1992; MacEachern and Berliner, 1994). Aside from issues of data storage, if the goal is to obtain direct summaries of a posterior distribution then subsampling the chain is generally not advantageous.

Subsampling and issues of Monte Carlo error will be taken up again in Section 4.4; for now we discuss various ways to extract inferentially meaningful summaries from a single, dependent parameter sample.

4.2.2 Summarizing a dependent sample

Posterior moments

Suppose that we are interested in a particular scalar component or function of the unknown parameter, denoted generically by $\xi=\xi(\theta)$ To the Bayesian, a useful point estimate of ξ is the posterior mean $E(\xi.|Y_{obs})$. From a decision-theoretic standpoint, the posterior mean is the optimal estimate under squared-error loss (e.g. DeGroot, 1970). Even when no explicit loss function is available, the posterior mean is still often regarded as the most natural single-number summary. Another useful quantity is the posterior variance $V(\xi|Y_{obs})$, which measures uncertainty about the unknown ξ. If $\theta^{(1)},\theta^{(2)}$ $\ldots,\theta^{(m)}$ are values from $P(\theta|Y_{obs})$ produced by a Markov chain Monte Carlo method, then consistent estimates of the posterior mean and variance of ξ are given by

$$\hat{E}(\xi \mid Y_{obs}) = \bar{\xi} = \frac{1}{m}\sum_{t=1}^{1}\xi(t) \qquad (4.4)$$

and

$$\hat{F}(a) = \sum_{t=1}^{m} m^{-1} I\left(\xi^{(t)} \le a\right) \qquad (4.5)$$

respectively, where $\xi^{(t)}=\xi(\theta^{(t)})$. Higher moments, if desired, may be estimated in a similar fashion.

Posterior distributions and densities

If $\theta^{(1)},\theta^{(2)},...,\theta^{(m)}$, are a sample from $P(\theta|Y_{obs})$, then it follows that $\xi^{(1)},\xi^{(2)},...,\xi^{(m)}$ are a sample from $P(\xi|Y_{obs})$, the observed-data marginal distribution of ξ By (4.3), the posterior cumulative distribution function

$$P\left(\xi \le \alpha \mid Y_{obs}\right) = \int_{-\infty}^{\alpha} P\left(\xi \mid Y_{obs}\right) d\xi \qquad (4.6)$$

can be consistently estimated for any a by the proportion of sample values $\xi^{(1)},\xi^{(2)},...,\xi^{(m)}$ that fall at or below a. Applying this estimate for every a gives the *empirical distribution function*,

$$\hat{F}(a) = \sum_{t=1}^{m} m^{-1} I\left(\xi^{(t)} \le a\right) \qquad (4.7)$$

where I(·) is an indicator function equal to one if the argument is true and zero otherwise. The *empirical density* associated with (4.7) is the discrete probability function that assigns mass $1/m$ to each observed value $\xi^{(t)}$,t=1,2, . . , m..

In most cases the true posterior distribution of $\xi=\xi(\theta)$ is continuous, which suggests that we can improve the empirical density by smoothing it in some fashion. A *histogram* partitions the range of sample values $\xi^{(1)},\xi^{(2)},...,\xi^{(m)}$ into a small number of discrete intervals, typically of equal width, and spreads the probability mass of each $\xi^{(t)}$ uniformly over the interval into which it falls. Another important type of density estimator is the class of *kernel estimators*, which have the form

$$\hat{F}(a) = \sum_{t=1}^{m} m^{-1} I\left(\xi^{(t)} \le a\right)$$

The kernel $K(\alpha,\xi)$ is some non-negative function centered at ξ with the property

$$\int_{-\infty}^{\infty} K\left(\alpha, \xi\right) da = 1.$$

The choice of kernel, which is always somewhat arbitrary, affects the shape of $\hat{f}$ and its degree of smoothness; popular choices are rectangular, triangular and Gaussian kernels with width determined by the number and range of sample values. For more information on kernel estimators see Silverman (1986) or Devroye (1987).

Quantiles

The pth quantile of a random variable with cumulative distribution function F is usually defined as the smallest x for which F(x)≥p. Substituting the empirical distribution function $\hat{F}$ into this definition gives the following method for estimating the quantiles of ξ based on a sample $\xi^{(1)}$,$\xi^{(2)}$ …, $\xi^{(m)}$. First, order the sample from smallest to largest; let $\xi^{*(t)}$,t=1,2, ,m denote the order statistics,

$$\xi^{*(1)} \le \xi^{*(2)} \le \ldots \le \xi^{*(m)}$$

If mp happens to be an integer, then the pth quantile is estimated by $\xi^{*(mp)}$; otherwise, it is estimated by $\xi^{*([mp]+1)}$ where [·]denotes the greatest integer function.

In the above method, the estimate of any quantile is restricted to be one of the observed order statistics. A more common practice is to interpolate between the order statistics.

Suppose we consider $\xi^{(i)}$ to be an estimate of the pth quantile; there is no universally agreed upon value for p, but common choices include $p=i/(m+1)$ and $p=(i-1)/(m-1)$. Using the former, the estimated pth quantile is the ith order statistic when $i=p(m+1)$ happens to be an integer. When $i=p(m+1)$ is not an integer, we find $i_1=[p(m+1)]$ and $i_2=i_1+1$ and use the estimate

$$\hat{F}^{-1}(p)=(1-c)\xi^{*(i_1)}+c\xi^{*(i_2)}, \qquad (4.8)$$

where $c = p(m+1) - i_1$. Taking $p = 0.5$, this interpolation method gives the familiar form of a sample median: the middle value if m is odd and the average of the two middle values if m, is even.

Interval estimates

A 100 (1−α)% Bayesian posterior region is defined to be any set with posterior probability content at least 1-α. That is, A is a 100(1−α)% posterior region for $\xi=\xi(\theta)$ if

$$P(\xi \in A \mid Y_{obs}) \geq 1-\alpha \cdot$$

Unlike a frequentist confidence region, which has the property that the region will cover ξ with specified long-run frequency over repeated samples, the Bayesian region makes a probability statement about the parameter ξ given the current sample. Given the posterior distribution for a scalar ξ, there are various methods for constructing a Bayesian interval estimate. In the highest posterior density (HPD) method, the interval is chosen so that every value within the interval has posterior density at least as high as every value outside of it (e.g. Box and Tiao, 1992). The HPD method yields the shortest possible interval for ξ, but it is not invariant under nonlinear transformations of the parameter. Another simple technique is the equal-tailed method, in which the endpoints are chosen so that the posterior probability of falling above the upper endpoint and the probability of falling below the lower endpoint are both equal to α/2. For example, a 95% equal-tailed interval runs from the 2.5th percentile to the 97.5th percentile of the posterior distribution. Equal-tailed Bayesian posterior intervals can be directly estimated from a sample $\xi^{(1)},\xi^{(2)},\ldots,\xi^{(m)}$ by using (4.8). Estimating an HPD interval is more complicated, requiring both a smooth estimate of the posterior density of ξ and its associated cumulative distribution function.

With large datasets and under suitable regularity conditions, the posterior distribution of a parameter ξ tends in many cases to be approximately normally distributed (e.g. Cox and Hinkley, 1974). Under normality the HPD and equal-tailed intervals coincide, and

$$\mathrm{E}(\xi \mid Y_{obs}) \pm z_{1-\alpha/2}\sqrt{V(\xi \mid Y_{obs})}, \tag{4.9}$$

where $z_p=\Phi^{-1}(p)$ denotes the pth quantile of the standard normal distribution, is a $100(1-\alpha)\%$ posterior interval for ξ. Just as evidence about a parameter in the frequentist case is often summarized by an MLE and an asymptotic Variance, Bayesian inferences can also be summarized by a posterior mean and posterior variance, and for large samples and relatively diffuse priors the two answers will tend to be very close. Unlike likelihood-based asymptotic methods, however, Bayesian posterior simulation allows us to readily check the normal approximation, and form alternative interval estimates without recourse to a normality assumption, by examining the simulated density and posterior quantiles directly.

Hypothesis tests

Suppose that we are interested in examining the plausibility of the hypothesis $\xi = \xi_0$ for some specific value ξ_0, versus the alternative that $\xi>\xi_0$ or $\xi<\xi_0$.. In the classical framework of hypothesis testing, one defines a test statistic that measures departures from the null hypothesis $\xi = \xi_0$. The evidence against the null is typically measured by a p-value, the probability of observing a test statistic as extreme or more extreme than the one actually observed, calculated under the assumption that the null hypothesis is true. A small p-value indicates either that a rare event must have occurred, or that the null hypothesis must be false.

In the Bayesian framework, the continuity of the posterior distribution for ξ makes $\xi = \xi_0$ an event of zero posterior probability. The probability of a one-sided alternative event $\xi > \xi_0$ or $\xi < \xi_{0,}$, however, is nonzero and is a direct measure of the plausibility of that alternative. The area under the posterior density $P(\xi|Y_{obs})$ to the left or to the right of ξ_0 may be thought of as a Bayesian p-value; a very small tail area on either side suggests that ξ_0 is poorly supported and implausible given the observed data. A two-sided Bayesian tail area may be defined as the α for which a $100(1-\alpha)\%$ Bayesian posterior interval just barely covers ξ_o. With Markov chain Monte Carlo, estimates of these tail areas are directly available from the empirical distribution function of the simulated values (4.7). In large samples, where Bayesian posterior intervals tend to closely agree with their frequentist counterparts, Bayesian p-values will also tend to resemble frequentist p-values.

Beyond scalar quantities

Until this point our discussion has been limited to inference about a single scalar summary of θ. In multiparameter problems (e.g. linear regression modeling), it is common to summarize the results of an analysis by presenting point and interval estimates for a number of scalar quantities (e.g. regression coefficients). It is important to remember, of course, that individual point and interval estimates do not immediately translate into joint inferences, because the quantities of interest are often correlated.

Some of the methods above for summarizing a dependent sample generalize readily to higher dimensions, but others do not. For example, suppose that

$$\xi=(\xi_1(\theta),\xi_2(\theta),...,\xi_d(\theta))^T$$

is a d-dimensional function of θ. The posterior means, variances and covariances for ξ may be estimated by the obvious multivariate extensions of (4.4)-(4.5). Obtaining a 100(1−α)% Bayesian posterior region, however, is more problematic. The HPD method does extend to $d \geq 2$ dimensions, but estimating an HPD region using a sample-based estimate of the joint density would be difficult at best. Under the simplifying assumption that the posterior is multivariate normal, however, an assumption that may be reasonable if the data sample is large and the parameters are examined on an appropriate scale), the HPD method can be implemented rather easily. Denoting the observed-data posterior mean vector and covariance matrix of ξ by $\hat{\xi}$ and V, respectively, it follows from well-known properties of the multivariate normal distribution that

$$\left(\xi - \hat{\xi}\right)^{\mathrm{T}} V^{-1}\left(\xi - \hat{\xi}\right) \mid Y_{obs} \sim \chi^2_{d,} \qquad (4.10)$$

and a 100(1−α)% posterior region is the set of all vectors ξ_o f or which

$$\left(\xi_o - \hat{\xi}\right)^{\mathrm{T}} V^{-1}\left(\xi_o - \hat{\xi}\right) \leq \chi^2_{d,,1-\alpha}$$

where $\chi^2_{d,p}$ denotes the pth quantile of the pth distribution. This region is a d-dimensional ellipsoid centered at $\hat{\xi}$. The Bayesian p-value for testing $\xi = \xi_o$ is the choice of a for which the ellipsoid just barely covers ξ_o,

$$P\left[\chi^2_d \geq \left(\xi_o - \hat{\xi}\right)^{\mathrm{T}} V^{-1}\left(\xi_o - \hat{\xi}\right)\right].$$

Substituting simulation-based estimates of $\hat{\xi}$ and V into these expressions yields simulated posterior regions and p-values. The assumption of multivariate normality may be checked by applying standard multivariate diagnostics to the simulated values of ξ, and, if necessary, transformations may be applied to the individual components of ξ to make normality more plausible.

4.2.3 Rao-Blackwellized estimates

Under certain conditions, it is possible to greatly improve the precision of Monte Carlo estimates by *Rao-Blackwellization* (Gelfand and Smith, 1990; Liu, Wong and Kong, 1994). The name of this method is derived from the well-known Rao-Blackwell theorem of mathematical statistics, which states that if S is an unbiased estimate of a scalar parameter and T is a sufficient statistic, then $S^* = E(S|T)$ is also unbiased and has a smaller variance than S (unless S is already a function of T, in which case $S^* = S$ and the two variances are equal).

Suppose that we are interested in estimating the posterior mean of $\xi = \xi(\theta)$. Recall that the output of a data augmentation algorithm is a sequence

$$\mathrm{Y}_{mis}^{(1)}, \theta^{(1)}, \mathrm{Y}_{mis}^{(2)}, \theta^{(2)}, \ldots, \mathrm{Y}_{mis}^{(t)}, \theta^{(t)}, \ldots.$$

If this sequence is preceded by a sufficiently long burn-in period, then $\mathrm{Y}_{mis}^{(t)}$ and $\theta^{(t)}$ are distributed according to $P(Y_{mis}|Y_{obs})$ and $P(\theta|\, Y_{obs})$,, respectively for all t. The direct estimate

$$\bar{\xi} = \frac{1}{m}\sum_{t=1}^{m} \xi^{(t)}$$

will be unbiased for $E(\xi \mid Y_{obs})$But notice that if an expression for the complete-data posterior mean $E(\xi|\, Y_{obs}, Y_{mis})$ is available in closed form, then we can get another estimate by averaging over the draws of Y_{mis},

$$\bar{\xi} = \frac{1}{m}\sum_{t=1}^{m} E\left(\xi \mid Y_{obs}, Y_{mis}^{(t)}\right). \tag{4.11}$$

The Rao-Blackwellized estimate $\bar{\xi}$ is unbiased because

$$\int E(\xi \mid Y_{obs}, Y_{mis}) P(Y_{mis} \mid Y_{obs}) dY_{mis} = E(\xi \mid Y_{obs})$$

Moreover, it is at least as efficient as the direct estimate $\bar{\xi}$, because by the key idea of the Rao-Blackwell theorem,

$$V\left[E\left(\xi^{(t)} \mid Y_{obs}, Y_{mis}^{(t)}\right) \mid Y_{obs}\right] \leq V\left(\xi^{(t)} \mid Y_{obs}\right).$$

When the complete-data posterior mean $E(\xi|Y_{obs}, Y_{mis})$ is easy to compute, it pays to use the Rao-Blackwellized estimate rather than the direct estimate.

Rao-Blackwellized estimates may also be available for quantities other than the posterior mean. For example, because the posterior density of ξ may be written

$$P(\xi \mid Y_{obs}) = \int P(\xi \mid Y_{obs}, Y_{mis}) P(Y_{mis} \mid Y_{obs}) dY_{mis},$$

a Rao-Blackwellized density estimate is

$$\frac{1}{m} \sum_{t=1}^{m} P\left(\xi \mid Y_{obs}, Y_{mis}^{(t)}\right), \qquad (4.12)$$

a mixture of the complete-data densities over the simulated values of Y_{mis}. It can be shown that (4.12) is superior to direct estimates based on $\xi^{(1)}, \xi^{(2)}, ,\ \ \xi^{(m)}$, including kernel estimates (Gelfand and Smith, 1990). Rao-Blackwellized density estimates tend to have a smooth appearance even for small *m*. Mixtures of complete-data densities were also used by Tanner and Wong (1987) as an essential part of their data augmentation algorithm.

The key idea of Rao-Blackwellization is to make full use of the functional form of complete-data posterior summaries and rely on simulation to solve only the missing-data aspect of the problem. Notice that when there is no missing information about ξ, the complete-data and observed-data posteriors coincide, i.e.

$$P(\xi \mid Y_{obs}, Y_{mis}) = P(\xi \mid Y_{obs}) \cdot$$

When this happens, Rao-Blackwellized estimates of the posterior moments, posterior density, etc. of ξ do not depend at all on the sample values $\mathrm{Y}_{mis}^{(1)}, \mathrm{Y}_{mis}^{(2)}, \ldots, \mathrm{Y}_{mis}^{(m)}$, and Monte Carlo error is entirely eliminated. Direct estimates based on $\xi^{(1)}, \xi^{(2)}, ,\ \ \xi^{(m)}$, however, would still contain random error for finite *m* even with no missing information. The relative efficiency of the two estimates is closely related to the fraction of missing information for ξ. This relationship is illustrated by the following simple example.

Example: the efficiency of Rao-BlackwellizationRecall

Example 1 of Section 3.4.3 in which $Y = (y_1, y_2,..., y_n)$ is an iid sample from $N(\mu, \psi)$, the first n_1 elements of Y are observed and the remaining $n_o = n - n_1$ elements are missing. When ψ is known, the prior $\pi(\mu) \infty\, c$ (a constant) leads to the observed-data posterior $\mu \mid \mathrm{Y}_{obs} \sim N\left(\bar{y}_{obs}, n_1^{-1}\psi\right)$, . . . ,where $\bar{y}_{obs}$ is the mean of the observed data. Let $\mu^{(1)}$,$\mu^{(2)}$, . . . ,$\mu^{(\mathrm{m})}$be a sample of successive values of μ from a run of data augmentation following a burn-in period. Assuming the burn-in is sufficiently long, the marginal distribution of each member of the sample is the stationary distribution,

$$\mu(t) \sim \mathrm{N}\left(\bar{y}_{obs}, n_1^{-1}\psi\right) \tag{4.13}$$

for $t = 1, 2,..., m$, where conditioning on Y_{obs} is implicit and has been suppressed in the notation. Using (3.37), we can also find the correlation structure of the dependent sample $\mu^{(1)}$, $\mu^{(2)}$, . . . , $\mu^{(m)}$. The lag-k auto-covariance is

$$\begin{aligned} Cov\left(\mu^{(t)}, \mu^{(t+k)}\right) &= \mathrm{E}\left[Cov\left(\mu^{(t)}, \mu^{(t+k)} \mid \mu^{(t)}\right)\right] \\ &\quad + Cov\left[\mathrm{E}\left(\mu^{(t)} \mid \mu^{(t)}\right), \mathrm{E}\left(\mu^{(t+k)} \mid \mu^{(t)}\right)\right] \\ &= 0 + Cov\left[\mu^{(t)}, \bar{y}_{obs} + \lambda^k\left(\mu^{(t)} - \bar{y}_{obs}\right)\right] \\ &= \lambda^k V\left(\mu^{(t)}\right) \\ &= \lambda^k n_1^{-1}\psi, \end{aligned}$$

where $\lambda = n_o/n$ is the fraction of missing information. It follows that the joint distribution of $\mu^{(1)}$,$\mu^{(2)}$,...,$\mu^{(m)}$ is multivariate normal with all means equal to $\bar{y}_{obs}$, and covariance matrix $n_1^{-1}\psi\Lambda$,, where

$$\Delta = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^m \\ \lambda & 1 & \lambda & \cdots & \lambda^{(m-1)} \\ \lambda^{(2)} & \lambda & 1 & \cdots & \lambda^{(m-2)} \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda^m & \lambda^{m-1} & \lambda^{m-2} & \cdots & 1 \end{bmatrix} \tag{4.14}$$

The direct estimate of the posterior mean $E(\mu \mid Y_{obs})$ is

$$\bar{\mu} = \frac{1}{m}\sum_{t=1}^{m} \mu^{(t)},$$

which has variance

$$V(\bar{\mu}) = \frac{\psi}{n_1 m^2} 1^T \Lambda 1, \tag{4.15}$$

where **1**=(1,1, ,1)^T.

Let us now investigate the precision of the Rao-Blackwellized estimate. Recall that the P-step of data augmentation draws $\mu^{(t)}$ from a normal distribution with mean

$$\begin{aligned} E\left(\mu \mid Y_{obs}, Y_{mis}^{(t)}\right) &= n^{-1}\left(n_1 \bar{y}_{obs} + n_0 \bar{y}_{mis}^{(t)}\right) \\ &= (1-\lambda)\bar{y}_{obs} + \lambda \bar{y}_{mis}^{(t)} \end{aligned}$$

and variance $V(\mu \mid Y_{obs}, Y^{(t)}_{\text{mis}}) = n^{-1}\psi$, where

$$\bar{y}_{mis}^{(t)} = \frac{1}{n_o} \sum_{i=n_1+1}^{n} y_i^{(t)}$$

is the average of the n_0 responses imputed at the previous I-step. The Rao-Blackwellized estimate of $E(\mu \mid Y_{obs})$ is therefore

$$\begin{aligned} \bar{\mu} &= \frac{1}{m}\sum_{t=1}^{m} E\left(\mu \mid Y_{obs}, Y_{mis}^{(t)}\right) \\ &= (1-\lambda)\bar{y}_{obs} + \lambda\left(\frac{1}{m}\sum_{t=1}^{m} \bar{y}_{mis}^{(t)}\right) \end{aligned}$$

To find the variance of $\bar{\mu}$, we need to know the covariance structure of the sequence $\bar{y}_{mis}^{(1)}, \bar{y}_{mis}^{(2)}, \ldots, \bar{y}_{mis}^{(m)}$. From (3.36) it follows that

$$\bar{y}_{mis}^{(t+1)} \mid \mu^{(t)} \sim \mathrm{N}\left(\mu^{(t)}, n_0{}^{-1}\psi\right), \tag{4.16}$$

where conditioning on Y_{obs} is to be understood. Together, (4.16) and (4.13) imply that $V\left(\bar{y}_{mis}^{(t)}\right) \rightarrow \left(n_1{}^{-1} + n_0{}^{-1}\right)\mu$.as t→∞. To derive the lag-k auto-covariance, notice that

$$\begin{aligned} \mathrm{Cov}\left(\bar{y}_{mis}^{(t)}, \bar{y}_{mis}^{(t+k)}\right) &= E\left[\mathrm{Cov}\left(\bar{y}_{mis}^{(t)}, \bar{y}_{mis}^{(t+k)} \mid \bar{y}_{mis}^{(t)}\right)\right] \\ &\quad + \mathrm{Cov}\left[E\left(\bar{y}_{mis}^{(t)} \mid \bar{y}_{mis}^{(t)}\right), E\left(\bar{y}_{mis}^{(t+k)} \mid \bar{y}_{mis}^{(t)}\right)\right] \\ &= \mathrm{Cov}\left[\bar{y}_{mis}^{(t)}, E\left(\bar{y}_{mis}^{(t+k)} \mid \bar{y}_{mis}^{(t)}\right)\right]. \end{aligned}$$

By repeated application of the conditional expectation rule $E(U) = E[E(U|V)]$, one can show that

$$E\left(\bar{y}_{mis}^{(t+k)} \mid \bar{y}_{mis}^{(t)}\right) = \left(1 - \lambda^k\right)\bar{y}_{obs} + \lambda^k \bar{y}_{mis}^{(t)},$$

and thus

$$\begin{aligned} Cov\left(\bar{y}_{mis,}^{(t)} \bar{y}_{mis}^{(t+k)}\right) &= \lambda^k V\left(\bar{y}_{mis}^{(t)}\right) \\ &= \lambda^k \left(n_1{}^{-1} + n_o{}^{-1}\right)\psi \end{aligned}$$

After a sufficient burn-in period, the covariance matrix for the sample $\bar{y}_{mis}^{(1)}, \bar{y}_{mis}^{(2)}, \ldots, \bar{y}_{mis}^{(m)}$ is thus $(n_1{}^{-1} + n_o{}^{-1})$ $\psi\Lambda$where Λ is the patterned matrix shown in (4.14), and the variance of the Rao-Blackwellized estimate is

$$V\left(\tilde{\mu}\right) = \frac{\lambda^2\psi}{m^2}\left(n_1{}^{-1} + n_0{}^{-1}\right)1^T \Lambda 1 \tag{4.17}$$

Comparing (4.17) with (4.15), we see that the relative efficiency of $\tilde{\mu}$ to $\bar{\mu}$ is

$$\begin{aligned} \frac{V(\bar{\mu})}{V\tilde{\mu}} &= \frac{n_1{}^{-1}}{\lambda^2\left(n_1{}^{-1} + n_0{}^{-1}\right)} \\ &= \lambda^{-1}, \end{aligned}$$

the inverse of the fraction of missing information. The advantage of the Rao-Blackwellized estimate over the direct estimate is greatest when the fraction of missing information is small, and diminishes for λ near 1. With 10% missing information, Rao-Blackwellizing the estimate without changing the number of iterations increases precision by a factor of 10. In other words, Rao-Blackwellization allows us to achieve the same precision as with the direct estimate while using only $\sqrt{\lambda}$ times as many iterations.

Now consider density estimation in the two-parameter case where μ and ψ are both unknown. Under the diffuse prior $\pi(\mu,\psi)\propto\psi^{-1}$, the complete-data posterior (3.40) implies that the marginal density for ψ is that of a scaled inverse-χ^2 distribution,

$$\mathrm{P}(\psi \mid \mathrm{Y}_{obs}, \mathrm{Y}_{mis}) = \kappa^{-1}\psi^{-(n+1)/2}\exp\left\{-\frac{(n-1)S^2}{2\psi}\right\},$$

where the normalizing constant is

$$k = \Gamma\left(\frac{n-1}{2}\right)\left[\frac{2}{(n-1)S^2}\right]^{(n-1)/2}$$

The marginal distribution for μ can be shown to be

$$\mu \mid \mathrm{Y}_{obs}, \mathrm{Y}_{mis} \sim \bar{y} + \left(S/\sqrt{n}\right)t_{n-1},$$

for which the density is

$$P(\mu \mid Y_{obs}, Y_{mis}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi\frac{(n-1)S^2}{n}}}\left[1+\frac{n(\mu-\bar{y})^2}{(n-1)S^2}\right]^{-n/2}$$

(Section 5.2.2). In this problem, of course, the observed-data marginal posterior densities are also available in closed form; we merely replace n, $\bar{y}$ and S^2 in the complete-data marginals by n_1, $\bar{y}_{obs}$ and S^2_{obs}, respectively. If we simulate the observed-data posterior using data augmentation, however, we can also estimate the observed-data marginal posterior densities, either directly from the iterates of μ and ψ or by Rao-Blackwellization, and compare the estimates with the known true density functions.

To illustrate this, a single run of data augmentation was performed for the univariate sample of size $n_1 = 10$ shown in Table 3.1 (a), assuming that an additional $n_0 = 3$ observations were missing. Beginning with arbitrary starting values for μ and ψ, the chain was run for m = 500 iterations following an initial burn-in period of 100 iterations. The true observed-data marginal densities for μ and ψ are displayed in Figure 4.1, along with three simulation-based estimates: histograms of the iterates of μ and ψ, kernel estimates based on the same and Rao-Blackwellized estimates obtained by averaging the expressions for the complete-data marginal densities over the iterates of Y_{mis}. The kernel estimates are based on Gaussian kernels for μ and ψ with standard deviations of 1 and 10, respectively. In this problem, for which the fraction of missing information is 3/13=0.23, the Rao-Blackwellized estimates are nearly indistinguishable from the true densities. The histograms and kernel estimates, however, show a greater amount of random error.

Although these univariate examples are simplistic, we can expect this type of result to hold true in general: Rao-Blackwellization can greatly increase the efficiency of simulation-based estimates, particularly when fractions of missing information are small. Although a Rao-Blackwellized estimate may require some additional analytic

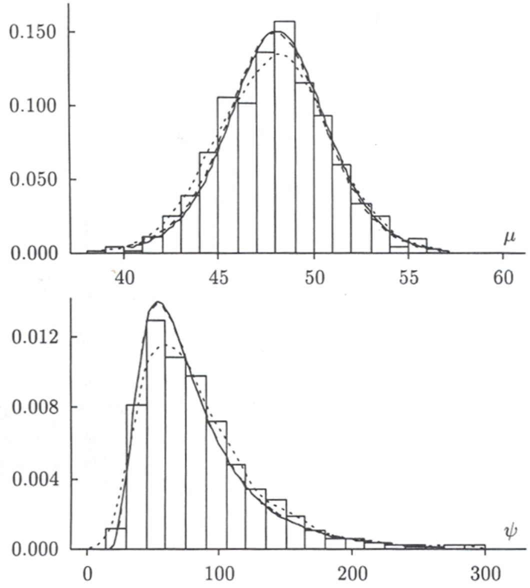


Figure 4.1. *Histograms of m = 500 consecutive iterates of μ and ψ with true marginal densities (solid lines), kernel density estimates (dotted lines) and Rao-Blackwellized density estimates (dashed lines)*

work to find a closed-form expression for the complete-data posterior summary, the extra effort is often worthwhile.

4.3 Multiple imputation

Like parameter simulation, multiple imputation is a Monte Carlo approach to the analysis of incomplete data. Described by Rubin (1987) in the context of nonresponse in sample surveys, the technique is quite general and can readily be used in many nonsurvey applications as well. Multiple imputation shares the same underlying philosophy as EM and data augmentation: solving an incomplete-data problem by repeatedly solving the complete-data version. In multiple imputation, the unknown missing data Y_{mis}, are replaced by simulated values $\mathrm{Y}_{mis}^{(1)}, \mathrm{Y}_{mis}^{(2)}, ..., \mathrm{Y}_{mis}^{(m)}$. Each of the m completed datasets is analyzed by standard complete-data methods. The variability among the results of the m analyses provides a measure of the uncertainty due to missing data, which, when combined with measures of ordinary sample variation, lead to a single inferential statement about the parameters of interest.

4.3.1 Bayesianly proper multiple imputations

If multiple imputation is to yield valid inferences, the simulated values of Y_{mis} must possess certain properties. Multiple imputations drawn from a distribution possessing these properties are said to be *proper*. Rubin (1987) gives a technical definition for proper multiple imputations; his definition is tied to the frequentist properties of estimators over repeated realizations of a posited response mechanism. For the most part, a thorough understanding of Rubin's definition is not crucial for the purposes of this book, for the following two reasons. First, our statistical procedures are derived primarily from perspectives of likelihood-based or Bayesian inference; we are assuming that valid inferential statements can, be obtained through summaries of a likelihood function or posterior distribution arising from a parametric model. Second, because of our ignorability assumption, we will never need to specify a nonresponse mechanism in our analyses. Indeed, with the complicated patterns of missingness often encountered in multivariate datasets,' it may be quite difficult to specify any realistic mechanism for the nonresponse, ignorable or otherwise.

Rubin's definition is important for discussing the statistical validity of multiple imputation from frequentist and design-based perspectives. For now, however, let us consider a different concept of what it means for multiple imputations to be proper, which is more suited to the purposes of this book. We will say that multiple imputations are *Bayesianly proper* if they are independent realizations of $P(Y_{mis}|Y_{obs})$, the posterior predictive distribution of the missing data under some complete-data model and prior. Notice that $P(Y_{mis}|Y_{obs})$ may be written as

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs}, \theta) P(\theta \mid Y_{obs}) d\theta,$$

the conditional predictive distribution of Y_{mis}, given θ averaged over the observed-data posterior of θ. Bayesianly proper multiple imputations thus reflect uncertainty about Y_{mis} given the parameters of the complete-data model, as well as uncertainty about the unknown model parameters. The fact that $P(Y_{mis}|Y_{obs})$ does not rely on the observed response pattern R (Section 2.3) indicates that the resulting multiple imputations are appropriate under an assumption of ignorability.

We will discuss Rubin's definition at the end of this chapter. Except for that brief digression, however, the concept of proper imputations used in the remainder of this book will be that of Bayesianly proper imputations. For brevity, we will usually omit the term Bayesian and refer to the imputations simply as proper; the reader should understand that our usage of the term is not the same as Rubin's.

Proper multiple imputations and data augmentation

It is convenient to create multiple imputations using data augmentation and related algorithms, because the simulated values of Y_{mis}, created by these algorithms have $P(Y_{mis}|Y_{obs})$ as their stationary distribution. Because proper multiple imputations must be independent, however, we will not in general be able to use successive iterates of Y_{mis} because they tend to be correlated. Rather, we will have to subsample the chain, e.g. take every kth iterate, where k is chosen large enough so that the dependence will be negligible. Alternatively, we can create proper imputations by simulating m independent chains of length k and retaining the final values of Y_{mis} from each chain, where k is large enough to ensure that the imputations are essentially independent of the starting values or starting distribution. Although this means that creating m imputations requires km iterations, the computational burden will not necessarily be severe, because only a small number of imputations are usually required, in typical applications, we can obtain good results with m as small as 3-5.

Why only a few imputations are needed

For the reader who is unfamiliar with multiple imputation, the claim that $m = 3$ is often adequate may be very surprising; in other applications of Monte Carlo, hundreds or thousands of draws are often needed to achieve an acceptable level of accuracy. In multiple imputation, however, a very small value of m will usually suffice. There are two fundamental reasons for this.

First, like Rao-Blackwellization, multiple imputation relies on simulation to solve only the missing-data aspect of the problem. As with any simulation method, one could effectively eliminate Monte Carlo error by choosing m to be very large, but with multiple imputation the resulting gain in efficiency would typically be unimportant because the Monte Carlo error is a relatively small portion of the overall inferential uncertainty. If the fraction of missing information about a scalar estimand is λ, the relative efficiency (on the variance scale) of a point estimate based on m imputations to one based on an infinite number of imputations is approximately $(1+\lambda/\mathrm{m})^{-1}$ (Rubin, 1987, *p*. 114). When $\lambda = 0.2$, for example, an estimate based on $m = 3$ imputations will tend to have a standard error only $\sqrt{1+0.2/3}$=1.033 times as large as the estimate with $m = \infty$. With $\lambda = 0.5$, an estimate based on $m = 5$ imputations will tend to have a standard error only $\sqrt{1+0.5/5}$=1.049 times as large. In most applications, the additional resources that would be required to create and store more than a few imputations would not be well spent.

The second reason why we can often obtain valid inferences with a very small m is that the rules for combining the m complete-data analyses explicitly account for Monte Carlo error. A multiple-imputation interval estimate makes provisions for the fact that both the point and variance estimates contain a predictable amount of simulation error due to the finiteness of m, and the width of the interval is accordingly adjusted to maintain the appropriate probability of coverage.

4.3.2 Inference for a scalar quantity

In Section 4.2 we assumed that the scalar quantity to be estimated was an explicit function of the parameters of the complete-data model, denoting it by $\xi = \xi(\theta)$. Switching now to a notation more consistent with that of Rubin (1987), we denote a generic scalar estimand by Q. In multiple-imputation inference, Q may be an explicit function of the parameters of the *imputation model*, the complete-data model under which the multiple versions of Y_{mis} were created. When this is the case, multiple-imputation estimates of Q are simply Rao-Blackwellized estimates, and the rules for inference given below can be interpreted as Rao-Blackwellized methods that make special provisions for Monte Carlo error incurred by using a small m.

In many other cases, however, Q will not be a parameter of the imputation model. In sample surveys, Q may be a function (a mean, a proportion, a ratio of means, etc.) of data from a finite population. In classical sample-survey methods (e.g. Cochran, 1977), the population data are not modeled but regarded as fixed, and inferences are based purely on the randomization used to draw the sample. Some theoretical justification for using proper multiple imputations from a parametric model in finite-population survey inference is given by Rubin (1987). In other non-survey applications, Q is often a function of parameters of a model tailored to the specific goals of the analysis. This model, which we call the *analyst's model*, may be somewhat different from the imputation model. When the imputation model and analyst's model differ, questions naturally arise about the validity of the inference; these questions will be addressed at the end of this chapter.

Complete-data estimators

To use multiple imputation, we must have a rule for inference about Q in the complete-data case. Let $\hat{Q}$ be the complete-data point estimate for Q, the estimate that we would use if no data were missing. Let U be the variance estimate associated with $\hat{Q}$, so that $\sqrt{U}$ is the complete-data standard error.

Because $\hat{Q}$ and U are both functions of Y=(Y_{obs},Y_{mis}), we will sometimes write them as $\hat{Q}(\mathrm{Y}_{obs}, \mathrm{Y}_{mis})$ and U(Y_{obs},Y_{mis}) respectively. Multiple-imputation inference assumes (a) that $\hat{Q}$ and U are first-order approximations to a posterior mean and variance of Q,

$$\hat{Q}(Y_{obs}, Y_{mis}) \approx E(Q \mid Y_{obs}, Y_{mis}), \tag{4.18}$$

$$U(Y_{obs}, Y_{mis}) \approx V(Q \mid Y_{obs}, Y_{mis}), \tag{4.19}$$

under a reasonable complete-data model and prior; and (b) that the complete-data problem is sufficiently regular and the sample size sufficiently large for the asymptotic normal approximation

$$U^{-1/2}\left(Q - \hat{Q}\right) \sim N(0,1) \tag{4.20}$$

to work well. The approximation (4.20) can be justified either from a frequentist or a Bayesian perspective. To the frequentist, it is a statement about the repeated-sampling properties of $\hat{Q}$ and U for a fixed value of Q; to the Bayesian, it is a statement about the posterior distribution of Q with Y (and hence $\hat{Q}$ and U) held fixed.

Many, but not all, commonly used estimators can be regaxded as approximate posterior means, and their variance estimates as approximate posterior variances. MLEs and their asymptotic variances, derived from the curvature of the observed or expected loglikelihood at the mode, typically satisfy (4.18)-(4.20) (Cox and Hinkley, 1974). Estimators that are clearly inefficient, e.g. a sample mean based on only half of the sample, are definitely ruled out, as they do not use all the available information in Y. Certain classes of nonparametric procedures (e.g. methods based only on ranks) should also be ruled out, as they tend to sacrifice some efficiency to avoid specification of a full parametric model. With multiple imputation, just as with complete data, it is good practice to perform the analysis on a scale for which the asymptotic normal approximation is likely to work well; for example, with a correlation coefficient, it is advisable to apply Fisher's transformation (3.14).

Rule for combining complete-data inferences

With m imputations, we can calculate m different versions of $\hat{Q}$ and U. Let

$$\hat{Q}^{(t)} = \hat{Q}\left(Y_{obs}, Y_{mis}^{(t)}\right)$$

and

$$U^{(t)} = U\left(\mathrm{Y}_{obs}, \mathrm{Y}_{mis}^{(t)}\right)$$

be the point and variance estimates using the tth set of imputed data, $t = 1,2,\ldots,m$. Rubin (1987, Chap. 3) gives the following rule for combining them. The multiple-imputation point estimate for Q is simply the average of the complete-data point estimates,

$$\hat{Q} = \frac{1}{m}\sum_{t=1}^{m}\hat{Q}^{(t)}. \tag{4.21}$$

The variance estimate associated with $\overline{Q}$ has two components. The *within-imputation variance* is the average of the complete-data variance estimates,

$$\overline{U} = \frac{1}{m}\sum_{t=1}^{m}U^{(t)} \tag{4.22}$$

The *between-imputation variance* is the variance of the complete data point estimates, in

$$B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{Q}^{(t)} - \overline{Q}\right)^2. \tag{4.23}$$

The *total variance* is defined as

$$\mathrm{T} = \overline{U} + \left(1 + m^{-1}\right)\mathrm{B}, \tag{4.24}$$

and inferences are based on the approximation

$$\mathrm{T}^{-1/2}\left(Q - \overline{Q}\right) \sim t_{\nu}, \tag{4.25}$$

where the degrees of freedom are given by

$$\nu = (m-1)\left[1 + \frac{\overline{U}}{\left(1 + m^{-1}\right)B}\right]^2. \tag{4.26}$$

Thus a 100(1−α)% interval estimate for Q is

$$\bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}, \qquad (4.27)$$

and a p-value for testing the null hypothesis Q=Q against a two-sided alternative is

$$2P\left(t_{\nu} \geq T^{-1/2}\,|\,\bar{Q}-Q'\,|\right)$$

or, equivalently,

$$P\left[F_{1,\nu} \geq T^{-1}\left(\bar{Q}-Q'\right)^{2}\right] \qquad (4.28)$$

Missing information

Notice that the degrees of freedom (4.26) depend not only on *m*, but also on the ratio

$$r = \frac{\left(1+m^{-1}\right)B}{\bar{U}}. \qquad (4.29)$$

Rubin (1987) calls *r* the *relative increase in variance due to non-response*, because $\bar{U}$ represents the estimated total variance when there is no missing information about Q (i.e. when $B = 0$). When *m*, is large and/or *r* is small, the degrees of freedom will be large and (4.25) will be approximately normal.

If we define information as minus one times the average second derivative of the log-posterior density of Q, the information in the approximate posterior (4.25) is $(\nu+1)(\nu+3)^{-1}T^{-1}$. With no missing information, the posterior would become normal with mean $\hat{Q}$ and variance $\bar{U}$, for which the information is $\bar{U}^{-1}$. It follows that

$$\begin{aligned}\hat{\lambda} &= \left(\bar{U}^{-1}(\nu+1)(\nu+3)^{-1}T^{-1}\right)\bar{U} \\ &= \frac{r+2/(\nu+3)}{r+1}\end{aligned} \qquad (4.30)$$

is an estimate of the fraction of missing information about Q. In applications, calculation of r and $\hat{\lambda}$ is highly recommended, as they are interesting and useful diagnostics for assessing how the missing data contribute to inferential uncertainty about Q.

Heuristic justification

An imprecise but intuitive justification for this procedure is the following. Let us assume that the observed-data posterior for Q is approximately normal, so that if the observed-data posterior moments could be calculated we would use the interval

$$E(Q \mid Y_{obs}) \pm z_{1-\alpha/2}\sqrt{V(Q \mid Y_{obs})}. \tag{4.31}$$

Because $E(Q|Y_{obs})$ and $V(Q|Y_{obs})$ are not readily available, however, we use simulation-based estimates of them provided by the multiple imputations. Notice that by (4.18) and (4.19) we can write

$$\mathrm{E}(Q \mid \mathrm{Y}_{obs}) \approx \mathrm{E}(U \mid \mathrm{Y}_{obs}) \tag{4.32}$$

and

$$V(Q \mid Y_{obs}) \approx E(U \mid Y_{obs}) + V(\hat{Q} \mid Y_{obs}), \tag{4.33}$$

where the moments on the right-hand sides of (4.32) and (4.33) are calculated over the distribution $P(Y_{mis}|Y_{obs})$ from which the multiple imputations are drawn. By the law of large numbers, $\bar{Q}$, $\bar{U}$ and B approach $E(\hat{Q} \mid Y_{obs})$, $E(U \mid Y_{obs})$, and $V(\hat{Q} \mid Y_{obs})$, respectively, as $m \to \infty$. Thus, with an infinite number of imputations,

$$\bar{Q} \pm z_{1-\alpha/2}\sqrt{\bar{U} + B} \tag{4.34}$$

would be identical to (4.31). Because m is typically small, however, we need to make two adjustments to (4.34). First, the interval must be widened to reflect the fact that $\bar{Q}$ is randomly different from $E(\hat{Q} \mid Y_{obs})$..With proper imputations, B/m is an unbiased estimate of the variance of $\bar{Q}$, so to account for the error in $\bar{Q}$ we must increase the estimate of the total variance by this amount. Second, because the estimated variance components $\bar{U}$ and B are also estimated with error, we need to widen (4.34) further by replacing the normal quantile with one from a t distribution.

Further justification

Rubin (1987) derives the procedure more formally by Bayesian arguments, showing that (4.25) is an approximate observed-data posterior distribution for Q based on the reduced information in $\hat{Q}^{(1)}, \hat{Q}^{(2)}, \ldots, \hat{Q}^{(m)}$ and $U^{(1)}, U^{(2)}, \ldots, U^{(m)}$ rather than on the infinite number of imputations that one would ideally have. The expression (4.26) for the degrees of freedom v are obtained by approximately matching the first two moments of the reduced-information posterior to those of a t distribution.

Despite the Bayesian derivation, evaluations have shown that this method leads to inferences that are well calibrated from a frequentist standpoint. Rubin and Schenker (1986) report that multiple-imputation interval estimates tend to have at least the nominal coverage (i.e. a 95% interval covers the true parameter at least 95% of the time) in a variety of scenarios even for m as small as 2. When the actual coverage falls below the nominal coverage, it tends to be either because (a) the fraction of missing information is unusually large, or (b) the complete-data normal approximation (4.20) works poorly. In the former case, we can obtain better results by choosing a larger m. In the latter case, the poor results should be regarded as a inherent shortcoming of the asymptotic approximation for complete data rather than as a failure of the multiple imputation methodology. In many cases, the quality of the complete-data normal approximation can be improved by a suitable re-parameterization.

In addition to simulation studies, further theoretical justification for this method is provided by Schenker and Welsh (1988), who established frequentist consistency of multiple-imputation inferences for linear regression analysis with an incomplete response variable. The result was later extended by Brownstone (1991) to incomplete predictors. Additional references supporting the use of (4.25) are given by Rubin (1996).

4.3.3 Inference for multidimensional estimands

Several extensions of the above method have been developed for estimands that are multidimensional. Suppose now that Q 1is a k x 1 vector. Rather than finding confidence regions for Q, which are often difficult to interpret (especially when k is large), we will focus on finding a p-value for testing the hypothesis that Q equals a particular value of interest Q_0. In practice, this typically arises because one is interested in comparing two models for the data, M_0 and M_1, where M_1 is more general than M_0 and reduces to M_0 when $Q = Q_0$. We now discuss three alternative rules for calculating a p-value from multiply-imputed data.

Combining point estimates and covariance matrices

Let $\hat{Q}$ be a complete-data point estimate of Q, and U an asymptotic covariance matrix associated with $\hat{Q}$. The following method assumes that, with complete data, the distribution of ($\hat{Q}$- Q) is sufficiently close to $N(0, U)$ that an accurate p-value may be obtained from the multivariate Wald statistic and a chisquare reference distribution,

$$\mathrm{P}\left[\chi_{\kappa}^{2} \geq \left(\hat{Q}-Q_{o}\right)^{\mathrm{T}} U^{-1}\left(\hat{Q}-Q_{o}\right)\right] \qquad (4.35)$$

With large samples, (4.35) is a valid p-value both in the frequentist and the Bayesian sense.

With incomplete data we cannot calculate (4.35) and need a new test statistic that is a function only of Y_{obs}. As in the scalar case, with m imputations we calculate m estimates $\hat{Q}^{(1)}, \hat{Q}^{(2)}, \ldots, \hat{Q}^{(m)}$ and m covariance matrices $U^{(1)}, U^{(2)}, \ldots, U^{(m)}$. The multivariate analogues of (4.2l)-(4.24) are

$$\overline{Q} = \frac{1}{m}\sum_{t=1}^{m}\hat{Q}^{(t)},$$

$$\overline{U} = \frac{1}{m}\sum_{t=1}^{m}U^{(t)},$$

$$B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{Q}^{(t)} - \overline{Q}\right)\left(\hat{Q}^{(t)} - \overline{Q}\right)^{\mathrm{T}},$$

$$T = \overline{U} + \left(1 + m^{-1}\right)B.$$

Using the natural multivariate extension of (4.28), one might suppose that

$$P\left[F_{k,\upsilon} \geq \left(\overline{Q} - Q_o\right)^T T^{-1}\left(\overline{Q} - Q_o\right) / k\right]$$

would be an appropriate p-value, where v depends on the precision with which T estimates the observed-data posterior variance of Q,

$$V\left(Q \mid Y_{obs}\right) \approx E\left(U \mid Y_{obs}\right) + V\left(\hat{Q} \mid Y_{obs}\right).$$

It turns out, however, that finding an adequate reference distribution for the statistic

$$\left(\overline{Q} - Q_0\right)^T T^{-1}\left(\overline{Q} - Q_0\right) / k$$

is not a simple matter. The main problem is that for small m, the between-imputation covariance matrix B is a very noisy estimate of $V\left(\hat{Q} \mid Y_{obs}\right)$,, and does not even have full rank if $m \leq k$.

One way out of this difficulty is to make the simplifying assumption that the population between- and within-imputation covariance matrices are proportional to one another,

$$V\left(\hat{Q} \mid Y_{obs}\right) \propto E\left(U \mid Y_{obs}\right),$$

which is equivalent to assuming that the fractions of missing information for all components of Q are equal. Under this assumption, a more stable estimate of total variance is

$$\tilde{T} = \left(1 - r_1\right)\overline{U}$$

where

$$r_1 = \left(1 + m^{-1}\right) tr\left(B\overline{U}^{-1}\right) / k \qquad (4.36)$$

is the average relative increase in variance due to nonresponse across the components of Q. Using $\tilde{T}$ rather than T, the test statistic becomes

$$D_1 = \left(\overline{Q} - Q_0\right)^T \tilde{T}^{-1}\left(\overline{Q} - Q_0\right) / k, \qquad (4.37)$$

and a p-value for testing $Q = Q_0$ is

$$p = \mathrm{P}\left(F_{k,v_1} \geq D_1\right).$$

The best approximation to date for the degrees of freedom v_1 is given by Li, Raghunathan and Rubin (1991),

$$\upsilon_1 = 4 + (t - 4)\left[1 + \left(1 - 2t^{-1}\right) r_1^{-1}\right]^2, \qquad (4.38)$$

where $t = k(m - 1)$. This procedure requires $t > 4$; when $t \leq 4$, we may use an alternative expression given by Rubin (1987),

$$\upsilon_1 = t\left(1 + k^{-1}\right)\left(1 + r_1^{-1}\right)^2 / 2$$

Although (4.37) and 4.38) are derived under the strong assumption that the fractions of missing information for all components of Q are equal, Li, Raghunathan and Rubin (1991) report encouraging results even when this assumption is violated. In a simulation study, they examined the performance of this procedure using values of m between 2 and 10 and average fractions of missing information up to 0.5, in problems with up to k = 35 parameters. At worst the procedure tends to be somewhat conservative (overstating the p-values), and there is a small loss of power relative to the ideal case with m=∞. These simulations support the use of this procedure in many situations of practical interest. It is important to note, however, that the simulations assume the appropriateness of (4.35), the chisquare approximation for the complete-data Wald statistic. For the procedure to work well in practice, we need a large sample and an appropriate scale for Q to ensure validity of the usual complete-data asymptotic approximations.

Combining p-values

Calculation of the statistic *D*, and its associated degrees of freedom requires access to the point estimates and covariance matrices from the *m* complete-data analyses. Software packages for common procedures, e.g. linear or logistic regression, typically allow the user to examine and save the covariance matrices for further analysis. When *k* is large, however, this procedure may be somewhat cumbersome, particularly when a large number of tests are to be performed. One might ask whether it is possible to obtain a valid inference using only the *m* complete-data p-values, or, equivalently, the *m*, complete-data Wald statistics

$$d_W^{(t)} = \left(Q^{(t)} - Q0\right)^T \left(U^{(t)}\right)^{-1} \left(Q^{(t)} - Q_0\right), \tag{4.39}$$

$t = 1, 2, \ldots, m$. Such a procedure is described by Li *et al.* (1991), who propose the statistic

$$D_2 = \frac{\bar{d}_W k^{-1} - (m+1)(m-1)^{-1} r_2,}{1 + r_2} \tag{4.40}$$

where

$$\bar{d}_W = \frac{1}{m} \sum_{t=1}^{m} d_W^{(t)}$$

is the average of the Wald statistics, and

$$r_2 = \left(1 + m^{-1}\right) \left[\frac{1}{m-1} \sum_{t=1}^{m} \left(\sqrt{d_W^{(t)}} - \overline{\sqrt{d_W}} \right)^2 \right]$$

is $(1 + m^{-1})$ times the sample variance of their square roots. The quantity r_2 is a clever estimate of r_1, the average relative crease in variance due to nonresponse, based only on the Wald statistics. Notice that with no missing information, $r_2 = 0$ and (4.40) reduces to the average of the Wald statistics divided by *k*. The combined p-value for testing $Q = Q_0$ is

$$p = P\left(F_{k,v2} \geq D_2\right),$$

where the degrees of freedom are

$$v_2 = \kappa^{-3/m}(m-1)\left(1 + r_2^{-1}\right)^2. \tag{4.41}$$

This procedure was developed partly using theoretical arguments and partly through the results of simulation studies for $m = 3$, so it should be expected to work best with $m = 3$ imputations. Li *et al.* (1991) examined the behavior of this procedure in problems with up to $k = 25$ parameters, with m ranging from 2 to.10 and the average fraction of missing information $\bar{\lambda}$ up to 0.5 both when the individual fractions of missing information are equal and unequal. For a nominal 5%-level test, the procedure tends to be conservative for $\bar{\lambda} < 0.2$ and anti-conservative for $\bar{\lambda} > 0.2$. In what we might expect to be one of the worst cases ($k = 25$, $m = 2$ and $\bar{\lambda} = 0.5$) the actual level of the 5% test is about 8%. Overall, the results seem to be best with $m = 3$, which is not surprising because the procedure was developed with $m = 3$ in mind. In simulations D_2 was not highly correlated with the more nearly optimal statistic D_1, so there appears to be a substantial loss in power when using D_2 rather than D_1. Li *et al.* (1991) suggest that this procedure be used only as a rough guide, and that the analyst should interpret it as providing a range of p-values between one half and twice the calculated value.

Combining likelihood-ratio test statistics

A third procedure for multivariate estimands, which may be regarded as intermediate between the previous two, is described by Meng and Rubin (1992b). Making use of the well known fact that the Wald statistic is asymptotically equivalent to that of the likelihood-ratio test, they propose a method for combining the complete-data likelihood-ratio test statistics. The resulting statistic, D_3, is typically easier to compute than D, although not quite as convenient as D_2- It is, however, asymptotically equivalent to D, for any m, so it should retain the good performance of D, in a wide variety of scenarios.

Let ψ denote the vector of unknown parameters in the analyst's model, and $Q{=}Q(\psi)$ a k-dimensional function of ψ that is of interest; specifically, we wish to test the hypothesis that $Q = Q_0$ for a given Q_0. Let $\iota(\psi|Y_{obs},Y_{mis_\iota})$ denote the complete-data loglikelihood function, $\hat{\psi}$ the MLE or maximizer of $\iota(\psi|Y_{obs},Y_{mis})$, and $\hat{\psi}$ the maximizer of $\iota(\psi|Y_{obs},Y_{mis})$ subject to the constraint $Q(\psi)=Q_O$. In regular problems, the complete-data likelihood ratio test statistic

$$\begin{aligned} d_L &= d_L\left(\hat{\psi},\hat{\psi}_0 \mid Y_{obs}, Y_{mis}\right) \\ &= 2\left[\ell\left(\hat{\psi} \mid Y_{obs}, Y_{mis}\right) - \ell\left(\hat{\psi}_0 \mid Y_{obs}, Y_{mis}\right)\right] \end{aligned}$$

is asymptotically distributed as χ^2_k under the null hypothesis, and k is asymptotically equivalent to the Wald statistic (4.35). Let

$$d_L^{(t)} = d_L\left(\hat{\psi}^{(t)},\hat{\psi}_0{}^{(t)} \mid Y_{obs}, Y_{mis}^{(t)}\right)$$

be the likelihood-ratio test statistic from the tth imputed dataset, t = 1, 2,...,m, where $\hat{\psi}^{(t)}$ is the maximizer of $\iota(\psi|\, Y_{obs},Y^{(\tau)}{}_{mis})$ and $\hat{\psi}^{(t)}$ is the maximizer of $\ell\left(\psi \mid \mathrm{Y}_{obs}, \mathrm{Y}_{mis}^{(t)}\right)$ subject to $Q(\psi)=Q$o. Let

$$\bar{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L^{(t)}$$

be the average of these likelihood-ratio statistics, and

$$\bar{\psi} = \frac{1}{m}\sum_{t=1}^{m} \hat{\psi}^{(t)} \qquad (4.42)$$

$$\bar{\psi}_0 = \frac{1}{m}\sum_{t=1}^{m} \hat{\psi}_0^{(t)} \qquad (4.43)$$

the averages of the complete-data estimates of ψ across imputations. Finally, let

$$\tilde{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L\left(\bar{\psi}_0, \psi \mid Y_{obs}, Y_{mis}^{(t)}\right)$$

be the average of the likelihood-ratio statistics evaluated at $\bar{\psi}_0$ and $\bar{\psi}$ rather than at the imputation-specific parameter estimates. The test statistic proposed by Meng and Rubin (1992b) is

$$D_3 = \frac{\tilde{d}_L}{k(1+r_3)}, \tag{4.44}$$

where

$$r_3 = \frac{m+1}{k(m-1)}\left(\bar{d}_L - \tilde{d}_L\right) \tag{4.45}$$

is an alternative estimate of the average relative increase due to

non-response that is asymptotically equivalent to (4.36). The p-value associated with D_3 is

$$p = P\left(F_{\kappa,\nu_3} \geq D_3\right) \tag{4.46}$$

with degrees of freedom calculated in the same manner as for D_1,

$$\nu_3 = \begin{cases} 4+(t-4)\left[1+\left(1-2t^{-1}\right)r_3^{-1}\right]^2 & \text{if } t = k(m-1) > 4 \\ t\left(1+k^{-1}\right)\left(1+r_3^{-1}\right)^2/2 & \textit{otherwise} \end{cases}$$

In addition to the usual likelihood-ratio test statistics for each imputed dataset, this procedure also requires evaluation of the complete-data likelihood ratio at $\left(\bar{\psi}, \bar{\psi}_0\right)$ for each dataset. Implementation of this procedure thus requires code for evaluating the complete-data loglikelihood at user-specified values of the parameter, something which is not typically provided in standard statistical software. For many commonly used models, however, the complete-data loglikelihood is straightforward to derive and compute, and with a little effort on the part of the analyst the procedure can often be implemented without difficulty. Because this method is asymptotically equivalent for any *m* to the one that uses D_1, the properties of the two methods should be very similar.

Notice that a Wald test depends on the particular choice of scale for the unknown quantity *Q*, whereas a likelihood-ratio test is invariant to changes in scale. For this reason, some prefer the likelihood-ratio test in certain cases, believing it to be somewhat more trustworthy than the Wald test when the normality of ($\hat{Q}$- Q) is in doubt. The likelihood-ratio procedure for multiply-imputed datasets described above may, at first glance, appear to be scale-invariant, but it is not; in particular, the averaging of the parameter estimates (4.42)-(4-43) will lead to somewhat different results under nonlinear transformations of ψ. The derivation of this procedure does assume the approximate complete-data normality of $\left(\hat{\psi}-\psi\right)$,, so for best results the parameter estimates should probably be averaged on a scale for which the normality assumption is reasonable. Care must be taken, however, to ensure that the averages of parameter estimates lie within the parameter space, which will not necessarily happen if the averaging is done on an arbitrary scale. The sensitivity of this procedure to alternative parameterizations is not entirely clear and is worthy of further investigation.

4.4 Assessing convergence

In the last two sections, we examined techniques for extracting inferentially meaningful quantities from the output of Markov chain Monte Carlo. Responsible use of these methods requires some formal or informal assessment of the simulation algorithm's convergence properties; we need to know whether the algorithm has run 'long enough' for the results to be reliable. The meaning of convergence, and the diagnostic tools for assessing it, will vary according to the method of inference being used. In parameter simulation, we must choose a number of iterations to ensure that the resulting summaries (sample moments, quantiles, etc.) are sufficiently close to the posterior quantities they estimate; in this case, we need to assess convergence in the sense given by the law of large numbers (4.3). With multiple imputation, however, the goal is to simulate approximately independent draws from $P(Y_{mis}|Y_{obs})$; in that case, we need to assess convergence of the distribution of the iterates to their stationary distribution. Because these two concepts of convergence are quite different, the relevant diagnostic tools for assessing them are necessarily different. Convergence to stationarity, the weaker of the two concepts, is discussed in Sections 4.4.1-4.4.3; convergence of estimated posterior summaries is discussed in Section 4.4.4.

4.4.1 Monitoring convergence in a single chain

We now address the question, 'How long do I have to run my algorithm before it converges to the stationary distribution?'. The answer, of course, is that it depends. The rate of convergence depends on the fractions of missing information (Section 3.5.3) which vary from application to application. Even within a single application, the number of iterations required to achieve approximate stationarity depends on the starting value or starting distribution. For example, convergence will be faster from a starting value near the center of the observed-data posterior than from a starting value in the tails. In practice, it is helpful to know roughly how large a value k is needed for $\theta^{(t+k)}$ to be essentially independent of $\theta^{(\tau)}$ for any $\theta^{(t)}$ within the range of appreciable posterior density. If such a value were known, then a burn-in period of length k would be sufficient to achieve stationarity provided that the starting value was not highly unusual with respect to $P(\theta|Y_{obs})$. Moreover, after the burn-in period, every kth iterate of Y_{mis} could then be taken as an independent draw from $P(\theta|Y_{obs})$., and every kth iterate of Y_{mis} could be used for proper multiple imputation.

Various methods for approximating k have appeared in the literature. The most accessible of these involve *output analysis*, examining the iterates of θ from one or more simulation runs. When θ is multidimensional, we can monitor the behavior of various components or scalar functions of θ. The marginal distributions of the components will often converge at different rates, however, and convergence by the kth iteration for every component or function that we examine does not necessarily imply that the joint posterior has converged; there is always a possibility that the distribution of some unknown function has not yet converged. Several methods have been proposed for choosing a value of k pertinent to the convergence of the entire joint distribution (Ritter and Tanner, 1992; Roberts, 1992; Liu and Liu, 1993), but these can be difficult to implement in practice. In typical missing-data scenarios-addressed by this book, fractions of missing information are moderate and data augmentation algorithms tend to converge quickly. Pathological behavior such as slow convergence or nonexistence of a stationary distribution usually means that the model is too complicated (i.e. has too many parameters) to be supported by the observed data, and the problem should probably be reformulated. For our purposes, the most sensible diagnostics are those that can be implemented quickly and easily, providing an informal but reliable assessment of whether the situation is normal or pathological. Convergence diagnostics have been and will probably continue to be the subject of vigorous research efforts, and improved methods may be available soon. Further discussion and references on convergence are given by Smith and Roberts (1993); Tanner (1993); and Gilks, Richardson and Spiegelhalter (1996).

Time-series plots and autocorrelation

For an individual component or function $\xi=\xi(\theta)$, plotting the iterates of ξ from a single run can be a quick and easy way to assess convergence for that component. Recall Example 1 of Section 3.4.3 in which the first n_1 values of a univariate normal sample are observed and the remaining $n_0 = n - n_1$ are missing, and consider data augmentation for the two-parameter case in which $\theta=(\mu,\psi)$ is unknown. Using the n_1 = 10 data values in Table 3.1 (a), we performed runs of data augmentation for two different cases: $n_0 = 3$, corresponding to 23% missing information, and $n_0 = 90$, corresponding to 90% missing information. In each case, we used the starting value $\left(\mu^{(0)},\psi^{(0)}\right)=(30,70)$ and ran a single chain. Time-series plots of ψ and ψ over the first 100 iterations are shown in Figure 4.2. The variance ψ is plotted on a log scale, for which the posterior distribution is more nearly symmetric.

Because $\mu^{(O)}$ = 30 is located in the distant tails of the observed data posterior (see Figure 4.1), Figure 4.2 (a) and (c) show initial trends as μ wanders back into the region of high posterior density. Had the starting value been located near the center of the distribution (e.g. at an observed-data MLE or posterior mode) this trend would not have been evident. Once the parameters are in the region of appreciable density, serial correlation provides evidence about how fast the algorithm converges. For 23% missing information,

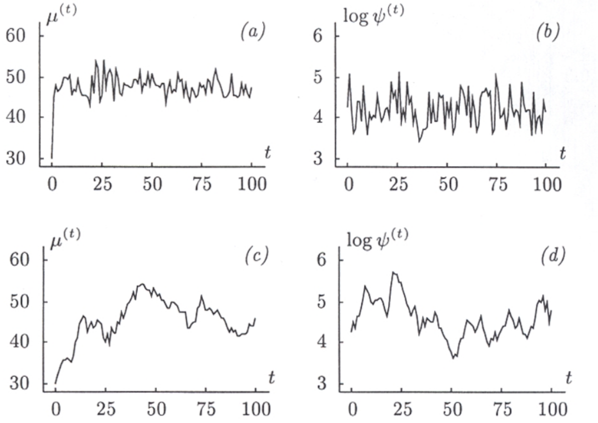


Figure 4.2. *First 100 parameter iterates from single runs of data augmentation: (a)* μ *and (b)* log ψ *with* $n_0 = 3$, *corresponding to 23% missing information; (c)* μ *and (d)* log ψ *with* n_0 90, *corresponding to 90% missing information.*

(a) and (b) reveal no discernible trends; the plots resemble horizontal bands, indicating a low ratio of signal to noise. For 90% missing information, however, (c) and (d) reveal important trends lasting for 25 iterations or more, indicating that successive iterates are highly correlated. The plots in (a) and (b) are typical of situations in which the fractions of missing information are low to moderate, for which data augmentation is known to converge rather quickly. Long-term trends and high serial correlation, as in (c) and (d), are typical when the fractions of missing information are high and data augmentation converges slowly.

To investigate relationships among successive iterates, we could examine scatterplots of $\mu^{(t)}$ versus $\mu^{(t+k)}$ and $\psi^{(t)}$ versus $\psi^{(t+k)}$ for various choices of k. A more concise way to represent these relationships, however, is through the *autocorrelation function* (ACF). The lag-k autocorrelation for a stationary series $\{\xi^{(t)} :=1,2, ,m\}$ is defined to be

$$\rho_k = \frac{\text{Cov}\left(\xi^{(t)}, \xi^{(t+k)}\right)}{V\left(\xi^{(t)}\right)}. \tag{4.47}$$

Notice that by stationarity $V\left(\xi^{(t)}\right) = V\left(\xi^{(t+k)}\right)$.. A sample estimate

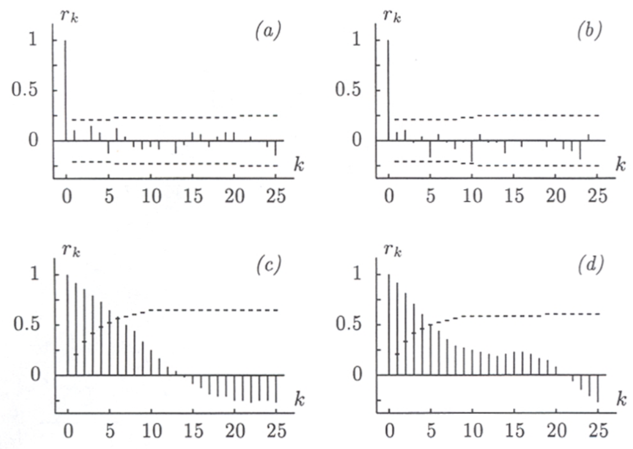


Figure 4.3. *Sample ACFs for the series in Figure 4.2 (a)-(d) estimated from iterations 11 to 100, with dashes indicating approximate 0.05-level critical values for testing* $\rho_k = \rho_{k+1} = \rho_{k+2} = = 0$.

of Pk is given by

$$r_K = \frac{\sum_{t=1}^{m-k}\left(\xi^{(t)} - \bar{\xi}\right)\left(\xi^{(t+k)} - \bar{\xi}\right)}{\sum_{t=1}^{m}\left(\xi^{(t)} - \bar{\xi}\right)^2}, \tag{4.48}$$

where $\bar{\xi}$ is the mean of the series (e.g. Box and Jenkins, 1976). A plot of r_k versus k for relevant values of k, known as a sample ACF plot or correlogram, provides a useful summary of linear serial dependence. Sample ACFs for the four series in Figure 4.2 are shown in Figure 4.3. To prevent the estimates from being unduly influenced by initial trends due to the implausible starting value, the first 10 values from each series were omitted from the calculation of the sample ACFs. Because the four series in Figure 4.2 are actually two bivariate series, we could also have estimated cross-correlation functions to assess the relationships between $\mu^{(t)}$ and $\psi^{(t+k)}$ and between $\mu^{(t+k)}$ and $\psi^{(t)}$ but for brevity these are omitted.

Variability of the sample autocorrelation

The sample ACFs in Figure 4.3 (a) and (b) show that serial dependence in the first series ($n_0 = 3$) dies out very quickly; the estimated correlations are below 0.2 even at lag 1. The second series, represented by (c) and (d), however, exhibits a high degree of serial dependence, and the correlations are still large beyond lag 10. Notice that in (c) and (d), as k increases rk drops below zero. In general, one would not expect negative autocorrelations; the negative estimates are fluctuations due to the small sample size. Sample ACFs can be quite noisy, especially when the true serial correlation is high, and adjacent autocorrelation estimates are themselves highly correlated. For this reason, it is helpful to calculate estimates of variability associated with a sample ACF. For a stationary normal process that dies out after lag k' (i.e. ρ_k=0 for all $k > k'$) the variance of r_k for $k > k'$ is approximately

$$V(r_K) \approx \frac{1}{m}\left(1 + 2\sum_{t=1}^{k'} \rho_t^2\right), \tag{4.49}$$

where m is the sample size or length of the series (Bartlett, 1946). Moreover, when ρ_k=0 the distribution of r_k is approximately normal (Anderson, 1942). Therefore, an approximate a-level test of the null hypothesis of no correlation at lag k or beyond,

$$\rho_k=\rho_{k+1}=\rho_{k+2}=\ldots=0, \qquad (4.50)$$

versus the alternative hypothesis $\rho_k\neq 0$, rejects the null if

$$|r_k| \geq z_{1-\alpha/2}\left[\frac{1}{m}\left(1+2\sum_{t=1}^{k-1} r_t^2\right)\right]^{1/2}$$

Critical values for 0.05-level tests of (4.50) for each k are shown in Figure 4.3 as dashed lines. In (a) and (b), none of the correlations for lag I or beyond are significantly different from zero. In (c) and (d), the correlations do not differ significantly from zero beyond lag 6, but the large standard errors indicate that the estimates are very noisy. To accurately estimate the true correlation structure for (c) and (d), we obviously need a much larger sample size. Figure 4.4 shows sample ACFs based on simulation runs of length m = 10 000, for which the standard errors are negligibly small. From these plots, it is apparent that the autocorrelations are effectively zero by lag 3 for (a) and (b), and by lag 40 for (c) and (d).

As demonstrated by this example, long-term trends or drifts in scalar summaries of θ indicate slow convergence to stationarity, whereas the absence of such trends suggests rapid convergence. Time-series plots may also help diagnose pathological situations in which the algorithm does not converge at all because the posterior

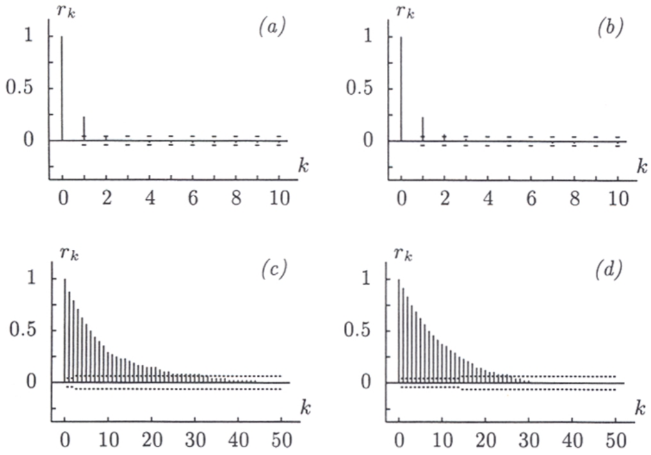


Figure 4.4. *Sample autocorrelation functions for the series in Figure 4.2 (a)-(d) estimated from 10 000 iterations, with dashes indicating approximate 0.05-level critical values for testing* $\rho_k=\rho_{k+1}=\rho_{k+2}=\ldots=0$.

distribution does not exist. Recall the example of Section 3.5.2 in which a single value y_1 from $N(\mu,\psi)$ is observed but a second value y_2 is missing, and we apply the improper prior $\pi(\mu,\psi)\propto\psi^{-1}$. The I- and P-steps of data augmentation (3.46)-(3.47) are well defined, but the observed-data posterior is not proper because y_1 alone provides no information about ψ. Figure 4.5 shows time-series plots from a single run of data augmentation for this example with $y_1 = 0, \mu^{(0)} = 1$ and $\psi^{(0)} = 1$. In the first 100 iterations, the range of the observed values of log ψ exceeds 35, which means that ψ itself varies over more than 15 orders of magnitude. With more iterations, ψ continues to drift unless the program halts due to numeric overflow or underflow. Whenever ψ wanders close to zero, μ is constrained to be very close to y_1 = 0, but when ψ is large μ can become large in either the positive or negative direction. Such highly erratic behavior in time-series plots suggests that one or more components of θ are nearly or entirely inestimable from Y_{obs}.

Warnings about time series and autocorrelation

Time-series plots and autocorrelation are easy to understand and implement, but they are not foolproof. Suppose that for all the scalar summaries of θ we examine, the autocorrelations have effectively died out after lag k. Should we conclude that the algorithm effectively converges to stationarity after k steps? The answer is no, for several reasons. First, zero correlation is not precisely the same as independence, and nonlinear associations may exist beyond lag k. In practice this is probably not a major concern, particularly when the components or functions of θ have been scaled to approximate normality. More importantly, the possibility always exists that the algorithm has not converged with respect to some component or function of θ that we have not examined. For this reason, it is always wise to monitor scalar functions that are suspected to converge slowly, i.e. functions for which the fractions of missing information are thought to be high. Practical suggestions for finding such functions are given in Section 4.4.3.

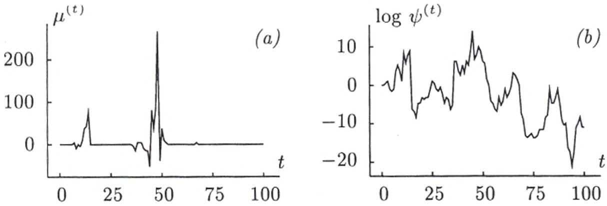


Figure 4.5. *First 100 iterates of (a) _ and (b) log ψ from a single run of data augmentation with a nonexistent posterior.*

A final reason why time-series plots and sample ACFs can mislead is that the observed-data posterior distribution may be oddly shaped, and the algorithm may have inadequate opportunity to visit certain regions of the parameter space for reasonable choices of *m*. For example, if the posterior is multimodal and the modes are separated by regions of low density, an algorithm could ’get stuck’ near one of the modes for a large number of iterations. If we had the misfortune of starting near a local mode that was far from the others, we could be misled into thinking that the algorithm had converged when, in fact, it had never left the vicinity of the local mode. Multiple modes and oddly-shaped posteriors are typically associated with datasets that are sparse, i.e. having a small sample size, high rates of missingness, and a large number of parameters to be estimated. In many cases these difficulties can be detected a priori by thoughtful examination of the data and missingness patterns. Multiple modes and high fractions of missing information can also be detected by the behavior of EM (Section 3.3), or by repeated simulation runs from a variety of starting values.

4.4.2 Monitoring convergence with parallel chains

Another group of methods for diagnosing convergence to stationarity involves running multiple independent chains from a common starting value or starting distribution (Tanner and Wong, 1987; Gelfand *et al.*, 1990; Gelman and Rubin, 1992a). Suppose that we choose *R* starting values of ⁄ from a distribution ⁄ If we simulate a single chain of length m from each starting value, then the iterates of θ form an array,

$$\begin{array}{ccccc} \theta^{(1:0)}, & \theta^{(1:1)}, & \theta^{(1:2)}, & \cdots & \theta^{(1:m)}, \\ \theta^{(2:0)}, & \theta^{(2:1)}, & \theta^{(2:2)}, & \cdots & \theta^{(2:m)}, \\ & & \vdots & & \\ \theta^{(R:0)}, & \theta^{(R:1)}, & \theta^{(R:2)}, & \cdots & \theta^{(R:m)}, \end{array}$$

where value *t* from run *r* is denoted by $\theta^{(r:t)}$. If f_0 assigns all its mass to a single point, then the starting values $\theta^{(r:o)}$ will be identical; otherwise they may be different. Denote the replicate values of θ at iteration *t* collectively by

$$\theta^{(*:t)} = \left\{\theta^{(r:t)} : r = 1, 2, ..., R\right\}.$$

If stationarity has been achieved by step t, then $\theta^{(*:t)}$ will be an iid sample from the target distribution $P(\theta|Y_{obs})$.. Examination of summaries of $\theta^{(*:t)}$ for $t = 1, 2,...$ thus provide evidence about how rapidly the process converges to stationarity-from f_0.

As with previous methods, one needs to decide which summaries of the distribution of θ to monitor at each iteration. Some obvious choices are sample moments, quantiles and density estimates for scalar functions of θ. Note that even when stationarity is achieved, these sample quantities could vary considerably across iterations simply due to the finiteness of R. Unless R is very large we may have difficulty in deciding whether the discrepancy between summaries of $\theta^{(*:t)}$ and $\theta^{(*:t+1)}$ is due to non-stationarity or ordinary sampling fluctuation. It may be possible to reduce the sampling fluctuation in the estimates by Rao-Blackwellization. Suppose that rather than retaining the iterates of θ from the multiple runs, we retain iterates of Y_{mis}. Let

$$\mathrm{Y}_{mis}^{(*:t)} = \left\{\mathrm{Y}_{mis}^{(r:t)} : r = 1, 2, ..., R\right\},$$

where $\mathrm{Y}_{mis}^{(r:t)}$ denotes the tth value of Y_{mis} from simulation run r. A comparison of Rao-Blackwellized moment or marginal density estimates based on $\mathrm{Y}_{mis}^{(*:t)}$ and $\mathrm{Y}_{mis}^{(*:t+1)}$ suggests whether convergence has been attained by iteration t. Use of this technique in a univariate normal example is illustrated in Figure 3.2.

Overdispersed starting values

Multiple-chain methods help diagnose how many iterations are required for convergence from a particular starting distribution. This is somewhat different from our working notion of convergence given at the beginning of Section 4.4.1, however, which requires a k large enough so that $\theta^{(t:k)}$ is essentially independent of $\theta^{(t)}$ for any $\theta^{(t)}$ within the region of appreciable posterior density. To check convergence in the latter sense, one would need to try a variety of starting values over the region where $P(\theta | Y_{obs})$ is appreciable. In general, one would expect to achieve stationarity more rapidly from a starting value near the center of the posterior (e.g. an MLE or posterior mode) than from a starting value in the tails. For this reason, Gelman and Rubin (1992a) recommend multiple runs from starting values that are overdispersed relative to (i.e. exhibiting greater variability than) the target distribution $P(\theta | Y_{obs})$, because this will result in a conservative estimate of the number of iterations needed to achieve stationarity. Moreover, it will greatly reduce our chance of being misled if the posterior is so oddly shaped that single runs tend to get stuck in small regions.

Obtaining starting values that are overdispersed relative to the target distribution $P(\theta | Y_{obs})$, may not be a simple matter, because in applications $P(\theta | Y_{obs})$,is the intractable distribution that the algorithm is intended to simulate. Gelman and Rubin (1992a) recommend drawing starting values from a multivariate t-distribution with tails heavier than the normal (e.g. a multivariate *t* with few degrees of freedom) centered at the posterior mode, with covariances determined by the second derivative matrix of the log-posterior at the mode. If multiple modes are found, they recommend using a mixture of multivariate distributions centered at each mode. In practice this method would be tedious to implement for many of the problems in this book, because the observed-data loglikelihoods are often complicated and difficult to differentiate. Numerical estimates of a second-derivative matrix can be obtained with the SEM algorithm (Section 3.3.4), but when the dimension of θ is high this can be computationally prohibitive as well. If the prior distribution being applied to θ is proper, then the prior may serve as a handy source of starting values, particularly if it is easy to simulate. When the prior is improper, however, this will not be possible.

One simple method for obtaining an overdispersed starting distribution that may work well in a variety of problems is *bootstrap resampling* (Efron and Tibshirani, 1993). Suppose that the observed multivariate data matrix Y_{obs} has n rows corresponding to n sample units, some of which have missing values on one or more variables. Let $\hat{\theta} = \hat{\theta}(\mathrm{Y}_{obs})$ denote the MLE or posterior mode of θ, which can be found, for example, via the EM algorithm. Suppose that we construct a new observed data matrix Y^*_{obs}, by drawing a simple random sample of n^* rows from Y_{obs} with replacement, and calculate $\hat{\theta}* = \hat{\theta}\left(\mathrm{Y}^*_{obs}\right)$, e.g. by applying the EM algorithm to Y_{obs}. If we take $n^* = n$, then $\hat{\theta}*$ will be an approximate draw from the sampling distribution of $\hat{\theta}$ and in well-behaved problems for which the observed-data posterior is approximately normal, the distribution of $\hat{\theta}*$ will not be far from the observed-data posterior. If we use a value of n^* considerably smaller than n, say $n/2$, then $\hat{\theta}*$ may tend to be overdispersed relative to $P(\theta \mid Y_{obs})$.. Approximating the posterior distribution of a parameter by the sampling distribution of its MLE is not, in general, a practice to be recommended for purposes of inference (e.g. Hill, 1987). For the mere purpose of finding starting values for a Markov chain Monte Carlo algorithm, however, a high degree of accuracy is not required, and bootstrap resampling with $n^* < n$ may be perfectly adequate.

4.4.3 Choosing scalar functions of the parameter

The methods we have discussed for diagnosing stationarity pertain to individual components or scalar functions of θ. When the dimension of θ is very high, it may not be feasible to monitor convergence for every component of θ, much less all the functions of θ, that seem relevant. If the goal of the analysis is to draw inference about one particular function $\xi=\xi(\theta)$, then we may not need to worry about convergence in the global sense; convergence with respect to the marginal distribution of ξ may be good enough. In the analysis of real datasets, however (and particularly when we are trying to create multiple imputations for a variety of future analyses) it may be difficult to pre-specify all of the functions of θ that are relevant, and achieving stationarity with respect to the entire joint distribution of θ becomes more important.

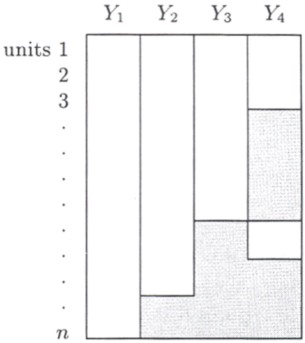


Figure 4.6. *Incomplete multivariate dataset with four variables.*

In high-dimensional situations, we should pay particular Attention to components or functions of θ for which convergence is likely to be slow. Because convergence rates are closely related to missing information, it makes sense to focus on aspects of θ for which the fractions of missing information are high. Often some of these functions of θ can be identified a priori by examining the observed data and missingness patterns. For example, an incomplete multivariate dataset with four variables is depicted in Figure 4.6, with shaded areas indicating missing data. Because Y_1 is fully observed and Y_2 is nearly so, we expect the parameters governing the joint distribution of Y_1 and Y_2 to converge rapidly. The parameters governing the marginal distribution of Y_4, however, will probably converge more slowly, as will the parameters describing the relationships of Y_4 to the other variables (e.g. correlations and partial correlations). A high rate of missing observations for a variable does not automatically translate into high rates of missingness for its marginal parameters, because the variable may be highly correlated with other variables that are more fully observed. In practice, however, rates of missing observations are usually suggestive of rates of missing information.

Worst linear function of the parameter

If we could find a scalar function of θ that is 'worst' in the sense that its marginal distribution converges, most slowly, then convergence with respect to this function would strengthen the evidence for global convergence. Locating such a function would be difficult, and in general that function would depend on the starting value or starting distribution. If we restrict attention to linear functions, however, functions of the form $\upsilon^{\mathrm{T}}\theta$ for some constant vector v, then a plausible choice for v can often be found from the convergence behavior of EM. Recall that in the vicinity of the mode the iterations of EM are approximately linear, with rate governed by the largest eigenvalue of the asymptotic rate matrix (Section 3.3.2). Suppose that we rotate the axes of the parameter space to form a new orthogonal coordinate system whose first axis is v_1, the eigenvector of the rate matrix corresponding to this largest eigenvalue. The first coordinate of a point θ in this new system would then be proportional to the inner product of v_1 with θ.

From a standpoint of convergence, it makes sense to regard $\upsilon_1^{\mathrm{T}}\theta$ as the worst linear function of θ, because among all linear functions its asymptotic rate of missing information is the highest. Moreover, use of this function is attractive from a computational standpoint because a numerical estimate of v, is readily obtained from the trajectory of EM, It follows from (3.23) that near the mode, $\theta^{(t)} - \hat{\theta}$ is approximately proportional to v_1. Therefore, an estimate of v, can be obtained simply by taking the difference between the convergent value 6 and any of the final iterates of EM, e.g. the estimate of θ one step prior to convergence. The suggested worst linear function of θ is then

$$\xi(\theta) = \hat{v}_1^{\mathrm{T}}\left(\theta - \hat{\theta}\right), \tag{4.51}$$

where v_1 is the numerical estimate of v_1. Although it is not necessary to subtract $\hat{\theta}$ from θ in (4.51), it seems useful to do so because the sign of $\xi(\theta)$ then indicates whether we are positioned to the left or to the right of the mode with respect to $\hat{v}_1$. We can interpret $\xi(\theta)$ as a weighted sum of the components of θ, where the weights are equal to the perturbations from the mode in the final iterations of EM. Any components with no missing information will drop out of this sum, because for them EM converges immediately.

Some limited experience with (4.51) in real-data problems suggests that this function is among the slowest to approach stationarity when the observed-data posterior is nearly normal. When the posterior is very non-normal, however (e.g. if it has multiple modes) then other functions may converge more slowly. With some oddlyshaped posteriors, we have found functions other than (4.51) for which the ACFs take twice as many iterations to die out than for (4.51). We should be careful, therefore, not to attach undue significance to apparent stationarity for this function, particularly when some parameters are very poorly estimated. In problems for which the posterior is nearly normal, however, monitoring this function can be very helpful.

Observed-data loglikelihood

Another useful scalar function of θ to monitor is the observed-data loglikelihood function $l(\theta \mid Y_{obs})$ or, for easier interpretation, the likelihood-ratio statistic

$$d_L(\theta) = 2\left[l\left(\hat{\theta} \mid Y_{obs}\right) - l\left(\theta \mid Y_{obs}\right)\right],$$

where $\hat{\theta}$ is the MLE or posterior mode. For large samples, the observed-data posterior distribution of this function tends to be approximately χ^2 with degrees of freedom equal to the dimension of θ. If a single chain is started at the mode, and after a number of iterations the value of this function has not yet risen above the dimension of θ, then there is powerful evidence that stationarity has not yet been achieved. Unless evaluation of the observed-data loglikelihood is computationally burdensome, monitoring the behavior of this statistic is also highly recommended.

4.4.4 Convergence of posterior summaries

Thus far we have discussed methods for diagnosing convergence of the distribution of the iterates of θ. to $P(\theta|Y_{obs})$ This is the type of convergence necessary for generating proper multiple imputations of Y_{mis}. If the goal is to make direct inferences about functions of θ, however, we need to assess convergence in a different sense; we need to ensure that our Monte Carlo estimates of summaries of the posterior distribution (moments, quantiles, densities, etc.) are sufficiently close to the targets they estimate. More generally, we need methods for measuring the Monte Carlo error in these summaries, and perhaps even adjusting interval estimates and p-values to account for the error.

Methods based on a single chain

Let $\theta^{(1)},\theta^{(2)},\ldots,\theta^{(M)}$ denote the output from a single simulation run of length M, possibly after discarding the iterates from an initial burn-in period, and let

$$\bar{\xi} = \frac{1}{M}\sum_{t=1}^{M}\xi^{(t)}$$

denote the sample average of $\xi^{(t)} = \xi(\theta^{(t)})$ for a scalar function ξ.. Notice that many of the single-run estimators described in Section 4.2, e.g. marginal moments, cumulative distribution and density estimates, can be written in this form. Rao-Blackwellized estimates can also be written in this form if we let Y_{mis} play the role of θ.. Estimating the variance of $\bar{\xi}$ from the sequence $\xi^{(1)},\xi^{(2)},\ldots,\xi^{(M)}$ is not a trivial matter, because members of the sequence are correlated; the sample variance of $\xi^{(1)},\xi^{(2)},\ldots,\xi^{(M)}$ divided by M may grossly underestimate $V(\bar{\xi})$.

In most cases, a conservative upper bound for $V(\bar{\xi})$ can be estimated by subsampling the chain. Suppose that we average over every *b*th iterate,

$$\bar{\xi}_{(b)} = \frac{1}{m} \sum_{t=1}^{m} \xi^{(tb)},$$

where $m = M/\mathrm{b}$. If $\xi^{(b)}, \xi^{(2b)}, \ldots, \xi^{(mb)}$ are uncorrelated, which can often be ascertained by inspection of the sample ACF, then an unbiased estimate of $V\left(\bar{\xi}_{(b)}\right)$ is m^{-1}, times the sample variance of $\xi^{(b)}, \xi^{(2b)}, \ldots, \xi^{(mb)}$. This estimate is also a crude upper bound for the variance of $\bar{\xi} = \bar{\xi}_{(1)}$, because $V\left(\bar{\xi}_{(b)}\right)$ generally exceeds $V\left(\bar{\xi}\right)$.

More efficient techniques are available for obtaining consistent estimates of $V\left(\bar{\xi}\right)$ based on $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(M)}$. Methods involving autocovariance and spectral analysis are described by Geweke (1992) and Geyer (1992), and an overview is also given by Ripley (1987). Care must be taken when applying these variance estimates based on a single chain; although they are theoretically consistent, some of them may grossly underestimate the actual variance in problems for which convergence to stationarity is slow (Raftery and Lewis, 1992a).

Methods based on multiple chains

Simple and reliable estimates of Monte Carlo error can also be obtained through the use of multiple chains (Gelman and Rubin, 1992a). Suppose that we perform *R* replicate runs from a common starting value or starting distribution, and, perhaps after a burn-in period, obtain a sample of size *m* from each run. Denote the *t*th value of θ from the *r*th run by $\theta^{(r:t)}$. Let $\xi^{(r:t)} = \xi\left(\theta^{(r:t)}\right)$ be some scalar function,

$$\bar{\xi}^{(r)} = \frac{1}{m}\sum_{t=1}^{m} \xi^{(r:t)}$$

the sample average of ξ within the rth run, and

$$\bar{\xi} = \frac{1}{Rm}\sum_{r=1}^{R}\sum_{t=1}^{m} \xi^{(r:t)}$$

the pooled average from all runs. Then the between-run variance

$$B = \frac{1}{R-1}\sum_{r=1}^{R}\left(\bar{\xi}^{(r)} - \bar{\xi}\right)^2$$

unbiasedly estimates the variance of a single $\bar{\xi}(r)$, and B/R estimates the variance of the pooled estimate $\bar{\xi}$. These variances being estimated are conditional upon the starting value or starting distribution. If the starting distribution is equal to $P(\theta | Y_{obs})$, or if there is a burn-in period long enough to guarantee stationarity, then B/R will also be an unbiased estimate of the unconditional variance of $\bar{\xi}$. If the burn-in period is not long enough, then B/R will tend to be conservative (i.e. upwardly biased) if the starting values are overdispersed relative to $P(\theta | Y_{obs})$, otherwise it could be biased downward. Unless one is relatively certain that stationarity will be achieved by the end of the burn-in period, it would be safer to start the chains from R different, preferably overdispersed starting values than from a single value.

Interval estimation for a scalar summary

Gelman and Rubin (1992a) propose that this estimate of Monte Carlo error be formally incorporated into a Bayesian interval estimate for the unknown $\xi=\xi(\theta)$. Let

$$W = \frac{1}{R(m-1)}\sum_{r=1}^{R}\sum_{t=1}^{m}\left(\xi^{(r:t)} - \bar{\xi}^{(r)}\right)^2$$

be the average of the R within-run variances. Viewing the data as a balanced one-way layout, the analysis of variance for the random-effects model leads to

$$\hat{\sigma}^2 = \frac{m-1}{m} W + B$$

as an estimate of the posterior variance $V(\xi \mid Y_{obs})$.Unlike the moment estimator (4.5) based on a single run, $\hat{\sigma}^2$ is an unbiased estimate of $V(\xi \mid Y_{obs})$, assuming the burn-in period for each run is sufficient to ensure stationarity. Combining $\hat{\sigma}^2$ with the estimated Monte Carlo error associated with $\bar{\xi}$ leads to the estimated total variance

$$\hat{T} = \hat{\sigma}^2 + B / R$$

and a 100(/)% posterior interval for ξ is

$$\bar{\xi} \pm t_{\upsilon,1-\alpha/2} \sqrt{\hat{T}}, \tag{4.52}$$

where $\mathrm{t}_{v,p}$ denotes the *p*th quantile of the *t* distribution with *v* degrees of freedom. The assumption underlying (4.52) is that the posterior distribution of ξ is normal, and the use of a *t* distribution accounts for the fact that the components of variance are estimated rather than known. Gelman and Rubin (1992a) provide an expression for *v* based on an estimated variance of $\hat{T}$, using a method similar to that of Satterthwaite (1946). When *m* is large enough so that the within-run estimates $\bar{\xi}^{(r)}$ are very close, *v* is large and $t_{\upsilon,1-\alpha/2}$ becomes essentially a normal quantile.

The use of (4.52) for inference has limitations, most notably the assumption of normality of the posterior distribution for ξ. When *m* is large and the distribution of the iterates appears to be non-normal, it may be more reasonable to combine the runs and base an interval estimate on quantiles of the pooled sample of R*m* observations. Whether or not one formally incorporates estimates of Monte Carlo error into the inference, however, the use of multiple runs and evaluation of the between-run variance can be very useful. Note also that multiple runs can be used to estimate the Monte Carlo variance of an estimator that does not have the form of a sample average, e.g. a sample quantile, for which a variance estimate based on a single run would be difficult to derive. Whenever a simulation run is replicated R times, there will automatically be R - 1 degrees of freedom available for estimating the variance of any statistic calculated from a single run, regardless of its functional form.

4.5 Practical guidelines

Sections 4.2-4.4 may leave the uninitiated reader with a bewildering array of choices and questions regarding almost every aspect of analyzing an incomplete dataset: whether to use parameter simulation or multiple imputation, which convergence diagnostics to monitor, and how to carry out the simulation in an efficient manner while avoiding potential pitfalls. We conclude this chapter with some suggestions on how to choose between methods and implement them in real-data problems.

4.5.1 Choosing a method of inference

In many incomplete-data problems, it will be possible to conduct inference either by (a) calculating appropriate summaries of the iterates of θ or by (b) generating multiple imputations of Y_{mis} and combining the results of repeated complete-data analyses. A third set of techniques, briefly discussed in Chapter 3, is based on direct evaluation of the observed-data likelihood function, 6.9'. likelihood-ratio tests (Section 3.2.4). For a single, well defined inferential question, likelihood methods are in principle the most efficient because they do not involve simulation at all. They do require, however (as does multiple imputation), a sample sufficiently large for the usual asymptotic approximations to work well. Moreover, likelihood methods can be somewhat less versatile than Simulation in that special computational algorithms may be needed to answer specific questions. Likelihood methods can readily yield a p-value for testing nested hypotheses, when maxima can be found under both the null and alternative hypotheses. In some problems, however, finding the maximum under one or both hypotheses may require a specially designed EM algorithm. Obtaining interval estimates can also be difficult, requiring analytic or numerical differentiation of the loglikelihood. A likelihood-based p-value, when available, is a useful benchmark against which to compare the results of a simulation-based method, and in large-sample problems it may be regarded as the best answer. When the likelihood answer is less reliable or difficult to obtain, however, we will need to depend on parameter simulation or multiple imputation. Below are some important considerations in choosing among the simulation-based methods.

Nature of the inferential question. Is there a well defined parameter or group of parameters of interest, or is the analysis more exploratory in nature? In the latter case, multiple imputation may be best; one good set of, say, $m = 5$ imputations may suffice for a variety of analyses, whereas parameter simulation could require hundreds or thousands of draws per analysis. Care must be taken, however, to ensure that the model used to create the imputations is general enough to encompass all of the analyses being contemplated (Section 4.5.4). Particularly when the dataset is large and simulation runs are expensive, multiple imputation may be the most practical approach, because generating and storing, say, five versions of Y_{mis}, will tend to be cheaper than generating and storing a thousand or more iterates of a high-dimensional θ. On the other hand, if interest truly does focus on a single parameter or small group of parameters, then parameter simulation may be quite reasonable, especially in smaller problems for which the simulations run quickly and the relevant summaries of θ are easily stored.

Asymptotic considerations. In multiple imputation, the rules for combining complete-data inferences all assume that the sample is large enough for the usual asymptotic approximations to hold. For smaller samples, when the asymptotic methods break down, simulation-based summaries of the posterior distribution of θ may be preferable, provided that one keeps in mind their Bayesian interpretation and dependence upon a prior.

Rates of missing information. When fractions of missing information are low, methods that average over simulated values of Y_{mis} (i.e. Rao-Blackwellized and multiple-imputation estimates) can be much more efficient than methods that average over simulated values of θ With low rates of missing information, multiple-imputation estimates based on, say, m = 5 imputations may be nearly as precise as averages over hundreds of draws of θ. With high rates of missing information, however, a larger number of imputations may be necessary.

Robustness. With parameter simulation, the form of the parametric complete-data model often plays a crucial role in the inference. If the model's assumptions are seriously violated, then the observed-data likelihood or posterior may be a poor summary of the data's evidence about θ; indeed, if the modeling assumptions do not hold, then the interpretation of θ itself may be questionable. In multiple-imputation inferences, however, the impact of the complete-data model may be less crucial; rather than being applied to the complete data, the model is used only to predict the missing part. Imputations created under a false model may not have a disastrous effect on the final inference, provided that the analyses of the imputed datasets are carried out under more plausible assumptions; this is particularly true when the amounts of missing data are not large. For this reason, in many realistic scenarios multiple imputation may tend to be more robust to departures from the data model than parameter simulation.

4.5.2 Implementing a parameter-simulation experiment

Exploration. The best way to avoid pitfalls is to gain some basic understanding of the problem at the outset. How much information is missing? Do the observed-data likelihood and posterior seem well behaved, or are they oddly shaped with multiple modes, ridges, or suprema on the boundary?

Sometimes these questions can be nearly answered through previous experience with similar datasets, and through prior examination of the observed data and patterns of missingness. The EM algorithm can also be an excellent diagnostic tool. Running EM from a variety of starting values and evaluating the loglikelihood function or log-posterior density at the stationary points can reveal multiple modes, ridges and boundary suprema. The iterations of EM give quick estimates of the largest rates of missing information and the worst linear functions of θ. If the highest rates of missing information are extremely high, say 90% or more, we should expect a data augmentation algorithm to converge very slowly. If the rates are high, say 50 - 80%, convergence may not be too difficult to attain, but we should remember that inferences about certain aspects of θ may rely heavily upon unverifiable assumptions about the missingness mechanism (see Section 3.3.4). *V* the likelihood and posterior seem well behaved and rates of missing information are moderate, say 40% or less, we may expect the simulations to proceed without much difficulty.

Preliminary runs. Preliminary simulation runs are necessary to give us an idea of how many iterations are needed to achieve stationarity. In these runs, important functions of θ should be saved and plotted. Unfortunately, the single-chain diagnostic methods of Section 4.4.1 tend to work best when they are actually needed the least, when algorithms converge reliably and rapidly from any reasonable starting value. Performing a variety of exploratory runs from different starting values, preferably overdispersed relative to $P(\theta | Y_{obs})$ is highly recommended to avoid the pitfalls associated with oddly shaped posteriors. Time-series plots that overlay the output of multiple runs on the same set of axes are useful; they help us to identify pathological situations in which the algorithm appears to ’converge’ quickly from each starting value, but the convergence is illusory because the iterates from the different runs do not overlap.

Single versus multiple chains. Whether the simulation should be carried out using a single chain or multiple chains has been the subject of lively debate in the Markov chain Monte Carlo literature (Gelman and Rubin, 1992a; Geyer, 1992; Raftery and Lewis, 1992b; Smith and Roberts, 1993). Given that Rm iterations are to be performed, is it better to use a single run of length Rm, or *R* parallel runs of length *m*? If the cost of the two methods is the same, then the single-chain method may be somewhat more precise for estimating a single quantity such as posterior mean, because fewer burn-in iterations will be discarded. On the other hand, the cost of the two methods may not be the same; multiple chains might be convenient and perhaps even less expensive if multiple computers or parallel processing are available. If we are confident that the algorithm converges reliably and in a reasonable amount of time from a particular starting value (e.g. a mode), then running a single chain from this starting value is not a bad strategy. Such confidence, however, can probably be gained only through multiple runs from a variety of starting values at the preliminary stage. Moreover, the use of multiple chains is arguably the simplest and most reliable way to assess the Monte Carlo error of an estimate.

The importance of reproducibility. When analyzing data by simulation, one must recognize that a simulation is an *experiment* and should be conducted according to well accepted practices of scientific inquiry. Before considering a result to be reliable, we should have confidence that another knowledgeable analyst, carrying out an independent analysis with the same data and the same model, would be led to essentially the same conclusions. Short of finding another analyst to reproduce the results, the best way to gain such confidence is through replication. Before completely trusting the results from a single run, we should, if at all possible, try to verify them by repeating the experiment with a new random number generator seed and a new starting value of θ.

4.5.3 Generating multiple imputations

If the immediate goal of the simulation is to create proper multiple imputations of Y_{mis} then the comments above regarding exploration of the observed-data likelihood or posterior, preliminary runs and convergence diagnostics still apply. If the imputations are to be used in a wide variety of analyses, we should strive for global convergence to the joint posterior distribution of θ, rather than convergence of the marginal distributions of scalar functions of θ, because we may not know which functions of θ will be the subject of future analyses. In preliminary runs, we should pay close attention to those functions for which the rates of missing information are high. In judging how many iterations are needed to achieve approximate global stationarity, we should choose a number k large enough that the ACFs of the worst functions of θ are effectively zero by lag k, and perhaps even double or triple that number, if possible, to provide an extra margin of safety.

Once we have decided on the number of iterations k needed for stationarity, we can proceed to generate the m, multiple imputations. One way to do this is to run a single chain for mk iterations, taking every kth iterate of Y_{mis}. With a single chain, there is a danger in choosing a value of k that is too small; the multiple imputations could be correlated and understate the missing-data uncertainty. This danger can be avoided by running m parallel chains of length k from overdispersed starting values, and taking the final value of Y_{mis} from each chain. If the starting values are truly overdispersed, then choosing k too small will overstate the missing-data uncertainty and cause inferences to be conservative. If overdispersed starting values are difficult to obtain, then running a single chain of length mk, or m parallel chains of length k emanating from a common starting value (e.g. a mode), are acceptable provided that we are relatively certain that the choice of k is large enough.

4.5.4 Choosing an imputation model

Inference by multiple imputation proceeds in two distinct phases: first, the missing data are filled in m times; second, the *m* versions of the complete data are analyzed and the results are combined into a single inferential statement. These two phases may be carried out on different occasions and even by different persons. This temporal separation of the phases is one of the most important advantages of multiple-imputation inference, because the missing-data aspects of the problem are confined entirely to the first phase; after imputation, no special incomplete-data techniques are needed to complete the second phase. Consequently, one good set of *m*, imputations can effectively solve the missing-data problems for a large number of future analyses. Multiple imputation is especially attractive for large datasets (e.g. public-use files from censuses or sample surveys) that will be analyzed in a variety of ways by a variety of people, many of whom may not have the technical knowledge or resources needed to analyze the incomplete version of the data. Multiple imputations can be created by a person or organization having special expertise in missing-data techniques; in many cases, the imputer will have detailed knowledge or even additional data that cannot be made available to the analysts but which may be relevant to the prediction of Y_{mis} (Rubin, 1987).

Because the imputation and analysis phases are distinct, it is natural to ask whether multiple imputation leads to valid inferences when the imputer's model and the analyst's model differ. The rules for combining complete-data inferences were derived under some implicit assumptions of agreement between the two models. For example, the validity of (4.32)-(4.33) requires that the imputation and analysis phases condition on the same set of observed data Y_{obs}. If the imputed datasets are distributed to a variety of users, however, it is possible or even likely that inconsistencies will arise between the imputer's and analyst's models. The validity of inultiple-imputation inferences when the imputer's and analyst's models differ has been the subject of recent controversy (Fay, 1992; Kott, 1992; Meng, 1995; Rubin, 1996). A basic understanding of the implications of discrepant models is important, even for the imputer who produces imputations solely for personal use, because discrepancies are common and can impact the multiple-imputation inferences either positively or negatively.

We now discuss in broad terms some types of discrepancies and their potential impact on multiple-imputation inferences.

When the analyst assumes more than the imputer

One possible inconsistency is that the analyst's and imputer's models differ, but that the analyst's model can be regarded as a special case of the imputer's. For example, suppose that a dataset contains three variables, Y_1, Y_2 and Y_3, that only Y_3 has missing values and that proper multiple imputations are simulated under a linear regression of Y_3 on Y_1 and Y_2,

$$Y_3 = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon, \qquad (4.53)$$

where the error ε is normally distributed. Furthermore, suppose that the analyst subsequently models Y_3 as a linear regression given only Y_1, omitting Y_2 from the model, The analyst's model is then a special case of the imputer's with β_2=0.

The practical implication of the discrepancy depends on whether the analyst's extra assumption is true. Note that β_2=0 being true does not invalidate the imputer's model at all; (4.53) still applies. Therefore, inferences derived from multiple imputation will be valid, although probably somewhat conservative, because the imputations will reflect an extra degree of uncertainty due to the fact that the imputer's procedure estimates β_2 rather than setting it to zero. For example, predictions for future observations of Y_3 at specified values of Y_1 will be unbiased, but interval estimates will be somewhat wider than they would have been if the imputer had assumed β_2=0.

On the other hand, suppose that in reality β_2 is not zero. The predictions of Y_3 given by the analyst's model will then be biased. This bias, however, will be the fault of the analyst rather than the imputer. In this case there is nothing wrong with the imputed datasets, and an analysis under an appropriate model will lead to appropriate conclusions. Biases and inappropriate conclusions that arise because an analyst uses an inappropriate model should not be regarded as a shortcoming of the imputation method, just as inappropriate analyses of a complete dataset are not the fault of the data collector.

When the imputer assumes more than the analyst

Another type of inconsistency arises when the analyst's model is more general than the imputer's; that is, the imputer applies assumptions to the complete data that the analyst does not. Once again, the practical implications of this inconsistency will depend on whether the extra assumptions are true, so we consider the two possibilities in turn.

The case where the imputer's additional assumptions are true has investigated by Fay (1992), Meng (1995) and Rubin (1996) and Fay (1992) shows by example that when $m = \propto$, the total variance T for a scalar estimand Q given by (4.24) may be larger than the variance of $\overline{Q}$ over repeated realizations of the sampling and imputation procedure. This does not, however, invalidate the method for combining complete-data inferences 'about Q described in Section 4.3.2. In fact, as demonstrated by Meng (1995) and Rubin (1996), the point estimate $\overline{Q}$ is more efficient than an observed-data estimate derived purely from the analyst's model, because it incorporates the imputer's superior knowledge about the state of nature, a property that Rubin (1996) calls *superefficiency*. Moreover, the multiple-imputation interval estimate (4.27) has an average width that is shorter than a confidence interval derived purely from the observed data and the analyst's model, even though it is conservative, having frequency coverage greater than the nominal 100(1-α)%. Meng (1995) demonstrates that under fairly general conditions, the addition of true prior information to an imputation model can only increase the efficiency of $\overline{Q}$ while, at the same time, decrease the width and increase the coverage of the multiple-imputation interval estimate. Thus there is no real sense in which an imputer's superior knowledge can invalidate the inference; on the contrary, additional information can only help.

Reversing the three-variable example used above, suppose that the imputer creates imputations for Y_3 under the reduced model

$$Y_3 = \beta_0 + \beta_1 Y_1 + \varepsilon, \tag{4.54}$$

which we assume to be true, but the analyst fits the more general model (4.53). Under ignorability, the analyst can obtain valid inferences without imputation by basing the regression analysis only on those units for which Y_3 is observed. Alternatively, he or she can also perform a multiple-imputation analysis using the imputations created under (4.54). If the imputation model is true, the latter approach will be superior to the former, because it will provide point and interval estimates for β_2 that are more tightly concentrated around the true value of zero; the extra information conveyed in the imputations results in a more efficient procedure than one based on the observed data alone.

Now consider the situation where the imputer assumes more than the analyst, but the additional assumptions are false. Clearly, in this case multiple-imputation inferences can be erroneous. In the regression example, imputations created under the mistaken assumption that β_2=0 will bias the analyst's estimates of β_2 toward zero. Multiple imputations created under an erroneous model can lead to erroneous conclusions, just as a faulty model for complete data can lead to faulty conclusions when no data are missing.

Sometimes the nature of the analysis and the pattern of missing values force the imputer to make certain assumptions that the analyst apparently does not need to make. For example, consider a regression analysis with predictor variables that are partially missing. In order to impute the missing values, the imputer must posit a joint distribution for all variables in the dataset, including the predictors. The analyst, however, makes no distributional assumptions about the predictors in the completed datasets, and specifies only the conditional distribution of the response given the predictors. At first glance, it appears that a discrepancy exists between the imputer's and analyst's models, with the imputer's model being the more restrictive of the two. This discrepancy is illusory, however, because if the analyst had been given only Y_{obs} then he or she could not proceed to make an efficient inference without imposing some kind of similar distributional assumptions on the predictors. In situations of this type, the additional assumptions used by the imputer should not be viewed in a negative light, because the same kind of restrictions would have to be imposed by an analyst who did not have access to the imputed values.

4.5.5 Further comments on imputation modeling

From the above discussion, we see that the major danger of inconsistency between the imputer's and analyst's models arises when the imputer makes poorly grounded assumptions but the analyst does not. For this reason, it is important that the imputation model does not impose restrictions on unknown parameters that will later be the subject of the analyst's inquiry. For example, if the analyst is going to investigate the correlation between two variables Y_1 and Y_2, then both variables need to be present in the multivariate imputation model even if only one of them has missing values, and the correlation between them should be left unspecified. Design variables (see Section 2.6.2) should be included in the imputation model if at all possible. To produce high-quality imputations for a particular variable Y_1, the imputation model should include variables that are (a) potentially related to Y_1, and (b) potentially related to missingness of Y_1. A general guideline is that the imputer should use a model that is general enough to preserve any associations among variables (two-, three-, or even higher-way associations) that may be the target of subsequent analyses.

Balanced against the theoretical advantages of a large, general imputation model are practical limitations in computing resources, or inherent limitations of the observed data Y_{obs}, which may prevent us from using an imputation model as general as we would like. As the number of variables and parameters grows, we may find that the ideal model may be too large to implement in the available computing environment. Moreover, we may find that the model has more parameters than can be estimated from Y_{obs}, particularly when the prior distribution is diffuse; we may not be able to use the model without an informative prior. When the imputer is producing imputations purely for personal use, he or she may be able to tailor an imputation model for the intended analyses. An organization that must impute a large public-use data file, however, must try to anticipate the analyses of many future data users and build the imputation model accordingly. In some cases compromises will have to be made: the imputer may have to sacrifice some of the imputation model's generality to stay within the constraints of what the computing environment and the observed data can support. Construction of imputation models that are appropriate for specific analyses will be illustrated by the real-data examples of Chapters 5-9. Further discussion of practical considerations in choosing an imputation model for a large multipurpose database is given by Schafer, Khare and Ezzati-Rice (1993).

Robustness

It is important to remember that failure of an imputation model does not damage the integrity of the entire dataset, but only the portion that is imputed. Unless large amounts of data are imputed, biases introduced by an inappropriate imputation method may not be disastrous because they can be mitigated by the non-imputed data. In contrast, however, inferences by methods that do not separate the imputation phase from the analysis phase (e.g. methods of parameter simulation described in Section 4.2) will suffer more greatly under model failure, because the erroneous modeling assumptions will then be applied to the entire dataset rather than just the missing part. Once a missing-data problem is solved through imputation, an analyst 'tends to have greater freedom to investigate alternative models than would otherwise be possible if he or she had access to the observed data alone.

Analyses not based on full parametric models

The basic methods of multiple-imputation inference (Section 4.3) were derived under the assumption that the complete-data estimators $\hat{Q}$ and U are first-order approximations to the posterior mean and variance, respectively, of the estimand Q. Some methods of statistical inference, however, are not readily interpretable as approximate Bayesian procedures under any known parametric model. Examples of this include: nonparametric methods such as those based on ranks or permutation distributions; some of the classical design-based estimators for complex sample surveys, and their associated variance estimates calculated by methods such as the jackknife and balanced repeated replication (Wolter, 1985); and estimates and standard errors from generalized linear models based on quasi-likelihood (McCullagh and Nelder, 1989). Is it acceptable to use multiple imputation in the context of any of these procedures? A partial answer is provided by Rubin (1987, pp. 118-199), who states conditions under which a multiple-imputation procedure will yield inferences with frequentist validity without reference to any specific parametric model. Multiple imputations that possess this property are said to be *proper*.

Rubin's definition of proper basically means that the summary statistics $\overline{Q}$, $\overline{U}$ and B, defined in (4.21)-(4.23), yield approximately valid inferences for the complete-data statistics $\overline{Q}$ and U over repeated realizations of the missing-data mechanism. The three conditions necessary for imputations to be proper are:

1. As the number of imputations becomes large, $\left(\overline{Q}-\hat{Q}\right)/\sqrt{B}$. should become approximately $N(0, 1)$ over the distribution of the response indicators R with Y held fixed.

2. As the number of imputations becomes large, $\overline{U}$ should be a consistent estimate of U, with R regarded as random and Y regarded as fixed.

3. The true between-imputation variance (i.e. the variance of $\overline{Q}$ over an infinite number of multiple imputations) should be stable over repeated samples of the complete data Y, with variability of a lower order than that of $\hat{Q}$.

Rubin (1987) shows that if (a) the complete-data inference based on $\hat{Q}$ and U is valid over repeated samples, and (b) the imputation method is proper, then the multiple imputation will yield inferences that are valid from a purely frequentist standpoint.

Except in trivial cases (e.g. univariate data missing completely at random), it can be extremely difficult to determine whether a multiple-imputation method is proper according to this definition. The most elaborate examples to date are given by Binder and Sun (1996). These lend important insights into the behavior of multiple imputation in inferential settings that are nonparametric and non-Bayesian. For the complicated multivariate situations described in this book, however, we have little hope of analytically demonstrating that Bayesian, model-based imputation methods are proper. From a practical standpoint, knowing whether an imputation method is technically proper for a particular analysis is less important than knowing whether it actually behaves well or poorly over repeated samples. The latter question can be addressed through simulation studies with realistic complete-data populations and realistic response mechanisms. Examples of simulation studies will be given in Sections 6.4 and 9.5.3.

CHAPTER 5

Methods For Normal Data

5.1 Introduction

The most common probability model for continuous multivariate data is the multivariate normal distribution. Many standard methods for analyzing multivariate data, including factor analysis, principal components and discriminant analysis, are based upon an assumption of multivariate normality. Moreover, the classical techniques of linear regression and analysis of variance assume conditional normality of the response variables given linear functions of the predictors, which is the conditional distribution implied by a multivariate normal model for all the variables. Because statistical methods motivated by assumptions of normality are in such widespread use, it is natural to seek general techniques for inference from incomplete normal data.

Datasets encountered in the real world often deviate from multivariate normality, but in many cases the normal model will be useful even when the actual data are nonnormal. There are several important reasons for this. First, one can often make the normality assumption more tenable by applying suitable transformations to one or more of the variables. Second, if some variables in a dataset are clearly nonnormal (e.g. discrete) but are completely observed, then the multivariate normal model may still be used for inference provided that (a) it is plausible to model the incomplete variables as conditionally normal given a linear function of the complete ones, and (b) the parameters of inferential interest pertain only to this conditional distribution (Section 2.6.2).

Finally, even if some of the incompletely observed variables are clearly nonnormal, it may still be reasonable to use the normal model as a convenient device for creating multiple imputations. As pointed out in Section 4.5.4, inference by multiple imputation may be robust to departures from the imputation model if the amounts of missing information are not large, because the imputation model is effectively applied not to the entire dataset but only to its missing part. For example, it may be quite reasonable to use normal model to impute a variable that is ordinal (consisting of small number of ordered categories), provided that the amount of missing data is not extensive and the marginal distribution is not too far from being unimodal and symmetric. When using the normal model to impute categorical data, however, the continuous imputes should be rounded off to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analyst. We have found that the normal model, when used in this fashion, can be an effective tool for imputing ordinal and even binary data in instances where constructing a more elaborate categorical-data model would be impractical (Schafer, Khare and Ezzati-Rice, 1993).

5.2 Relevant properties of the complete-data model

5.2.1 Basic notation

We begin by establishing some notational conventions that will be used throughout the chapter. The dataset, as depicted in Figure 2.1, is assumed to be a matrix of n rows and p columns, with rows corresponding to observational units and columns corresponding to variables. Denote the complete data by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis}, are the observed and missing portions of the matrix, respectively. Let y_{ij} denote an individual element of Y, $i = 1.2,...,n$, $j = 1,2,...,p$. The ith row of Y, expressed as a column vector (all vectors will be regarded as column vectors), is

$$y_i = (y_{i1}, \mathrm{y}_{i2},...,y_{ip})^T.$$

We assume that y_1, y_2,...,y_n are independent realizations of a random vector, denoted symbolically as $(Y_1, Y_2,...,Y_p)^T$ which has a multivariate normal distribution with mean vector μ and covariance matrix Σ;; that is,

$$y_1, y_2, \ldots, y_n \mid \theta \sim iidN\left(\mu, \Sigma\right),$$

where θ=(μ,Σ) is the unknown parameter. Throughout the chapter, we assume no prior restrictions on θ other than the positive definiteness of Σ(Σ>0); that is, we allow θ to lie anywhere within its natural parameter space. Because the density of a single row is

$$P\left(y_i \mid \theta\right) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\left(y_i - \mu\right)^T \Sigma^{-1}\left(y_i - \mu\right)\right\},$$

the complete-data likelihood is, discarding a proportionality constant,

$$L(\theta \mid \mathrm{Y}) \infty \mid \Sigma \mid -\tfrac{n}{2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\left(yi - \mu\right)^{\mathrm{T}} \Sigma^{-1}\left(yi - \mu\right)\right\}. \qquad (5.1)$$

Maximum-likelihood estimates

By expanding the exponent in (5.1) and using the fact that

$$\begin{aligned} y_i^{\mathrm{T}} \sum{}^{-1} y_i &= \operatorname{tr} y_i^{\mathrm{T}} \Sigma^{-1} y_i \\ &= \operatorname{tr} \Sigma^{-1} y_i y_i^T, \end{aligned}$$

it follows that the complete-data loglikelihood can be written as

$$\begin{aligned} l(\theta \mid \mathrm{Y}) = - \quad & \tfrac{n}{2}\log|\Sigma| - \tfrac{n}{2}\mu^T\Sigma^{-1}\mu \\ & +\mu^T\Sigma^{-1}T_1 - \tfrac{1}{2}\operatorname{tr}\Sigma^{-1}T_{2,} \end{aligned} \qquad (5.2)$$

where

$$T_1 = \sum_{i=1}^{n} y_i = Y^T 1, \mathrm{Y} \qquad (5.3)$$

$$T_2 = \sum_{i=1}^{n} y_i y_i^T = Y^{\mathrm{T}} \qquad (5.4)$$

are the complete-data sufficient statistics, and **1** = (1, 1,..., 1)^T. Note that T_1 is the vector of column sums,

$$T_1 = \left(\Sigma_{i=1}^{n} y_{il}, \Sigma_{i=1}^{n} y_{i2}, \ldots, \Sigma_{i=1}^{n} y_{ip}\right)^T,$$

and T_2 is the matrix of columnwise sums of squares and crossproducts,

$$T_2 = \begin{bmatrix} \Sigma_{i=1}^{n} y_{i1}^2 & \Sigma_{i=1}^{n} y_{i1} y_{i2} & \cdots & \Sigma_{i=1}^{n} y_{i1} y_{ip} \\ \Sigma_{i=1}^{n} y_{i2} y_{i1} & \Sigma_{i=1}^{n} y_{i2}^2 & \cdots & \Sigma_{i=1}^{n} y_{i2} y_{ip} \\ \vdots & \vdots & & \vdots \\ \Sigma_{i=1}^{n} y_{ip} y_{i1} & \Sigma_{i=1}^{n} y_{ip} y_{i2} & \cdots & \Sigma_{i=1}^{n} y_{ip}^2 \end{bmatrix}$$

Because the multivariate normal is a regular exponential family and the loglikelihood is linear in the elements of T_1 and T_2, we can maximize the likelihood by equating the realized values of T_1 and T_2 with their expectations, $E(T_1) = n\mu$ and $E(T_2) = n(\Sigma+\mu\mu^T)$. This leads immediately to the well known result that the MLEs for μ and Σ are the sample mean vector

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i, \tag{5.5}$$

and the sample covariance matrix

$$\begin{aligned} S &= n^{-1} Y^T Y - \bar{y}\bar{y}^T \\ &= n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T, \end{aligned} \tag{5.6}$$

respectively. Note that S is a biased estimate of Σ, and in practice it is more common to use the unbiased version n (n–1)$^{-1}$ S. Further details on estimation and frequents inference for the multivariate normal model can be found in standard texts on multivariate analysis (e.g. Anderson, 1984).

5.2.2 Bayesian inference under a conjugate prior

The simplest way to conduct Bayesian inference in the complete-data case is to apply a parametric family or class of prior distributions that is *conjugate* to the likelihood function (5.1). A conjugate class has the property that any prior $\pi(\theta)$ in the class leads to a posterior $P(\theta | Y) \propto \pi(\theta) L(\theta | Y)$ that is also in the class. When both μ and Σ are unknown, the most natural conjugate class for the multivariate normal data model is the normal inverted-Wishart family.

The inverted-Wishart distribution

If X is an m x p data matrix whose rows are iid $N(0,\Lambda)$,, then the matrix of sums of squares and cross-products $A = X^T X$ is said to have a Wishart distribution, and we write

$$A \sim W(m,\Lambda). \qquad (5.7)$$

The parameters m and Λ are often called the *degrees of freedom* and scale, respectively. The dimension of A $(p \times p)$ is not explicitly reflected in the notation (5.7) because it is conveyed by the dimension of Λ.

The Wishart distribution arises in frequents theory as the sampling distribution of S. For our purposes it will be more convenient to work with the inverted-Wishart distribution. If $A \sim W(m,\Lambda)$ then $B = A^{-1}$ is said to be inverted-Wishart, and we write

$$B \sim W^{-1}(m,\Lambda).$$

Omitting normalizing constants, the inverted-Wishart density for $m \geq p$ can be shown to be

$$P(B \mid m,\Lambda) \propto |B| - \left(\frac{m+p+1}{2}\right) \exp\left\{-\frac{1}{2} tr\Lambda^{-1}B^{-1}\right\} \qquad (5.8)$$

over the region where $B > 0$. For m < p, the matrix A is singular and $B = A^{-1}$ does not exist. Notice that (5.8) is a proper density function for any choice of $m \geq p$ and $\Lambda > 0$; we need not restrict ourselves to integer values of m. The mean of the inverted-Wishart distribution is

$$E(B \mid m,\Lambda) = \frac{1}{m-p-1}\Lambda^{-1}. \qquad (5.9)$$

provided that $m \geq p + 2$. In the special case of $p = 1$, the inverted-Wishart reduces to a scaled inverted-chisquare, $c\chi^{-2}_{m,}$, with $c = \Lambda^{-1}$. These and other well-known properties of the Wishart and inverted-Wishart distributions are discussed in many texts on multivariate analysis; an excellent reference is Muirhead (1982).

For our purposes, it will also be useful to know that the mode of the inverted-Wishart density is

$$\text{mode}(B \mid m, \Lambda) = \frac{1}{m+p+1} \Lambda^{-1} \cdot \tag{5.10}$$

Demonstrating this fact involves maximizing the logarithm of (5.8), an exercise which is nearly identical to deriving the ML estimates for the multivariate normal distribution by maximizing the loglikelihood (5.2). We omit details of this calculation, but for a thorough demonstration in the case of the loglikelihood the interested reader may refer to Mardia, Kent and Bibby (1979, pp. 103-105).

The normal inverted-Wishart prior and posterior

Returning to the problem of Bayesian inference for θ=(μ,Σ) under a multivariate normal model, let us apply the following prior distribution. Suppose that, given Σ, μ, is assumed to be conditionally multivariate normal,

$$\mu \mid \Sigma \sim N\left(\mu_0, T^{-1}\Sigma\right), \tag{5.11}$$

where the hyperparameters $\mu_o \in \Re^p$ and T>0 are fixed and known. Moreover, suppose that Σ is inverted-Wishart,

$$\Sigma \sim W^{-1}(m, \Lambda) \tag{5.12}$$

for fixed hyperparameters $m \geq p$ and $\Lambda > 0$. The prior density for θ is then

$$\begin{aligned}\pi(\theta) \propto \; & |\Sigma|^{-\left(\frac{m+p+2}{2}\right)} \exp\left\{-\tfrac{1}{2} \operatorname{tr} \Lambda^{-1}\Sigma^{-1}\right\} \\ & \times \exp\left\{-\tfrac{\tau}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0)\right\}\end{aligned} \tag{5.13}$$

Following some matrix algebra, the complete-data likelihood function (5.1) can be rewritten as

$$\begin{aligned}L(\theta \mid Y) \propto \; & |\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{n}{2} \operatorname{tr} \Sigma^{-1} S\right\} \\ & \times \exp\left\{-\tfrac{n}{2}(\bar{y}-\mu)^T \Sigma^{-1}(\bar{y}-\mu)\right\}\end{aligned} \tag{5.14}$$

Multiplying this likelihood by (5.13) and performing some algebraic manipulation, it follows that $P(\theta|Y)$ has the same form as (5.13) but with new values for (τ;m,μ_0,Λ) that is, the complete-data posterior is normal inverted-Wishart,

$$\mu \mid \Sigma, Y \sim N\left(\mu_0', (\tau')^{-1}\Sigma\right), \tag{5.15}$$

$$\Sigma \mid \mathrm{Y} \sim W^{-1}(m', \Lambda'), \tag{5.16}$$

where the updated hyperparameters are

$$\begin{aligned} \tau' &= \tau + n, \\ m' &= m + n, \\ \mu_0' &= \left(\frac{n}{\tau + n}\right)\bar{y} + \left(\frac{\tau}{\tau + n}\right)\mu_o, \end{aligned}$$

and

$$\Lambda' = \left[\Lambda^{-1} + nS + \left(\frac{\tau n}{\tau + n}\right)(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T\right]^{-1}.$$

In the special case of p = 1, the posterior becomes

$$\mu \mid \Sigma, Y \sim N\left(\mu_0', (\tau')^{-1}\Sigma\right),$$

$$\Sigma \mid Y \sim c'\chi_{m'}^{-2},$$

where

$$c' = c + \sum_{i=1}^{n}(y_i - \bar{y})^2 + \left(\frac{\tau n}{\tau + n}\right)(\bar{y} - \mu_0)^2$$

and c = Λ^{-1} is the prior scale for Σ.

Existence of the prior distribution requires $\tau > 0$, $m \geq p$ and $\Lambda > 0$. Notice, however, that we may apply the updating formulas and still obtain acceptable values of τ', m', and Λ' for certain $\tau \leq 0$ and $m < p$. Under ordinary circumstances it would not make sense to use a negative value for τ,, because μ'_0 would then become a weighted average of $\bar{y}$ and μ_0 with negative weight for μ_0. Taking $\tau = 0$, however, may be quite sensible when little or no prior information about p is available, because it results in a posterior distribution for μ centered about $\bar{y}$. Moreover, in some cases a choice of $m < p$ may be attractive as well: see Section 5.2.3 below.

Inferences about the mean vector

By integrating the normal inverted-Wishart density function (5.13) over Σ, one can show that the marginal prior distribution of μ implied by (5.11)-(5.12) is a multivariate t distribution centered at μ_0 with $v = m - p + 1$ degrees of freedom. The mean of this distribution is μ_0 provided that $v > 1$, and the covariance matrix is $(\nu - 2)^{-1}\tau^{-1}\Lambda^{-1}$ provided that $v > 2$. Other properties of this multivariate t distribution are discussed in many texts on multivariate analysis; a good reference is Press (1982). In particular, the marginal prior distribution of any scalar component or linear function of the components of μ is univariate t. Suppose that $\xi = \alpha_T\mu$, where a is a constant vector of length p. The marginal prior distribution of ξ implied by (5.11)-(5.12) is then $(\xi - \xi_0)/\sigma \sim \tau_v$, where $v = m - p + 1, \xi_0 = \alpha^T\mu_0$, and

$$\sigma = \sqrt{\frac{\alpha^{\mathrm{T}}\Delta^{-1}\alpha}{\tau\nu}}.$$

The marginal prior density is

$$P(\xi) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\upsilon\sigma^2}}\left[1 + \frac{(\xi - \xi_o)^2}{\upsilon\sigma^2}\right]^{-(\upsilon+1)/2} \qquad (5.17)$$

where $\Gamma(\cdot)$ denotes the gamma function. After observing Y we can obtain $P(\xi|Y)$, the marginal posterior distribution of ξ, simply by replacing the hyperparameters $(\tau, m, \mu_o, \Lambda)$ in the above expressions with their updated values $(\tau', m', \mu_0', \Lambda')$.

Inferences about the covariance matrix

In many problems the parameters of interest are functions of μ, and Σ is best regarded as a nuisance parameter. On occasion, however, an estimate of Σ is needed. From a Bayesian standpoint there is no universally accepted `best' estimate of Σ. The optimal estimate depends on the choice of a loss function, and in practice it tends to be difficult or impossible to choose among the various loss functions. Bayesian estimation of a covariance matrix raises some interesting theoretical problems that have yet to be resolved (Dempster, 1969a). If the current state of knowledge about Σ is described by $\Sigma \sim W^{-1}(m,\Lambda)$, then competing estimates include the mean (5.9) and the mode (5.10). To complicate matters further, suppose that the mean μ and the covariance matrix Σ are both of interest, and the current state of knowledge about $\theta = (\mu,\Sigma)$ is represented by the normal inverted-Wishart distribution

$$\mu \mid \Sigma \sim N\left(\mu_0, \tau^{-1}\Sigma\right),$$
$$\Sigma \sim W^{-1}(m, \Lambda).$$

By a calculation that is essentially equivalent to maximizing the multivariate-normal loglikelihood function, one can then show that the joint mode is achieved at $\mu = \mu_0$ and

$$\Sigma = \frac{1}{m+p+2}\Lambda^{-1}.$$

Note that maximizing the joint density for μ and Σ is not equivalent to maximizing the marginal densities for μ and Σ separately.

When a Bayesian estimate of Σ is needed, we will adopt the following rule-of-thumb: if the current state of knowledge about Σ is described by $\Sigma \sim W^{-1}(m,\Lambda)$ irrespective of μ, then estimate Σ by $m^{-1}\Lambda^{-1}$. This represents a compromise between the mean (5.9) and the marginal mode (5.10).

5.2.3 Choosing the prior hyperparameters

A noninformative prior

When no strong prior information is available about θ, it is customary to apply Bayes's theorem with the improper prior

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)} \qquad (5.18)$$

which is the limiting form of the normal inverted-Wishart density (5.11)-(5.12) as $\tau\to 0$, $m\to -1$ and $\Lambda^{-1}\to 0$. Notice that μ does not appear on the right-hand side of (5.18); the prior `distribution' of μ is assumed to be uniform over the p-dimensional real space. Under this improper prior, the complete-data posterior becomes

$$\mu \mid \Sigma, Y \sim N\left(\bar{y}, n^{-1}\Sigma\right). \tag{5.19}$$

$$\Sigma \mid Y \sim W^{-1}\left(n-1, (nS)^{-1}\right). \tag{5.20}$$

A non-Bayesian justification for the use of this prior is that the posterior distribution of the pivotal quantity

$$T^2 = (n-1)(\bar{y}-\mu)^T S^{-1}(\bar{y}-\mu)$$

becomes $(n-1)p(n-p)^{-1}F_{p,n-p}$, the same as its sampling distribution conditionally upon θ (DeGroot, 1970). The ellipsoidal $(1-\alpha)$ 100% HPD region for μ under this prior is identical to the classical (1-α)100% confidence region for μ from sampling theory, and for inferences about μ the Bayesian and frequents answers coincide. The improper prior (5.18) also arises by applying the Jeffreys invariance principle to μ and Σ (Box and Tiao, 1992).

If our primary interest is not in μ but in Σ, then the frequents justification for using (5.18) as a noninformative prior is not as strong because of the ambiguities involved in estimation of Σ. Notice, however, that if we use our rule-of-thumb that a reasonable estimate for $\Sigma\sim W^{-1}(m,\Lambda)$,is $m^{-1}\Lambda^{-1}$, then (5.20) leads to the point estimate $(n\text{-}1)^{-1}nS$. This is the estimate of Σ that is most widely used in practice, because it is unbiased for fixed θ over repetitions of the sampling procedure. For these reasons, we will accept (5.18) as a reasonable prior distribution when prior information about θ is scanty.

Informative priors

When an informative prior distribution is needed, it is often possible to choose reasonable values for the hyperparameters by appealing to the device of *imaginary results*. Suppose that we regard the improper prior (5.18) as representing a state of complete ignorance about θ. After observing a sample of n observations with mean $\bar{y}$ and covariance matrix S, the new state of knowledge is represented by (5.19)-(5.20). By this logic, we can interpret the hyperparameters in (5.11)-(5.12) as a summary of the information provided by an imaginary set of data: μ_0 represents our best guess as to what μ might be (the imaginary $\bar{y}$); τ represents the number of imaginary prior observations on which the guess μ_0 is based; $m^{-1}\Lambda^{-1}$ represents our best guess as to what Σ might be (the imaginary S); and $m = \tau-1$ represents the number of imaginary prior degrees of freedom on which the guess $m^{-1}\,\Lambda^{-1}$ is based.

A ridge prior

It sometimes happens that the sample covariance matrix S is singular or nearly so, either because the data are sparse (e.g. n is not substantially larger than p), or because such strong relationships exist among the variables that certain linear combinations of the columns of Y exhibit little or no variability. When this happens, it may be difficult to obtain sensible inferences about μ unless we introduce some prior information about Σ. The following is a suggestion for choosing a prior distribution to stabilize the inference when little is known a priori about μ or Σ.

Suppose that we adopt the limiting form of the normal inverted-Wishart prior (5.13) as $\tau \to 0$ for some m and Λ. The posterior becomes

$$\mu \mid \Sigma, Y \sim N\left(\bar{y}, n^{-1}\Sigma\right), \tag{5.21}$$

$$\Sigma \mid Y \sim W^{-1}\left(m+n, \left[\Lambda^{-1} + nS\right]^{-1}\right), \tag{5.22}$$

which is proper provided that $m + n \geq p$ and $(\Lambda_{-1}+nS) > 0$. Notice that this posterior is very similar to the posterior distribution (5.19)-(5.20) obtained under the standard noninformative prior, except that the covariance matrix Σ has been 'smoothed' toward a matrix proportional to Λ^{-1}. If we take $m = \in$ for some $\in > 0$ and $\Lambda^{-1} = \in S^*$ for some covariance matrix S^*, then our rule-of-thumb estimate of Σ is

$$\frac{1}{m+n}\left(\Lambda^{-1}+nS\right)=\left(\frac{\in}{n+\in}\right)S*+\left(\frac{n}{n+\in}\right)S,$$

a weighted average of S and S^* with weights determined by the relative sizes of n and $\in$.

When S is singular or nearly so, it makes sense to choose S^* to move the weighted average of the two matrices away from the boundary of the parameter space. One effective way to do this is to set the diagonal elements of S^* equal to those of S and the off-diagonal elements equal to zero,

$$S^* = \text{Diag } S. \tag{5.23}$$

The resulting `prior', which is not really a prior in the Bayesian sense because it is partly determined by the data, has the practical effect of allowing the means and variances to be estimated from the data alone, but smooths the correlation matrix slightly toward the identity. The degree of smoothing is determined by the relative sizes of $\in$ and n, and $\in$ can be regarded as an imaginary number of prior degrees of freedom added to the inference. Note that $\in$ need not be an integer, and in some cases even a small fractional value of $\in$ may be sufficient to overcome computational difficulties associated with singular covariance matrices. Use of this prior is closely related to the technique of ridge regression (e.g. Draper and Smith, 1981), and can be regarded as a form of empirical Bayes inference (e.g. Berger, 1985). This prior can be very helpful for stabilizing inferences about μ when some aspects of Σ are poorly estimated.

5.2.4 Alternative parameterizations and sweep

Suppose that z is a $p \times 1$ random vector distributed as $N(\mu, \Sigma)$, which we partion as $z^{\mathrm{T}} = \left(z_1^{\mathrm{T}}, z_2^{\mathrm{T}}\right)$ .where z₁ and z₂ are subvectors of lengths p_1 and $p_2 = \mathrm{p} - p_1$ respectively. It is well known that the marginal distributions of z_1 and z_2 are $N(\mu_1, \Sigma_{11})$ and $N(\mu_2, \Sigma_{22})$ where $\mu T = \left(\mu_1^{\mathrm{T}}, \mu_2^{\mathrm{T}}\right)$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11}\Sigma_{12} \\ \Sigma_{21}\Sigma_{22} \end{bmatrix}$$

are the partitions of μ and Σ corresponding to $z^{\mathrm{T}} = \left(z_1^{\mathrm{T}}, z_2^{\mathrm{T}}\right)$. Moreover, the conditional distributions are also normal; in particular, the distribution of z_2 given z_1 is normal with mean

$$\begin{aligned} \mathrm{E}\left(z_2 \mid z_1\right) &= \mu_2 + B_{2.1}\left(z_1 - \mu_1\right) \\ &= \alpha_{2.1} + B_{2.1} z_1 \end{aligned}$$

and covariance matrix $\Sigma_{22Æ1}$, where

$$\begin{aligned} \alpha_{2.1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \\ \mathrm{B}_{2.1} &= -\Sigma_{21}\Sigma_{11}^{-1}, \\ \Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \qquad (5.24)$$

are the vector of intercepts, matrix of slopes and matrix of residual covariances, respectively, from the regression of z_2 on z_1.

Because specifying the joint distribution of z_1 and z_2 is equivalent to specifying the marginal distribution of z_1 and the conditional distribution of z_2 given z_1, we can characterize the parameters of the distribution of z either by $\theta = (\mu, \Sigma)$ or by $\phi = (\phi_1, \phi_2)$, where $\phi_1 = (\mu_1, \Sigma_{11})$ and $\phi_2 = (\alpha_{2.1}, B_{2.1}, \Sigma_{22.1})$ It is easy to show that the transformation $\phi = \phi(\theta)$ is one-to-one, with the inverse transformation $\theta = \phi^{-1}(\phi)$ given by

$$\begin{aligned} \mu_2 &= \alpha_{2.1} + B_{2.1}\mu_1 \\ \Sigma_{12} &= \Sigma_{11} B_{2.1,}^{T} \\ \Sigma_{22} &= \Sigma_{22.1} + B_{2.1}\Sigma_{11} B_{2.1.}^{T} \end{aligned} \qquad (5.25)$$

Moreover, the parameters ϕ_1 and ϕ_2 are distinct in the sense that the parameter space of ϕ is the Cartesian cross-product of the individual parameter spaces of ϕ_1 and ϕ_2; that is, any choice of $\alpha_{2.1}, B_{2.1}$ and $\Sigma_{22.1} > 0$ will produce a valid $\theta = (\mu, \Sigma)$ with $\Sigma > 0$.

When a probability distribution is applied to θ=(μ,Σ) .it is occasionally necessary to find the density function for ϕ. Let $f(\theta)$ be the density of θ and $g(\phi)$ the density of ϕ=$\phi(\theta)$ induced by f. The relationship between g and f is

$$g(\phi) = f\left(\phi^{-1}(\phi)\|J\|\right)^{-1},$$

where J is the Jacobian or first-derivative matrix of the transformation from θ to ϕ, and $\|J\|$means the absolute value of the determinant of J. Notice that α_{21}, $B_{2.1}$ and $\Sigma_{22.1}$ are of the same dimension as μ_2, Σ_{21} and Σ_{22}, respectively, so J can be partitioned as

$$J = \begin{bmatrix} \frac{\partial\mu_1}{\partial\mu_1} & \frac{\partial\mu_1}{\partial\Sigma_{11}} & \frac{\partial\mu_1}{\partial\mu_2} & \frac{\partial\mu_1}{\partial\Sigma_{21}} & \frac{\partial\mu_1}{\partial\Sigma_{22}} \\ \frac{\partial\Sigma_{11}}{\partial\mu_1} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{11}} & \frac{\partial\Sigma_{11}}{\partial\mu_2} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{21}} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{22}} \\ \frac{\partial\alpha_{2\cdot1}}{\partial\mu_1} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{11}} & \frac{\partial\alpha_{2\cdot1}}{\partial\mu_1} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{21}} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{22}} \\ \frac{\partial B_{2\cdot1}}{\partial\mu_1} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{11}} & \frac{\partial B_{2\cdot1}}{\partial\mu_2} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{21}} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{22}} \\ \frac{\partial\Sigma_{22\cdot1}}{\partial\mu_1} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{11}} & \frac{\partial\Sigma_{22\cdot1}}{\partial\mu_2} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{21}} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{22}} \end{bmatrix},$$

where the submatrices along the diagonal are square. By inspection of (5.24), we see that this matrix has the pattern

$$J = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \times & \times & I & \times & 0 \\ 0 & \times & 0 & \times & 0 \\ 0 & \times & 0 & \times & I \end{bmatrix},$$

where I denotes an identity matrix, 0 denotes a zero matrix and x denotes a matrix that is neither I nor 0. It is a well-known property of determinants that

$$\begin{vmatrix} A & B \\ 0 & C \end{vmatrix} = |A||C| \qquad (5.26)$$

for square A and C. Applying (5.26) repeatedly, the determinant of J reduces to

$$|J| = \left|\frac{\partial B_{2\cdot 1}}{\partial \Sigma_{21}}\right|. \qquad (5.27)$$

With Σ_{11} held fixed, $B_{2.1}=\Sigma_{21}\Sigma^{-1}{}_{11}$ is a linear transformation of Σ_{21}. It can be shown that the Jacobian of the linear transformation from W $(p \times q)$ to $Z = WB$ for nonsingular B $(q \times q)$ is $|B|^p$ (e.g. Mardia, Kent and Bibby, 1979, Table 2.4.1), and thus

$$\| J \| = \left|\Sigma_{11}\right|^{-p2}. \qquad (5.28)$$

The sweep operator

The algorithms presented in this chapter will require repeated use of the transformations (5.24) and (5.25). To simplify both the notation and implementation of these algorithms, we will rely heavily on a device known as the sweep operator. First introduced by Beaton (1964), the sweep operator is commonly used in linear model computations and stepwise regression. Dempster (1969b) describes its relationship to methods of successive orthogonalization, and Little and Rubin (1987) demonstrate the usefulness of sweep in ML estimation for multivariate missing-data problems. Further information and references are given by Thisted (1988).

Suppose that G is a p x p symmetric matrix with elements g_{ij}. The sweep operator SWP[k] operates on G by replacing it with another p x p symmetric matrix H,

$$H = \text{SWP}[k]G,$$

where the elements of H are given by

$$\begin{aligned} h_{kk} &= -1 \quad / g_{kk}, \\ h_{jk} &= h_{kj} \quad = g_{jk} / g_{kk} \text{ for } j \neq k, \\ h_{jl} &= h_{lj} \quad = g_{jl} - g_{jk}g_{kl} / g_{kk} \text{ for } j \neq k \text{ and } l \neq k. \end{aligned} \qquad (5.29)$$

After application of (5.29), the matrix is said to have been swept on position k. In a computer program, sweep can be carried out as follows: first, replace g_{kk} with $h_{kk} = -1/g_{kk}$; next, replace the remaining elements g_{jl}=g_{jl} in row and column k with $h_{jk} = g_j l$-$g_{jl} h_{jk}$. and finally, replace the remaining elements $g_{jl} = g_{lj}$ in the other rows and columns by $h_{jl} = g_{jl}$ - $g_{kl} h_{jk}$. This method is efficient both in terms of computation time and memory, because no storage locations other than the matrix itself are necessary. Because both G and H are symmetric, further savings can be achieved by computing and retaining only the upper-triangular portion of the matrix.

Suppose that a $p \times p$ matrix G is partitioned as

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where G_{11} is $p_1 \times p_1$. After sweeping on positions 1, 2,..., p_1, the matrix becomes

$$\text{SWP}[1,2,\ldots,p_1]G = \begin{bmatrix} -G_{11}^{-1} & G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{bmatrix}$$

which is recognizable as a matrix version of (5.29). The notation SWP[1,2,...,p_1] indicates successive application of (5.29),

$$\text{SWP}[1,2,\ldots,p_1]G = \text{SWP}[p_1]\cdots\text{SWP}[2]\text{SWP}[1]G.$$

Sweeps on multiple positions need not be carried out in any particular order, because the sweep operator is commutative,

$$\text{SWP}[k_2]\text{SWP}[k_1]G = \text{SWP}[k_1]\text{SWP}[k_2]G.$$

Sweeping a $p \times p$ matrix G on positions 1, 2,..., p has the effect of replacing G by $-G^{-1}$. This inverse exists if and only if none of the attempted sweeps involve division by zero. When inverting a matrix with sweep, we can also readily obtain the determinant. Let Υ_{k} denote the kth diagonal element of the matrix after it is swept on positions 1, 2,..., $k - 1$,

$$\gamma_k = \left(\text{SWP}[1,2,..,k-1]G\right)_{kk}.$$

Then

$$|G| = \prod_{k=1}^{p} \gamma_k, \tag{5.30}$$

where γ_1 is taken to be g_{11}, the first element of G. Thus the determinant can be found by computing the product of the pivots (i.e. the diagonal elements of the matrix) as they appear immediately before the matrix is swept on them (Dempster, 1969b).

It is also convenient to define a *reverse-sweep* operator that returns a swept matrix to its original form. The reverse-sweep operator, denoted by

$$H = \mathrm{RSW}[k]G,$$

replaces the elements of G with

$$\begin{aligned} h_{kk} &= -1 \quad / \, g_{kk}, \\ h_{jk} &= h_{kl} \;\; = g_{jk} / g_{kk} \text{ for } j \neq k, \\ h_{jl} &= h_{lj} \;\; = g_{jl} - g_{jk} g_{kl} / g_{kk} \text{ for } j \neq k \text{ and } l \neq k. \end{aligned} \tag{5.31}$$

Notice that reverse sweep is remarkably similar to sweep, with the only difference being a minus sign in the calculation of $h_{jk} = h_{kj}$. It is easy to verify that reverse sweep is indeed the inverse of sweep,

$$\mathrm{RSW}[k]\;\mathrm{SWP}[k]\;G = G$$

and that reverse sweep is commutative,

$$\mathrm{RSW}[k_2]\;\mathrm{RSW}[k_1]\;G = \mathrm{RSW}[k_1]\;\mathrm{RSW}[k_2]\;G.$$

Computing alternative parameterizations

From a statistical viewpoint, the sweep operator is highly useful for the following reason: when applied to the parameters of the multivariate normal model, sweep converts a variable from a response to a predictor. Suppose that z is a $p \times 1$ random vector distributed as $N(\mu,\Sigma)$, and we partition it as $z^T = \left(z_1^T, z_2^T\right)$ where z_1 has length p_1. Let us arrange the parameters $\theta = (\mu,\Sigma)$ as a $(p+1) \times (p+1)$ matrix in the following manner,

$$\theta = \begin{bmatrix} -1 & \mu^T \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} -1 & \mu_1^T & \mu_2^T \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{5.32}$$

The reason for placing -1 in the upper-left corner will be explained shortly. To simplify book-keeping, we will allow the row and column indices to run from 0 to p rather than from 1 to $p + 1$, so that the parameters pertaining to the jth variable will appear in row and column *j*. Suppose that we sweep this θ-matrix on positions 1, 2,...,p_1; the result will be, by the matrix analogue of (5.29),

$$\begin{bmatrix} -1-\mu_1^{\mathrm{T}}\Sigma_{11}^{-1}\mu_1 & \mu_1^{\mathrm{T}}\Sigma_{11}^{-1} & \mu_2^{\mathrm{T}}-\mu_1^{\mathrm{T}}\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1}\Sigma_{12} \\ \mu_2-\Sigma_{21}\Sigma_{11}^{-1}\mu_1 & \Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{bmatrix}.$$

Comparing this to (5.24), we see that the last p - p_1 rows and columns contain $\alpha_{2.1}$, $B_{2.1}$, and $\Sigma_{22.1}$, the parameters of the conditional distribution of z_2 given z_1,

$$\mathrm{SWP}[1,\ldots p_1]\theta = \begin{bmatrix} -1-\mu_1^T\Sigma_{11}^{-1}\mu_1 & \mu_1^T\Sigma_{11}^{-1} & \alpha_{2\cdot1}^T \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & B_{2\cdot1}^T \\ \alpha_{2\cdot1} & B_{2\cdot1} & \Sigma_{22\cdot1} \end{bmatrix}$$

Moreover, the upper-left (p_1 + 1) × (p_1 + 1) submatrix contains in swept form the parameters of the marginal distribution of z_1,

$$\begin{bmatrix} -1 & \mu_1^{\mathrm{T}} \\ \mu_1 & \Sigma_{11} \end{bmatrix} = \mathrm{RSW}[1,\ldots,p_1]\begin{bmatrix} -1-\mu_1^{\mathrm{T}}\Sigma_{11}^{-1}\mu_1 & \mu_1^{\mathrm{T}}\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} \end{bmatrix}$$

We have thus shown that ϕ=(μ_1,Σ_{11},$\alpha_{2.1}$,$B_{2.1}$,$\Sigma_{22.1}$), expressed in matrix form as

$$\phi = \begin{bmatrix} -1 & \mu_1^T & \alpha_{2\cdot1}^T \\ \mu_1 & \Sigma_{11} & B_{2\cdot1}^T \\ \alpha_{2\cdot1} & B_{2\cdot1} & \Sigma_{22\cdot1} \end{bmatrix}. \qquad (5.33)$$

can be computed from the θ-matrix by first sweeping the full matrix on positions 1, 2,..., p_1, and then reverse sweeping the upper-left (p_1 + 1) × (p_1 + 1) submatrix on the same positions.

The reason for placing -1 in the upper-left corner of the θ matrix (5.32) is that this matrix can be considered to be already swept on position 0. Notice that if we reverse-sweep θ on position 0, we obtain

$$\mathrm{RSW}[0]\begin{bmatrix} -1 & \mu^T \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} 1 & \mu^T \\ \mu & \Sigma+\mu\mu^T \end{bmatrix}. \qquad (5.34)$$

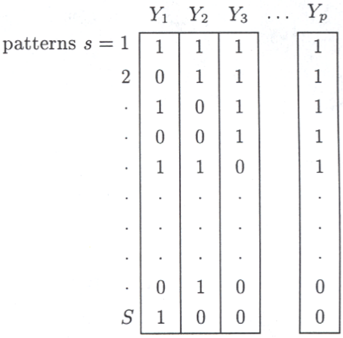
the parameters of the multivariate normal distribution expressed in terms of the first two moments of z about the origin. This unswept version of θ is quite useful because it is the natural representation for computing ML estimates. Suppose that Y is an $n \times p$ data matrix whose rows are independent realizations of the random vector z. If we arrange the sufficient statistics $T_1=Y^T1$ and $T_2=Y^TY$ into a $(p+1) \times (p+1)$ matrix

$$T=[1,Y]^T[1,Y]=\begin{bmatrix} n & T_1^T \\ T_1 & T_2 \end{bmatrix}, \tag{5.35}$$

then the moment equations for ML estimation set (5.34) equal to $n^{-1}T$. Hence the ML estimate of θ may be computed from the sufficient statistics by

$$\hat{\theta}=\text{SWP}[\theta]n^{-1}T.$$

Because ML estimates are invariant under transformations of the parameter, the MLE for an alternative parameterization ϕ can be obtained by sweeping $\hat{\theta}$ on the appropriate positions.



	Y_1	Y_2	Y_3	...	Y_p
patterns $s=1$	1	1	1		1
2	0	1	1		1
·	1	0	1		1
·	0	0	1		1
·	1	1	0		1
·	·	·	·		·
·	·	·	·		·
·	·	·	·		·
·	0	1	0		0
S	1	0	0		0

Figure 5.1. *Matrix of missingness patterns associated with Y with 1 denoting an observed variable and 0 denoting a missing variable.*

5.3 The EM algorithm

When portions of the data matrix Y are missing, ML estimates cannot in general be obtained in closed form; we must resort to iterative computation. The EM algorithm for a multivariate normal data matrix with an arbitrary pattern of missing values was described by Orchard and Woodbury (1972); Beale and Little (1975); Dempster, Laird and Rubin (1977); and Little and Rubin (1987). Because of its usefulness and its similarities to the simulation algorithms that follow, we describe in detail one possible implementation of EM for incomplete multivariate normal data.

5.3.1 Preliminary manipulations

To simplify notation and facilitate computations, it is helpful at the outset to group the rows of Y by their missingness patterns. A matrix of missingness patterns corresponding to Y is shown in Figure 5.1. We will index the missingness patterns by $s = 1, 2,..., S$, where S is the number of unique patterns appearing in the data matrix. The trivial pattern with all variables missing should be omitted from consideration. Rows of Y that are completely missing contribute nothing to the observed-data likelihood and would only slow the convergence of EM by increasing the fractions of missing information (Section 3.3.2).

For book-keeping purposes it will be helpful to define the following quantities. Let R be an $S \times p$ matrix of binary indicators with typical element rsj, where

$$r_{sj} = \begin{cases} 1 \text{ if } Y_j \text{ isobservedinpattern } s, \\ 0 \text{ if } Y_j \text{ ismissinginpattern } s. \end{cases}$$

The matrix R is shown in Figure 5.1. For each missingness pattern s, let O(s) and M(s) denote the subsets of the column labels $\{1, 2,..., p\}$ corresponding to variables that are observed and missing, respectively,

$$O(s) = \left\{ j : r_{sj} = 1 \right\},$$

$$M(s) = \left\{ j : r_{sj} = 0 \right\}.$$

Finally, let I(s) denote the subset of $\{1, 2,..., n\}$ corresponding to the rows of Y that exhibit pattern s. For example, suppose that the data matrix has ten rows with no missing values, and after sorting these rows are labeled 1,..., 10; the first row of R is then (1, 1,..., 1), and

$$\mathcal{O}(1) = \{1, 2, \ldots, p\},$$
$$\mathcal{M}(1) = \varnothing,$$
$$\mathcal{I}(1) = \{1, 2, ..10\}.$$

5.3.2 The E-step

Recall that in the E-step of EM, one calculates the expectation of the complete-data sufficient statistics over $P(Y_{mis}|Y_{obs},\theta)$ for an assumed value of θ. These statistics are of the form $\Sigma_i y_{ij}$ and $\Sigma_i y_{ij} y_{ik}$, so to perform the E-step we need to find the expectations of y_{ij} and $y_{ij} y_{ik}$ over $P(Y_{mis}|Y_{obs},\theta)$.

Because the rows $y_1, y_2,...,y_n$ of Y are independent given θ, we can write

$$P(Y_{mis} \mid Y_{obs}, \theta) = \prod_{i=1}^{n} P\left(y_{i(mis)} \mid y_{i(obs)}, \theta\right),$$

where $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing subvectors of y_i, respectively. The distribution $P(y_{i(mis)}|y_{i(obs)}, \theta)$ is a multivariate normal linear regression of $y_{i(mis)}$ on $y_{i(obs)}$, and the parameters of this regression can be calculated by sweeping the θ–matrix on the positions corresponding to the variables in $y_{i(obs)}$. If row i is in missingness pattern s, then the parameters of $P(y_{i(mis)} |y_{i(obs)}, \theta)$ are contained in SWP[$O(s)$]θ in the rows and columns labeled $M(s)$ Let A denote the swept parameter matrix

$$A = \text{SWP}[\mathcal{O}(s)]\theta,$$

and let a_{jk} denote the (j, k)th element of A, $j, k = 0, 1,..., p$. Using the results of Section 5.2.4, the reader may verify that the first two moments of $y_{i(mis)}$ with respect to $P(Y_{mis}|Y_{obs},\theta)$ are given by

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = a_{oj} + \sum_{\kappa \in O(s)} a_{kj} y_{ik},$$

$$\mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) = a_{jk}$$

for each $i \in I(s)$and j, $k \in M.(s)$ For any $j \in O(s)$, of course, the moments are

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = y_{ij},$$

$$\mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) = 0,$$

because y_{ij} is regarded as fixed. Applying the relation

$$\begin{aligned} E\left(y_{ij} y_{i\kappa} \mid Y_{obs}, \theta\right) = \; & \mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) \\ & + E\left(y_{ij} \mid Y_{obs}, \theta\right) E\left(y_{i\kappa} \mid Y_{obs}, \theta\right) \end{aligned}$$

it follows that

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = \begin{cases} y_{ij} \text{ for } j \in O(s), \\ y_{ij}^{*} \text{ for } j \in M(s), \end{cases}$$

and

$$E\left(y_{ij} y_{i\kappa} \mid Y_{obs}, \theta\right) = \begin{cases} y_{ij} y_{ik} \text{ for } j, k \in O(s), \\ y_{ij}^{*} y_{ik} \text{ for } j \in M(s), k \in \mathrm{O}(s), \\ \alpha_{jk} + y_{ij}^{*} y_{ik}^{*} \text{ for } j, k \in M(s), \end{cases}$$

where

$$y_{ij}^{*} = \alpha_{oj} + \sum_{k \in O(s)} \alpha_{kj} y_{ik}. \qquad (5.36)$$

The E-step consists of calculating and summing these expected values of y_{ij} and $y_{ij} y_{ik}$ over i for each j and k. The output of an E-step can then be written as $E(T | Y_{obs}, \theta)$ where T is the matrix of complete-data sufficient statistics

$$T = \begin{bmatrix} n & 1^{T} Y \\ Y^{T} 1 & Y^{T} Y \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} n & y_{i1} & y_{i2} & \cdots & y_{ip} \\ & y_{i1}^{2} & y_{i1} y_{i2} & \cdots & y_{i1} y_{ip} \\ & & y_{i2}^{2} & \cdots & y_{i2} y_{ip} \\ & & & \ddots & \vdots \\ & & & & y_{ip}^{2} \end{bmatrix}.$$

The elements below the diagonal are not shown and may be omitted from the calculations because they are redundant.

Notice that the matrix A=SWP[$O(s)$]θ needed for the E-step depends on the missingness pattern s, and thus in practice the elements of $E(T|Y_{obs},\theta)$ must be calculated by first summing expected values of y_{ij} and $y_{ij}y_{ik}$ for $i\in I(s)$, and then summing across patterns s = 1, 2,..., S, with a new A-matrix being calculated for each missingness pattern.

5.3.3 Implementation of the algorithm

Once $E(T|Y_{obs},\theta)$ has been found, carrying out the M-step is relatively trivial. For a given value of T the complete-data MLE is $\hat{\theta} = \text{SWP}[0]n^{-1}T,$ and the M-step merely carries out this same operation on $E(T|Y_{obs},\theta)$ rather than T. A single iteration of EM can thus be written succinctly as

$$\theta^{(t+1)} = \text{SWP}[0]n^{-1}E\left(T\,|\,Y_{obs},\theta^{(t)}\right). \tag{5.37}$$

In principle the EM algorithm for incomplete multivariate normal data is completely defined by (5.37), but from a practical standpoint we should still consider how to implement the algorithm in an efficient manner. It is beneficial to keep both processing time and memory usage down, but trade-offs between the two are inevitable; one can always reduce processing time at the expense of additional memory by storing rather than recomputing quantities that must be used repeatedly. The implementation suggested here stores rather than recomputes the portions of $E(T|Y_{obs},\theta)$ that do not depend on θ and thus remain the same for every E-step. This method may not be optimal for any particular dataset, but it is not difficult to program and seems to perform well in a wide variety of situations.

Observed and missing parts of the sufficient statistics

We can express the matrix T as the sum of matrices corresponding to the individual missingness patterns. Let

$$\mathrm{T}(s)=\begin{bmatrix} n_s & \Sigma y_{i1} & \Sigma y_{i2} & \cdots & \Sigma y_{ip} \\ & \Sigma y_{i1}^2 & \Sigma y_{i1}y_{i2} & \cdots & \Sigma y_{i1}y_{ip} \\ & & \Sigma y_{i2}^2 & \cdots & \Sigma y_{i2}y_{ip} \\ & & & \ddots & \vdots \\ & & & & y_{ip}^2 \end{bmatrix},$$

where all sums are taken over $i\in I(s)$, and $\mathrm{n}_s=\Sigma_{i\in I(s)}$ is the sample size in missingness pattern s; then

$$T=\sum_{s=1}^{S}T(s).$$

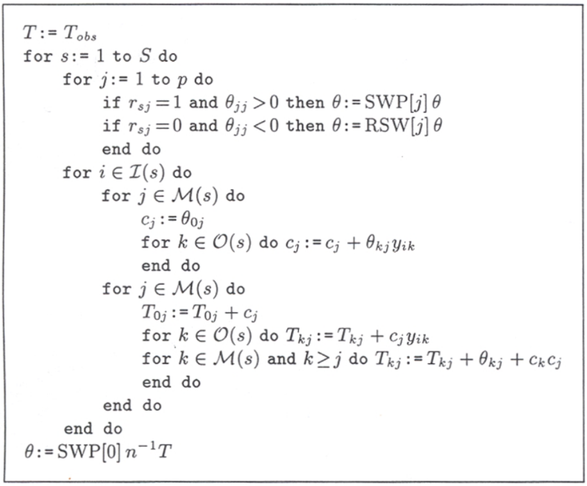
Each $T(s)$ can be further partitioned into an observed part and a missing part. Notice that the elements of $T(s)$ in the rows and columns labeled $M(s)$ are functions of Y_{mis} and perhaps Y_{obs} whereas the remaining elements of $T(s)$ are functions of Y_{obs} only. Define a new matrix T_{mis} (s) which has the same elements as $T(s)$ in the rows and columns labeled $M(s)$, but with all other elements set to zero, and define $T_{obs}(s)$ to be $T(s)$ $-T_{mis}(s)$. For example, consider a dataset with p = 3 variables, and suppose that missingness pattern s has Y_1 and Y_3 observed but Y_2 missing; then

$$T_{obs}(s)=\begin{bmatrix} n_s & \Sigma y_{i1} & 0 & \Sigma y_{i3} \\ & \Sigma y_{i1}^2 & 0 & \Sigma y_{i1}y_{i3} \\ & & 0 & 0 \\ & & & \Sigma y_{13}^2 \end{bmatrix},$$

$$T_{mis}(s)=\begin{bmatrix} 0 & 0 & \Sigma y_{i2} & 0 \\ & 0 & \Sigma y_{i1}y_{i2} & 0 \\ & & \Sigma y_{i2}^2 & \Sigma y_{i2}y_{i3} \\ & & & 0 \end{bmatrix},$$

where all sums are taken over $i\in I(s)$. Finally, define

$$T_{obs}=\sum_{s=1}^{S}T_{obs}(s)\,\text{and}\;T_{mis}=\sum_{s=1}^{S}T_{mis}(s),$$



```
$T := T_{obs}$
for $s := 1$ to $S$ do
    for $j := 1$ to $p$ do
        if $r_{sj} = 1$ and $\theta_{jj} > 0$ then $\theta := \mathrm{SWP}[j]\,\theta$
        if $r_{sj} = 0$ and $\theta_{jj} < 0$ then $\theta := \mathrm{RSW}[j]\,\theta$
        end do
    for $i \in \mathcal{I}(s)$ do
        for $j \in \mathcal{M}(s)$ do
            $c_j := \theta_{0j}$
            for $k \in \mathcal{O}(s)$ do $c_j := c_j + \theta_{kj} y_{ik}$
            end do
        for $j \in \mathcal{M}(s)$ do
            $T_{0j} := T_{0j} + c_j$
            for $k \in \mathcal{O}(s)$ do $T_{kj} := T_{kj} + c_j y_{ik}$
            for $k \in \mathcal{M}(s)$ and $k \geq j$ do $T_{kj} := T_{kj} + \theta_{kj} + c_k c_j$
            end do
        end do
    end do
$\theta := \mathrm{SWP}[0]\, n^{-1} T$
```

Figure 5.2. *Single iteration of EM for incomplete multivariate normal data, written in pseudocode*

so that $T = T_{obs} + T_{mis}$. The E-step may then be written

$$\begin{aligned} E(T \mid Y_{obs}, \theta) &= T_{obs} + E(T_{mis} \mid Y_{obs}, \theta) \\ &= \sum_{s=1}^{S} T_{obs}(s) + \sum_{s=1}^{S} E(T_{mis}(s) \mid Y_{obs}, \theta). \end{aligned}$$

The elements of T_{obs} can be calculated once at the outset of the program and stored for all future iterations of EM.

An implementation in pseudocode

One possible implementation of an iteration of EM is shown in Figure 5.2. It is written in *pseudocode*, a shorthand language that can be understood by anyone with programming experience and is easily converted into standard languages like Fortran or C. In this pseudocode, the symbol `: =' indicates the operation of assignment; for example, `$a := b$' means `set a equal to b.' This implementation requires two $(p+1) \times (p+1)$ matrix workspaces: T, into which the expected sufficient statistics are accumulated, and θ, which holds the current estimate of the parameter. For simplicity, the rows and columns of these matrices are labeled from 0 to p rather than from 1 to $p + 1$. In addition, a single vector of length p, denoted by $c = (c_1,...,c_p)$, is needed as a temporary workspace to hold the values of y_i^*j given by (5.36). The iteration begins by setting T equal to T_{obs} which we assume has already been computed. The expectations of y_{ij} and $y_{ij}y_{ik}$ that contribute to T_{mis} are then calculated and added into T, one missingness pattern at a time. In order to calculate these expectations within a missingness pattern s, the θ-matrix must be put into the required SWP[O(s)] condition; for this, we use the convenient book-keeping device that a diagonal element θ_{jj} is negative if and only if θ has been swept on position j. Finally, after the expected sufficient statistics are fully accumulated into T, the new parameter estimate is calculated and stored in θ in preparation for the next iteration.

For efficiency, the code in Figure 5.2 does not calculate the off-diagonal elements of T more than once. If θ and T are stored as two-dimensional arrays, then only the upper-triangular portions should be used, and T_{jk} or θ_{jk} should be interpreted as the (j, k)th element if $\leq k$ or the (k, j)th element if $j > k$. Memory requirements can be reduced by retaining only the upper-triangular parts of T and 0 in packed storage. To reduce the impact of rounding errors, T, θ, and c should be stored in double precision. Rounding errors can also be reduced by centering and scaling the columns of Y at the outset; for example, we could transform the observed data in each column of Y to have mean zero and unit variance before running EM. If the data are centered and scaled, however, we should remember that θ will be expressed on this transformed scale, and for interpretability we may need to transform the estimate of θ back to the original scale at the end of the program.

Starting values

EM requires a starting value $\theta^{(0)}=(\mu^{(0)},\Sigma^{(0)})$for the first iteration. Any starting value may be used provided that $\Sigma^{(0)}$ is positive definite, but in practice it helps to choose a value that is likely to be close to the mode. Several choices for starting values are described by Little and Rubin (1987). The mean vector and covariance matrix calculated only from the completely observed rows of Y may work well, provided that there are at least $p + 1$ such rows. Another easy method is to use the observed data from each variable to supply starting values for the means and variances, and set the initial correlations to zero; if the columns of Y have been centered and scaled at the outset to have mean 0 and variance 1, then this corresponds to taking $\mu^{(0)}=(0,0,...,0)^T$ and $\Sigma^{(0)}=I$.

Unless the fractions of missing information for some components of θ are very high, the choice of starting value is usually not crucial; when the missing information is low to moderate, the first few iterations of EM tend to bring θ to the vicinity of the mode from any sensible starting value. When writing a program for general use, it is helpful to give the user the option of supplying a starting value, because restarting EM from a variety of locations helps to diagnose unusual features of the observed-data likelihood, such as ridges and multiple modes.

Estimates on the boundary

It sometimes happens, particularly with sparse datasets, that the observed-data likelihood function increases without limit as θ approaches the boundary of the parameter space (i.e. as Σ approaches a singular matrix). When this occurs, the EM algorithm may behave in a variety of ways. In some problems, the elements of θ stabilize and EM appears to converge to a solution on the boundary. In other problems, the program halts due to numeric overflow or attempted division by zero. In yet other problems, the sweeps required for the E-step become numerically unstable as the iterates approach the boundary, and substantial rounding errors are introduced. We have found that these rounding errors sometimes `deflect' θ away from the boundary, causing a sudden large drop in likelihood from one iteration to the next. The iterates may approach the boundary for a number of steps, deflect away, approach again, and deflect away again in a recurring fashion. If the elements of θ do not appear to have converged after a large number of iterations, then it is advisable to monitor both the loglikelihood (Section 5.3.5) and some aspect of Σ (e.g. the determinant, or the ratio of the largest eigenvalue to the smallest) to determine whether the iterates are approaching the boundary.

When an ML estimate falls on the boundary, it is often helpful to apply a ridge prior and use EM to find the posterior mode as described below.

5.3.4 EM for posterior modes

This EM algorithm can be easily altered to compute a mode of the observed-data posterior distribution rather than an MLE. As discussed in Section 3.2.3, the E-step is no different; only the M-step needs to be modified. The exact form of this modification will depend on the prior distribution applied to θ.

Priors for incomplete data

At this point, it is worthwhile to consider what prior distributions may be appropriate for an incomplete dataset. Because a prior distribution by definition reflects one's state of knowledge about θ before any data are observed, the fact that some data are missing should from a strictly Bayesian viewpoint have no effect whatsoever on the choice of a prior. To the Bayesian purist, any prior that is appropriate for complete data will be equally appropriate for incomplete data. Most statisticians would agree, however, that choosing a prior distribution (including its analytic form) purely by introspection can be difficult, and in practice most priors are chosen at least partly for computational convenience. The normal inverted-Wishart family of prior distributions, described in Sections 5.2.2 and 5.2.3, is computationally convenient for the EM and data augmentation algorithms in this chapter. In general, this family is not conjugate when data are incomplete; the observed data posterior $P(\theta \mid Y_{obs})$ under a normal inverted-Wishart prior is tractable only in special cases. Yet EM and data augmentation are both easy to implement under this family of priors, because the simplicity of these algorithms depends upon the tractability of the complete-data problem.

When prior information about θ is scanty, we suggest that the customary diffuse prior for complete data,

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)},$$

may also be reasonable when some data are missing. Recall from Section 5.2.3 that one important justification for this prior with complete data is that Bayesian and frequents inferences about p coincide. This result does not immediately generalize to incomplete data, but limited experience suggests that Bayesian inferences under this prior may also be approximately valid from a frequents point of view. Little (1988) reports that in the case of bivariate datasets with missing values on one variable generated by an ignorable mechanism, this prior leads to Bayesian inferences about μ that are well-calibrated; the HPD regions tend to have frequency coverage close to the nominal levels. Because this prior treats the variables Y_1, Y_2,..., Y_p in a symmetric fashion, we conjecture that similar results may hold for more complicated multivariate scenarios as well.

When data are sparse and certain aspects of $\sum$ are poorly estimated, we suggested in Section 5.2.3 that a useful prior for complete data was the limiting form of the normal inverted-Wishart with τ=0, m=$\in$ for some $\in$ > 0, and Λ^{-1}=$\in$Diag S, where S is the complete-data sample covariance matrix. With incomplete data S cannot be calculated, but a useful substitute is the matrix with diagonal elements equal to the sample variances among the observed values in each column of Y. This prior effectively smooths the variances in $\sum$ toward the observed-data variances and the correlations toward zero. If the observed data in each column of Y have been scaled at the outset of the program to have unit variances, then this prior will simply take Λ^{-1}=$\in I$.

Modications to the M-step

The joint mode of the normal inverted-Wishart distribution,

$$\begin{aligned} \mu \mid \Sigma &\sim N\left(\mu_0, \tau^{-1}\Sigma\right), \\ \Sigma &\sim W^{-1}(m, \Lambda), \end{aligned}$$

is achieved at μ_0 and $(m+p+2)^{-1}\Lambda^{-1}$ for μ and Σ, respectively (Section 5.2.2). Thus the complete-data posterior mode for $\theta = (\mu, \Sigma)$ under the normal inverted-Wishart prior with hyperparameters $(\tau, m, \mu_0, \Lambda)$, denoted by $\tilde{\theta} = \left(\tilde{\mu}, \tilde{\Sigma}\right)$, is

$$\tilde{\mu} = \mu_0' \text{ and } \tilde{\Sigma} = \frac{1}{m' + p + 2}(\Lambda')^{-1},$$

where μ_0, m' and Λ' are the updated versions of the hyperparameters given in Section 5.2.2. By reverse-sweeping the mode on position 0 and equating the result to a matrix of modified sufficient statistics,

$$\text{RSW}[0]\begin{bmatrix} -1 & \tilde{\mu}^T \\ \tilde{\mu} & \tilde{\Sigma} \end{bmatrix} = \begin{bmatrix} 1 & \tilde{\mu}^T \\ \tilde{\mu} & \tilde{\Sigma} + \tilde{\mu}\tilde{\mu}^T \end{bmatrix} = n^{-1}\begin{bmatrix} n & \tilde{T}_1^T \\ \tilde{T}_1 & \tilde{T}_2 \end{bmatrix}$$

the mode can be computed as if it were an ML estimate based on $\tilde{T}_1$ and $\tilde{T}_2$ rather than T_1 and T_2. Solving for $\tilde{T}_1$ and $\tilde{T}_2$ and substituting expressions for the updated hyperparameters gives

$$\tilde{T}_1 = \left(\frac{n}{n+\tau}\right)T_1 + \left(\frac{\tau}{n+\tau}\right)n\mu_0$$

and

$$\tilde{T}_2 = \frac{n}{n+m+p+2}\left(T_2 - \tfrac{1}{n}T_1T_1^T + \Lambda^{-1} + A\right) + \tfrac{1}{n}\tilde{T}_1\tilde{T}_1^T$$

as the modified sufficient statistics, where

$$A = \frac{\tau}{n(\tau+n)}\left(T_1 - n\mu_0\right)\left(T_1 - n\mu_0\right)^T.$$

To modify the EM algorithm shown in Figure 5.2 to compute a posterior mode rather than an MLE, we need only to replace the expected sufficient statistics T_1 and T_2 in the workspace T by the modified versions $\tilde{T}_1$ and $\tilde{T}_2$ immediately before executing the final step $\theta := \text{SWP}[0]n^{-1}T$.

5.3.5 Calculating the observed-data loglikelihood

One of the great advantages of the EM algorithm is that it never requires calculation of the observed-data loglikelihood function or its derivatives. The observed-data likelihood for this problem, discussed in Example 3 of Section 2.3.2, or its logarithm $l(\theta|Y_{obs})$, would be very tedious to differentiate or maximize by gradient-based methods. Evaluation of $l(\theta|Y_{obs})$, at a specific value of θ, however, is not overwhelmingly difficult; the computations required for a single evaluation are comparable to those needed for a single iteration of EM.

It follows from (2.10) that the observed data-loglikelihood function may be written as

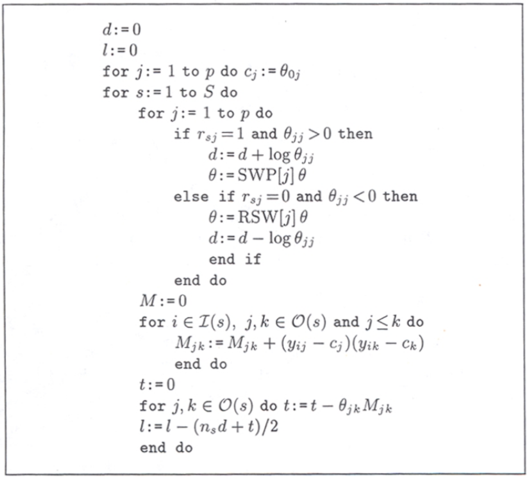
$$\sum_{s=1}^{S}\sum_{i\in I(s)}\left\{-\tfrac{1}{2}\log\left|\Sigma_{s}^{*}\right|-\tfrac{1}{2}\left(y_{i(obs)}-\mu_{s}^{*}\right)^{T}\Sigma_{s}^{*-1}\left(y_{i(obs)}-\mu_{s}^{*}\right)\right\},$$

where $y_{i(obs)}$ denotes the observed part of y_i and μ_s^* and Σ_s^* denote the subvector of μ and the submatrix of Σ, respectively, that pertain to the variables that are observed in pattern s. An equivalent but computationally more convenient expression is

$$l(\theta\mid Y_{obs})=\sum_{s=1}^{S}\left\{-\tfrac{n_s}{2}\log\left|\Sigma_{s}^{*}\right|-\tfrac{1}{2}\,tr\Sigma_{s}^{*-1}M_{s}\right\}, \tag{5.38}$$

where n_s is the number of observations in missingness pattern s and

$$M_{s}=\sum_{i\in I(s)}\left(y_{i(obs)}-\mu_{s}^{*}\right)\left(y_{i(obs)}-\mu_{s}^{*}\right)^{T}.$$



```
$d := 0$
$l := 0$
for $j := 1$ to $p$ do $c_j := \theta_{0j}$
for $s := 1$ to $S$ do
    for $j := 1$ to $p$ do
        if $r_{sj} = 1$ and $\theta_{jj} > 0$ then
            $d := d + \log \theta_{jj}$
            $\theta := \mathrm{SWP}[j]\,\theta$
        else if $r_{sj} = 0$ and $\theta_{jj} < 0$ then
            $\theta := \mathrm{RSW}[j]\,\theta$
            $d := d - \log \theta_{jj}$
            end if
        end do
    $M := 0$
    for $i \in \mathcal{I}(s)$, $j, k \in \mathcal{O}(s)$ and $j \le k$ do
        $M_{jk} := M_{jk} + (y_{ij} - c_j)(y_{ik} - c_k)$
        end do
    $t := 0$
    for $j, k \in \mathcal{O}(s)$ do $t := t - \theta_{jk} M_{jk}$
    $l := l - (n_s d + t)/2$
    end do
```

Figure 5.3. *Calculation of observed-data loglikelihood function.*

Pseudocode for calculating l($\theta|Y_{obs}$) is shown in Figure 5.3. This algorithm requires a $p \times p$ matrix workspace M to hold values of $M_{<i>s}$</i>, and a $p \times 1$ vector c for temporary storage of μ. The constants d and t hold $\log\left|\Sigma_s^*\right|$ and $tr\Sigma_s^{*-1}M_s$, respectively, and after execution the loglikelihood value is contained in 1. This program modifies the parameter matrix θ; if necessary, however, the single line

$$\theta := \mathrm{RSW}[\mathcal{O}(S)\theta$$

may be added at the end of the program, which will return θ to its original state except for rounding errors.

Notice that the algorithm for evaluating L($\theta|Y_{obs}$) bears a strong resemblance to a single step of EM. An obvious question to ask is whether the two sets of code can be combined, so that an evaluation of the loglikelihood is efficiently woven into EM itself This is certainly possible, but subject to the following caveats. First, the loglikelihood would have to be evaluated at the parameter estimate from the *previous* iteration; that is, we would have to evaluate l($\theta^{(t)}$ $|Y_{obs}$) as we computed $\theta^{(t+1)}$. Second, notice that a loglikelihood evaluation requires accumulation of the observed parts of the complete-data sufficient statistics, rather than the expected values of the missing parts. Recall that the EM code in Figure 5.2 assumes that T_{obs} the portion of the expected value of T that does not change over the iterations, has already been computed and stored at the outset of the program. Evaluation of the observed-data loglikelihood, however, requires access to the individual matrices $T_{obs}(s)$ for $s = 1,2,\ldots, S$ which could be very cumbersome to store. If, as in Figure 5.3, the matrices $T_{obs}(s)$ are not stored but effectively recomputed at each iteration, then the proportionate reductions in computing time achieved by combining the two algorithms over running them separately would not be overwhelming.

When EM is used to find a posterior mode rather than an MLE, the function that is guaranteed to be non-decreasing at each iteration is no longer the observed-data likelihood but the observed-data posterior density. The logarithm of the observed-data posterior density is

$$\log P(\theta \mid Y_{obs}) = l(\theta \mid Y_{obs}) + \log \pi(\theta),$$

where unnecessary normalizing constants have been omitted. Thus the log-posterior density may be evaluated by adding log $\pi(\theta)$ to the result of the algorithm in Figure 5.3. Under a normal inverted-Wishart prior with hyperparameters $(\tau, m, \mu_0, \Lambda)$, this additional term is

$$\log \pi(\theta) = -\frac{m+p+2}{2}\log|\Sigma| - \tfrac{1}{2}\, tr\Sigma^{-1} M_0,$$

where

$$M_0 = \Lambda^{-1} + \tau(\mu - \mu_0)(\mu - \mu_0)^T,$$

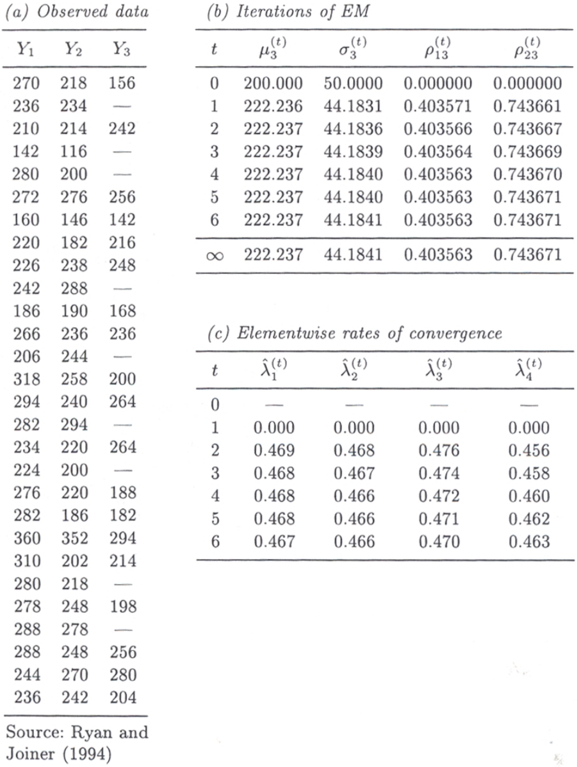
and unnecessary constants have again been omitted.

5.3.6 Example: serum-cholesterol levels of heart-attack patients

Ryan and Joiner (1994, Table 9.1) report serum-cholesterol levels for $n = 28$ patients treated for heart attacks at a Pennsylvania medical center. For all patients in the sample, cholesterol levels were measured 2 days and 4 days after the attack. For 19 of the 28 patients, an additional measurement was taken 14 days after the attack. The data are displayed in Table 5.1 (a), with readings at 2, 4 and 14 days denoted by Y_1, Y_2 and Y_3, respectively.

Regarding the complete data as a random sample from a trivariate normal distribution, we applied EM to find the observed-data ML estimates of the nine parameters in $\theta=(\mu,\Sigma)$ (ML estimates for this dataset could also be calculated noniteratively; see Section 6.5). Denote the elements of μ and Σ by μ_j and $\sigma_{j\kappa}$,

Table 5.1. *EM algorithm applied to cholesterol levels for heart-attack patients measured 2, 4 and 14 days after attack*



(a) Observed data

Y_1	Y_2	Y_3
270	218	156
236	234	—
210	214	242
142	116	—
280	200	—
272	276	256
160	146	142
220	182	216
226	238	248
242	288	—
186	190	168
266	236	236
206	244	—
318	258	200
294	240	264
282	294	—
234	220	264
224	200	—
276	220	188
282	186	182
360	352	294
310	202	214
280	218	—
278	248	198
288	278	—
288	248	256
244	270	280
236	242	204

Source: Ryan and Joiner (1994)

(b) Iterations of EM

t	$\mu_3^{(t)}$	$\sigma_3^{(t)}$	$\rho_{13}^{(t)}$	$\rho_{23}^{(t)}$
0	200.000	50.0000	0.000000	0.000000
1	222.236	44.1831	0.403571	0.743661
2	222.237	44.1836	0.403566	0.743667
3	222.237	44.1839	0.403564	0.743669
4	222.237	44.1840	0.403563	0.743670
5	222.237	44.1840	0.403563	0.743671
6	222.237	44.1841	0.403563	0.743671
∞	222.237	44.1841	0.403563	0.743671

(c) Elementwise rates of convergence

t	$\hat{\lambda}_1^{(t)}$	$\hat{\lambda}_2^{(t)}$	$\hat{\lambda}_3^{(t)}$	$\hat{\lambda}_4^{(t)}$
0	—	—	—	—
1	0.000	0.000	0.000	0.000
2	0.469	0.468	0.476	0.456
3	0.468	0.467	0.474	0.458
4	0.468	0.466	0.472	0.460
5	0.468	0.466	0.471	0.462
6	0.467	0.466	0.470	0.463

respectively, for j, k = 1, 2, 3, and let $\rho_{j\kappa} = \sigma_{j\kappa}(\sigma_{jj}\sigma_{\kappa\kappa})^{-1/2}$ denote the correlations. From starting values chosen based on a crude guess, $\mu^{(o)}=(200,200,200)^T$ and $\Sigma^{(o)}=(50)^2 I$, convergence within four significant digits to

$$\hat{\mu} = \begin{bmatrix} 253.9 \\ 230.6 \\ 222.2 \end{bmatrix}, \hat{\Sigma} = \begin{bmatrix} 2195 & 1455 & 835.4 \\ & 2127 & 1515 \\ & & 1953 \end{bmatrix}$$

was achieved in just three iterations. Because no data, e missing for Y_1 or Y_2, the five parameters $(\mu_1,\mu_2,\sigma_{11},\sigma_{22},\rho_{12})$ converge in a single step regardless of the starting value. Iterates of the four remaining parameters, expressed as μ_3, $\sigma_3=\sqrt{\sigma_{33}}$, ρ_{13} and ρ_{23}, are displayed to six significant digits in Table 5.1 (b).

For estimation of θ, the iterations beyond t = 4 are superfluous because precision beyond three or four digits is rarely necessary. As discussed in Section 3.3.4, however, these additional iterations can be used to estimate elementwise rates of convergence, which are typically equal to the largest fraction of missing information. Elementwise rates of convergence for the four parameters that do not converge in one step, estimated using (3.27), are displayed in Table 5.1 (c). These estimates, which are all close to 47%, do not measure the individual rates of missing information for the four parameters μ_3, σ_3, ρ_{13} and ρ_{23}; rather, they pertain to the function of θ for which the rate of missing information is highest.

Notice that the 47% rate of missing information is somewhat higher than the 9/28 = 32% rate of missing observations for Y_3. Because we know that the parameters pertaining to the joint distribution of (Y_1, Y_2) have no missing information, the 47% rate must pertain to some function of the parameters of the regression of Y_3 on Y_1 and Y_2. It is instructive to consider why the largest rate of missing information exceeds the rate of missing observations for Y_3. A hint is provided by the scatterplot of Y_1 versus Y_2 displayed in

Figure 5.4 (a). The cases having missing values for Y_3 tend to be slightly farther, on average, from the center of the (Y_1, Y_2) distribution than do the cases for which Y_3 is observed. Because they are farther from the center, they exert more influence on the estimates of the regression parameters. A well known measure of influence in linear regression models is provided by the *leverage values*, the diagonal elements of the hat matrix (e.g. Draper and Smith, 1981).

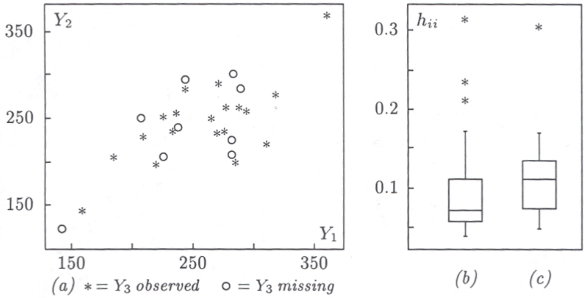


Figure 5.4. *(a) Scatterplot of Y_1 versus Y_2 for all cases, and boxplots of leverage values h_{ii} for cases having (b) Y_3 observed and (c) Y_3 missing.*

The hat matrix for linear regression is defined to be

$$H{=}X(X^TX)^{-1}XT,$$

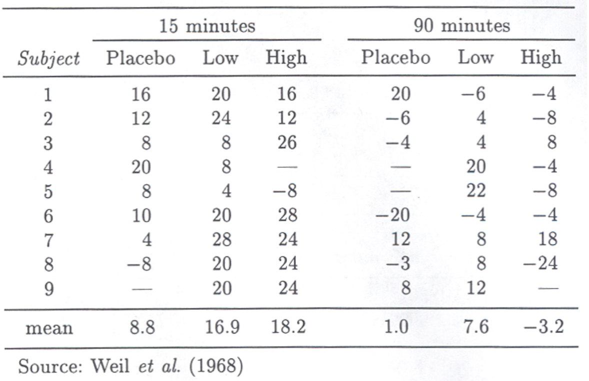
where X is the matrix of predictor variables, in this case a 28 × 3 matrix containing the observed values of Y_1 and Y_2 and the column vector $1=(1,1,...,)^T$. Boxplots of the diagonal elements hii of H for the cases having Y_3 observed and the cases having Y_3 missing are shown in Figures 5.4 (b) and (c), respectively. The incomplete cases tend to have slightly higher values of h_{ii} and thus exert greater influence on an average, per-case basis over the estimates of the regression parameters.

The parameters of greatest interest in this problem appear to be functions of μ, such as comparisons or contrasts among μ_1, μ_2 and μ_3. Although the rate of missing observations for Y_3 is 32%, we might conjecture that the rate of missing information for μ_3 or a contrast involving μ_3 is substantially lower, because of the high correlations between Y_3 and the completely observed variables Y_1 and Y_2. The rate of missing information for μ_3, a contrast involving μ_3 or any other function of θ may be estimated in a straightforward manner by multiple imputation; see Section 6.2.1.

5.3.7 Example: changes in heart rate due to marijuana use

Weil et al. (1968) describe a pilot study to investigate the clinical and psychological effects of marijuana use in human subjects. Nine healthy male subjects, all of whom claimed never to have used marijuana before, received doses in the form of cigarettes of uniform size. Each subject received each of the three treatments (low dose, high dose and placebo) and the order of treatments within subjects was balanced in a replicated 3 × 3 Latin square. Changes in heart rate for the $n = 9$ subjects measured 15 and 90 minutes after the smoking session are displayed in Table 5.2. Because the article does not specify the order in which the treatments were given to the individual subjects, we will ignore this feature of the data and proceed as if the order effects are negligible.

Table 5.2. *Change in heart rate recorded 15 and 90 minutes after marijuana use, measured in beats per minute above baseline*



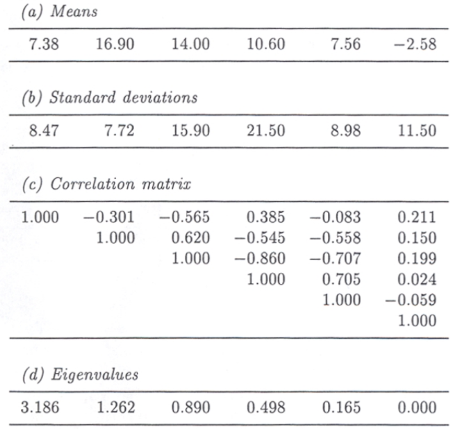
	15 minutes			90 minutes		
Subject	Placebo	Low	High	Placebo	Low	High
1	16	20	16	20	−6	−4
2	12	24	12	−6	4	−8
3	8	8	26	−4	4	8
4	20	8	—	—	20	−4
5	8	4	−8	—	22	−8
6	10	20	28	−20	−4	−4
7	4	28	24	12	8	18
8	−8	20	24	−3	8	−24
9	—	20	24	8	12	—
mean	8.8	16.9	18.2	1.0	7.6	−3.2

Source: Weil *et al.* (1968)

At first glance, it appears that missing data are only a minor problem here; only 5 of the 54 data values are missing. Yet, the EM algorithm converges very slowly. Depending on the starting values and convergence criterion, several hundred iterations may be needed to obtain convergence. The elementwise rates of convergence indicate that the largest fraction of missing information is approximately 97%. Moreover, the ML estimate of θ lies on the boundary of the parameter space. The ML estimates of the means, standard deviations and correlations are displayed in Table 5.3, along with the eigenvalues of the estimated correlation matrix. The smallest eigenvalue is zero to three decimal places, indicating that the estimated covariance matrix is singular or nearly so.

Why do so few missing values create such difficulty in this example?

Table 5.3. *ML estimates of means, standard deviations and correlations for the columns of Table 5.2, with eigenvalues of the estimated correlation matrix*



(a) Means

7.38	16.90	14.00	10.60	7.56	−2.58

(b) Standard deviations

8.47	7.72	15.90	21.50	8.98	11.50

(c) Correlation matrix

1.000	−0.301	−0.565	0.385	−0.083	0.211
	1.000	0.620	−0.545	−0.558	0.150
		1.000	−0.860	−0.707	0.199
			1.000	0.705	0.024
				1.000	−0.059
					1.000

(d) Eigenvalues

3.186	1.262	0.890	0.498	0.165	0.000

There are two primary reasons. First, the incomplete cases appear to be very influential. A comparison of the ML estimates of the means in Table 5.3 (a) with the means of the observed data in the columns of Table 5.2 is quite revealing. The large discrepancy for the fourth column (10.6 versus 1.0) demonstrates that a disproportionate amount of information about the mean for that column is provided by subjects 4 and 5. Further examination of Table 5.2 reveals that these two subjects have rather extreme values in some of the other columns, which gives them high leverage. When these two subjects are deleted, EM converges rapidly and the estimated largest fraction of missing information drops to 45%.

A second reason why this example is problematic is that the complete-data estimation problem is poorly conditioned. The number of subjects $n = 9$ is not much greater than the number of variables $p = 6$. When n and p are nearly equal, it becomes likely that certain linear combinations of the columns of Y will show little or no variability, particularly when the columns are correlated. The multivariate normal model for this example has 27 parameters, too many to be estimated well from a dataset of this size even with complete data. Although certain aspects of θ are poorly estimated, however, we can still make reasonable inferences about the parameters of interest; see Section 5.4.4.

5.4 Data augmentation

5.4.1 The I-step

Data augmentation for incomplete multivariate normal data is remarkably similar to the EM algorithm. The deterministic E- and M-steps are replaced by stochastic I- and P-steps, respectively, where the I-step simulates

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right),$$

and the P-step simulates

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right).$$

Because the rows y_1, $y_2,\ldots,y_{\mathrm{n}}$ of Y are conditionally independent given θ, the I-step is carried out by drawing

$$y_{i(mis)}^{(t+1)} \sim P\left(y_{i(mis)} \mid y_{i(obs)}, \theta^{(t)}\right)$$

independently for $i = 1, 2,\ldots, n$. As discussed in Section 5.3.2, if row i is in missingness pattern s then the conditional distribution of $y_{i(mis)}$ given $y_{i(obs)}$ and θ is multivariate normal with means

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = a_{0j} + \sum_{k \in O(s)} a_{kj} y_{ik} \qquad (5.39)$$

and covariances

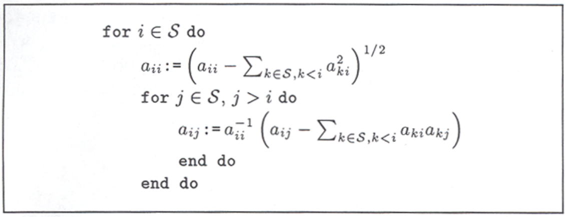
$$\mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) = a_{jk} \qquad (5.40)$$

for $j,k \in M(s)$, where a_{jk} denotes an element of the matrix

$$A = \mathrm{SWP}[O(s)]\theta. \qquad (5.41)$$

Thus the I-step of data augmentation involves nothing more than the independent simulation of random normal vectors for each row of the data matrix, with means and covariances given by (5.39) and (5.40).

A convenient way to simulate random normal vectors within the I-step is to create a *Cholesky factorization* routine that operates



```
for $i \in S$ do
    $a_{ii} := \left(a_{ii} - \sum_{k \in S, k<i} a_{ki}^2\right)^{1/2}$
    for $j \in S$, $j > i$ do
        $a_{ij} := a_{ii}^{-1}\left(a_{ij} - \sum_{k \in S, k<i} a_{ki} a_{kj}\right)$
    end do
end do
```

Figure 5.5. *Calculation of* A:= Chol$_s A$.

on square submatrices of (5.41). The Cholesky factor of a positive definite matrix A, denoted by

$$C = \mathrm{Chol} A,$$

is an upper-triangular matrix of the same dimension of A having the property that $C^{\mathrm{T}}C = A$. To simulate a random vector z from $N(b, A)$, we may take

$$z = b + (\mathrm{Chol} A)^T z_0,$$

where z_0 is a vector of the same length as z containing independent standard normal variates. A typical Cholesky factorization routine operates on the upper-triangular portion of a symmetric matrix, overwriting it with its Cholesky factor, To draw from the distribution of $y_{i(mis)}$ given $y_{i(obs)}$ and θ, however, we need to calculate the Cholesky factor of only the square submatrix of (5.41) corresponding to the rows and columns in M(s). For a set S of row labels of a matrix A, let us use

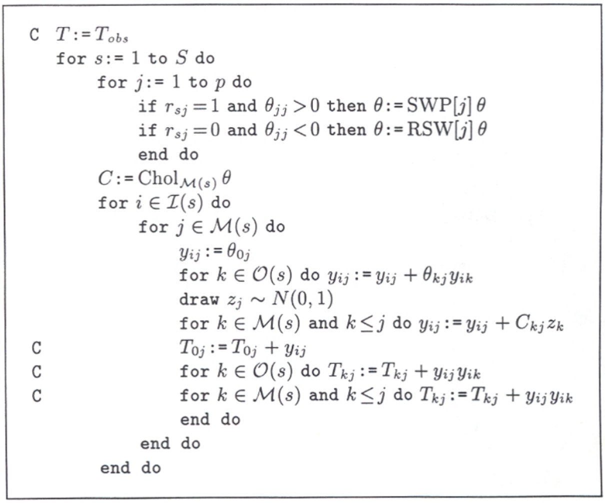
$$A{:=}\mathrm{Chol}_s A \tag{5.42}$$

to indicate the operation that overwrites (the upper triangular portion of) the square submatrix $\{a_{j\kappa}: j,k\in S\}$ with its Cholesky factor, while leaving the remaining elements of A unchanged. A simple algorithm for this operation, adapted from pseudocode given by Thisted (1988, p. 83), is shown in Figure 5.5.

Once the Cholesky factorization is available, the I-step becomes a simple matter of cycling through the missingness patterns $s = 1,\dots, S$, calculating

$$\mathrm{Chol}_{M(s)}\mathrm{SWP}[O(s)]\theta$$

for each s, and simulating $y_{i(mis)}$ for each $i\in I(s)$. An implementation of the I-step is shown in Figure 5.6. The code simulates the



```
C  $T := T_{obs}$
   for $s$ := 1 to $S$ do
      for $j$ := 1 to $p$ do
         if $r_{sj} = 1$ and $\theta_{jj} > 0$ then $\theta := \mathrm{SWP}[j]\,\theta$
         if $r_{sj} = 0$ and $\theta_{jj} < 0$ then $\theta := \mathrm{RSW}[j]\,\theta$
         end do
      $C := \mathrm{Chol}_{\mathcal{M}(s)}\,\theta$
      for $i \in \mathcal{I}(s)$ do
         for $j \in \mathcal{M}(s)$ do
            $y_{ij} := \theta_{0j}$
            for $k \in \mathcal{O}(s)$ do $y_{ij} := y_{ij} + \theta_{kj} y_{ik}$
            draw $z_j \sim N(0,1)$
            for $k \in \mathcal{M}(s)$ and $k \le j$ do $y_{ij} := y_{ij} + C_{kj} z_k$
C           $T_{0j} := T_{0j} + y_{ij}$
C           for $k \in \mathcal{O}(s)$ do $T_{kj} := T_{kj} + y_{ij} y_{ik}$
C           for $k \in \mathcal{M}(s)$ and $k \le j$ do $T_{kj} := T_{kj} + y_{ij} y_{ik}$
            end do
         end do
      end do
```

Figure 5.6. *I-step for incomplete multivariate normal data.*

missing values in Y_{mis} and stores them in the appropriate elements of Y. In addition, the code contains four lines preceded by the single character 'C' which accumulate the simulated complete-data sufficient statistics and store them in a $(p + 1) \times (p + 1)$ matrix workspace T. If the I-step is to be followed by a P-step, then these sufficient statistics will be needed to describe the complete-data posterior distribution of θ. If the I-step will not be followed by a P-step (e.g. if it is the final step of a chain for producing an imputation of Y_{mis}) then these four lines may be omitted. The code in Figure 5.5 requires two temporary workspaces: a $p \times p$ matrix C for storing Cholesky factors, and a $p \times 1$ vector z for holding simulated $N(0, 1)$ variates.

5.4.2 The P-step

Under the prior distributions discussed in Sections 5.2.2 and 5.2.3, the complete data posterior $P(\theta|Y_{obs}, Y_{mis})$ is a normal inverted-Wishart distribution. The P-step of data augmentation, therefore, is merely a simulation of the normal inverted-Wishart distribution,

$$\mu \mid \quad \Sigma \sim N\left(\mu_0, \tau^{-1}\Sigma\right),$$
$$\Sigma \sim W^{-1}(m, \Lambda),$$

for some $(\tau, m, \mu_o, \Lambda)$determined by the prior, the observed data Y_{obs}, and the missing data $Y_{mis}^{(t)}$ imputed at the last I-step. The specific values of $(\tau, m, \mu_o, \Lambda))$ are calculated using the formulas for updating hyperparameters given in Section 5.2.2.

The most obvious way to generate $\Sigma \sim W^{-1}(m, \Lambda)$ is to take $\Sigma = (X^T X)^{-1}$, where X is an $m \times p$ random matrix whose rows are independent draws from $N(0, \Lambda)$. This method cannot be used for non-integer values of m, however, and may be cumbersome for large m because it requires mp random variates. More efficient methods for generating random Wishart matrices are available that require simulation of only p $(p + 1)/2$ random variates. One such method relies on a characterization of the Wishart distribution known as the

Bartlett decomposition (e.g. Muirhead, 1982). If $A \sim W(m,I)$ where I is a $p \times p$ identity matrix and $m \geq p$, then we can write $A = B^T B$ where B is an upper triangular matrix whose elements are independently distributed as

$$b_{jj} \sim \sqrt{X^2_{m-j+1,}} j = 1, \ldots, p, \tag{5.43}$$

$$b_{jk} \sim N(0,1), j < k. \tag{5.44}$$

Suppose that we generate an upper-triangular matrix B according to (5.43)-(5.44), so that $B^T B \sim W(m,I)$, and take

$$M = (B^T)^{-1} C,$$

where C is the Cholesky factor of Λ^{-1}(i.e.$C^T C = \Lambda^{-1}$). Then $\Sigma = M^T M$ will be distributed as $W^{-1}(m,\Lambda)$ because

$$\begin{aligned}\left(M^T M\right)^{-1} &= C^{-1} B^T B \left(C^T\right)^{-1} \\ &\sim W\left(m, \left(C^T C\right)^{-1}\right).\end{aligned}$$

(Here we have made use of the property that $D \sim W(n,\Gamma)$ implies $C^T D C \sim W(n, C^T \Gamma C)$ which follows immediately from the definition of the Wishart distribution.) Moreover, taking

$$\mu = \mu_0 + \tau^{-1/2} M^{\mathrm{T}} z.$$

where $z \sim N(0,I)$ is a $p \times 1$ vector of independent standard normal variates, results in $\mu | \Sigma \sim N(\mu_0, \tau^{-1}\Sigma)$This method requires the inversion of only the triangular matrix B^T, which can be accomplished via a simple backsolving operation. Note that with the exception of M, all matrices used here are either symmetric or triangular, so memory requirements can be reduced by retaining only their upper-triangular portions in packed storage.

5.4.3 Example: cholesterol levels of heart-attack patients

Recall the example of Section 5.3.6 in which cholesterol measurements were recorded for patients 2, 4 and 14 days after heart attack. The EM algorithm converged rapidly with an estimated largest fraction of missing information equal to 47%. We applied data augmentation to this example under the noninformative prior (5.18). Output analysis from preliminary runs suggested that the data augmentation algorithm also converged rapidly. For illustration, we ran a single chain for 1100 iterations starting from the ML estimate of θ, discarded the first 100 iterations, and estimated ACFs for a variety of scalar functions of θ over the remaining 1000 iterations. We deliberately chose functions of θ for which the rates of missing information were thought to be high, including:

1. μ_3 and σ_3, the mean and standard deviation of Y_3, respectively;

2. the parameters of the linear regression of Y_3 on Y_1 and Y_2, including the slopes

$$\begin{bmatrix} \beta_{31\cdot12} \\ \beta_{32\cdot12} \end{bmatrix}^T = \begin{bmatrix} \sigma_{31}\sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix}^{-1},$$

the intercept

$$\beta_{30\cdot12} = \mu_3 - \begin{bmatrix} \sigma_{31}\sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

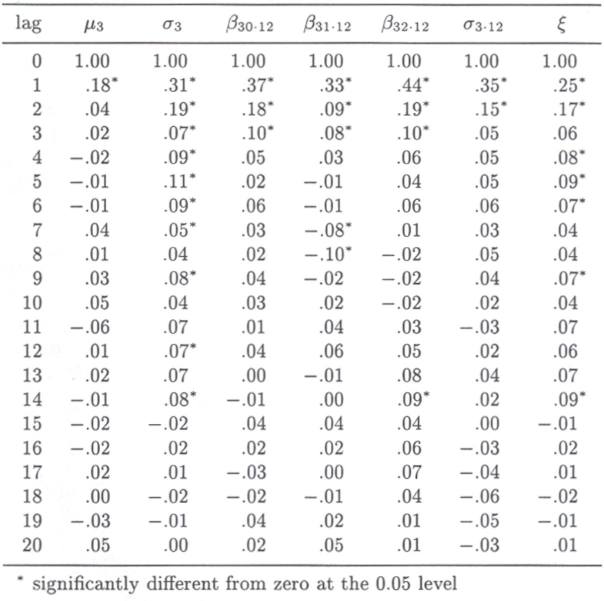
and the residual standard deviation $\sigma_{3\cdot12} = \sqrt{\sigma_{33\cdot12},}$, where

$$\sigma_{33\cdot12} = \sigma_{33} - \begin{bmatrix} \sigma_{31}\sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix};$$

and

3. the worst linear function $\xi=\xi(\theta)$ estimated from the final iterations of EM, as described in Section 4.4.3. This is the inner product of θ and the estimated eigenvector corresponding to the largest eigenvalue of EM's asymptotic rate matrix. Because there are no missing values on Y_1 or Y_2, ξ is a weighted sum of μ_3, σ_{13}, σ_{23} and σ_{33}, where the weights are the perturbations from the ML estimates in the final iterations of EM.

Table 5.4 *Sample ACFs of selected scalar parameters estimated over iterations of data augmentation*



lag	μ_3	σ_3	$\beta_{30\cdot 12}$	$\beta_{31\cdot 12}$	$\beta_{32\cdot 12}$	$\sigma_{3\cdot 12}$	ξ
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.18*	.31*	.37*	.33*	.44*	.35*	.25*
2	.04	.19*	.18*	.09*	.19*	.15*	.17*
3	.02	.07*	.10*	.08*	.10*	.05	.06
4	−.02	.09*	.05	.03	.06	.05	.08*
5	−.01	.11*	.02	−.01	.04	.05	.09*
6	−.01	.09*	.06	−.01	.06	.06	.07*
7	.04	.05*	.03	−.08*	.01	.03	.04
8	.01	.04	.02	−.10*	−.02	.05	.04
9	.03	.08*	.04	−.02	−.02	.04	.07*
10	.05	.04	.03	.02	−.02	.02	.04
11	−.06	.07	.01	.04	.03	−.03	.07
12	.01	.07*	.04	.06	.05	.02	.06
13	.02	.07	.00	−.01	.08	.04	.07
14	−.01	.08*	−.01	.00	.09*	.02	.09*
15	−.02	−.02	.04	.04	.04	.00	−.01
16	−.02	.02	.02	.02	.06	−.03	.02
17	.02	.01	−.03	.00	.07	−.04	.01
18	.00	−.02	−.02	−.01	.04	−.06	−.02
19	−.03	−.01	.04	.02	.01	−.05	−.01
20	.05	.00	.02	.05	.01	−.03	.01

* significantly different from zero at the 0.05 level

Sample ACFs for these functions of θ up to lag 20 are displayed in Table 5.4. Correlations that are significantly different from zero at the 0.05 level, as determined by Bartlett's formula (4.49), are marked with an asterisk. Because the series is so long and the serial dependence is not high, the standard errors are small and even very small correlations are deemed significant. Even for the worst functions examined, however, the correlations are effectively zero by lag 10, and definitely negligible by lag 20. Time-series plots of these functions showed no unusual features and resembled those of the rapidly-converging series displayed in Figure 4.2 (a) and (b). Based on this evidence, we feel safe in concluding that the algorithm effectively achieves stationarity by 20 iterations.

The parameters of greatest interest in this problem are functions of μ=(μ_1,$\mu_2\mu_3$)^T. For illustration, we will focus attention on three quantities: μ_3, the average cholesterol level at 14 days;

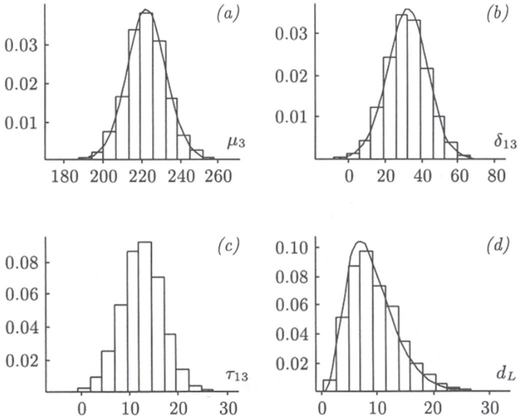


Figure 5.7. *Histograms of sample values of (a)μ_3, (b) δ_{13}, (C) τ_{13} and (d) d_L from 5000 consecutive iterations of data augmentation.*

$\delta_{13} = \mu_1 - \mu_3$, the average decrease in cholesterol level from day 2 to day 14; and τ_{13}=100(μ_1−μ_3)/μ_1, the relative percentage decrease in average cholesterol level from day 2 to day 14. To draw inferences about these quantities, we simulated another single chain of 5100 iterations starting from the ML estimate, discarded the first 100, and saved the 5000 remaining values of μ_3, δ_{13} and τ_{13}. Histograms of the sample values for these three quantities are shown in Figure 5.7 (a)-(c). Because μ_3 and δ_{13} are linear combinations of the elements of μ, obtaining Rao-Blackwellized estimates of the marginal densities of these quantities is straightforward. Under the prior (5.18), the complete-data posterior is given by (5-19)-(5.20). Using (5.17), it follows that the complete-data posterior density of a linear combination $\eta - a^T\mu$ is

$$P(\eta \mid Y_{obs}, Y_{mis}) = k\left[1+\frac{\left(\eta - a^T\bar{y}\right)^2}{(n-p)\sigma^2}\right]^{-(n-p+1)/2}, \qquad (5.45)$$

where n = 28 and p = 3 are the number of observations and variables, respectively; $\sigma^2 = (n-p)^{-1} a^T Sa$; $\bar{y}$ and S are the sample mean vector (5.5) and covariance matrix (5.6) computed from Y=(Y_{obs},Y_{mi}s); and

$$k = \frac{\Gamma\left(\frac{n-p+1}{2}\right)}{\Gamma\left(\frac{n-p}{2}\right)\sqrt{\pi(n-p)\sigma^2}}.$$

Rao-Blackwellized density estimates for μ_3=(0,0,1)μ and δ_{13}=(1,0,–1)μ estimated from the first 1000 iterations after the τ_{13} initial burn-in period are shown superimposed over the histograms in Figure 5.7 (a) and (b). Because τ_{13} is nonlinear its density is somewhat less easy to find, and Rao-Blackwellized estimates for this quantity are not shown.

In addition to μ_3, δ_{13} and τ_{13}, we also calculated and stored values of the likelihood-ratio statistic

$$d_L = d_L(\theta) = 2\left[l\left(\hat{\theta} \mid Y_{obs}\right) - l\left(\theta \mid Y_{obs}\right)\right]$$

over the 5000 iterations, where $\hat{\theta}$ is the ML estimate. For large samples, the posterior distribution of d_L is approximately $\chi^2_{d,}$, where d is the dimension of θ (in this case, 9). A histogram of the sample values of d$_L$ is displayed in Figure 5.7 (d) with the χ^2_9 density function superimposed over it, showing that the actual posterior matches the theoretical approximation quite closely.

Simulated posterior means for μ_3, δ_{13} and τ_{13} were found by averaging the 5000 iterates of each parameter. Simulated 95% posterior intervals were found by calculating the 2.5 and 97.5 percentiles of each sample using (4.8). To obtain a rough assessment of the random error in these estimates, a second chain was generated in an identical fashion with a different random-number generator seed. The simulated posterior means and 95% intervals (in parentheses) for the two replicate runs are shown below.

μ_3	δ_{13}	τ_{13}
222.2	31.8	12.4
(201.6,244.0)	(8.9,55.4)	(3.7,20.9)
222.4	31.4	12.3
(201.7,242.6)	(8.9,53.3)	(3.7,20.3)

Inferences about μ_3, δ_{13} and τ_{13} can also be conducted through multiple imputation. This will be demonstrated in Section 6.2.1.

5.4.4 Example: changes in heart rate due to marijuana use

Returning to the data in Table 5.2, let μ_j denote the population mean corresponding to column j, and let $\delta_{j\kappa}=\mu_j-\mu_\kappa; j,k=1,\ldots,6$. Following the original article be Weil et al. (1968), we will focus attention on the six treatment comparisons below.

15 minutes		90 minutes	
Low *vs.* Placebo	δ_{21}	Low *vs.* Placebo	δ_{54}
High *vs.* Placebo	δ_{31}	High *vs.* Placebo	δ_{64}
High *vs.* Low	δ_{32}	High *vs.* Low	δ_{65}

Data augmentation under the usual noninformative prior (5.18) does not work for this problem; the iterates of θ quickly wander to the boundary of the parameter space, causing numeric overflow. This pathological behavior suggests that the posterior is not proper. To stabilize the inference, we applied a ridge prior as described in Sections 5.2.3 and 5.3.4. After centering and scaling the columns of Y so that the observed data in each column have mean zero and unit variance, we set the hyperparameters of the normal inverted-Wishart prior to $\tau=0$, $m=\epsilon$, and $\Lambda^{-1}=\epsilon I$ for $\epsilon=0.5$. Under this weak prior, EM converges slowly but reliably to a posterior mode in the interior of the parameter space, with the largest fraction of missing information estimated at 95%.

The slow convergence of EM in this example suggests that data augmentation will also converge slowly, and output analysis from a preliminary run confirmed this. Using the same ridge prior, we simulated a single chain beginning at the posterior mode and monitored a variety of scalar summaries of θ. Time-series plots for δ_{21} and δ_{54} (on the original scale) from the first 100 iterations are shown in Figure 5.8 (a) and (b), respectively. The iterates of δ_{21} appear to approach stationarity quickly, whereas the series for δ_{54} shows long-range dependence. This is not surprising, because δ_{54} is a function of μ_4, and our earlier analysis led us to conjecture that the rate of missing information for μ_4 was very high. Sample ACFs for δ_{21} and δ_{54} estimated from 10 000 iterations are displayed in Figure 5.8 (c) and (d), respectively. Figure 5.8 (d) is typical of the ACFs for other slowly converging functions of θ. For all the functions we examined, the serial correlations effectively died out by lag 50.

The slow convergence in this example should lead us to use extra caution in designing the simulation experiment. Running independent chains from overdispersed starting values would be attractive,

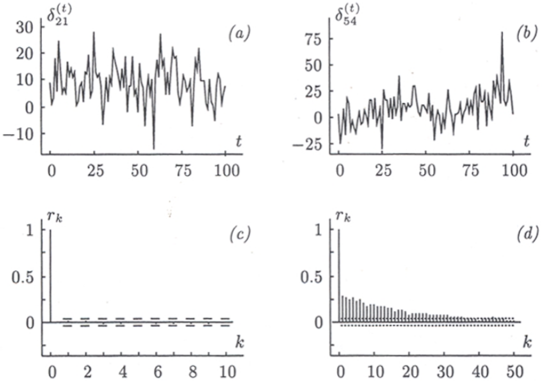


Figure 5.8. *Time-series plots of (a) δ_{21} and (b) δ_{54} over the first 100 iterations of data augmentation, and sample AM for* (C) *δ_{21} and (d) δ_{54} estimated from 10 000 iterations, with dashes indicating approximate 0.05-level critical values for testing $\rho_{\kappa}=\rho_{\kappa+1}=\ldots=0$.*

but obtaining overdispersed starting values is not easy. Bootstrap resampling is unlikely to work well, because n is not much larger than p, so the distribution of θ over bootstrap samples will probably bear little resemblance to the observed-data posterior. Sampling from the prior is not possible, because the prior is not a proper probability distribution. Because convergence to stationarity tends to be fastest when the starting value is near the center of the observed-data posterior, we decided to run ten independent chains of 5500 iterations each, starting each chain at the posterior mode. After discarding the first 500 values from each chain, the pth sample quantile for each contrast $\delta_{j\kappa}$, was calculated for $p = 0.025$, 0.25, 0.5, 0.75 and 0.975 from the remaining 5000 values. Finally, the sample quantiles were averaged across the ten chains. For each of these averages, the variance of the quantiles across chains was used to estimate a standard error with nine degrees of freedom. The estimated quantiles for all six parameters are displayed in Figure 5.9. All of the simulated 95% posterior intervals cover zero, indicating that there is no strong evidence that any of the contrasts is different from zero. Standard errors for the simulated quantiles

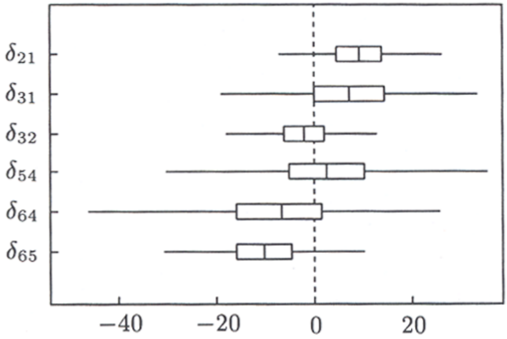


Figure 5.9. *Simulated posterior medians, quantiles and 951% equal-tailed intervals for six contrasts.*

ranged from 0.02 to 0.72, which is quite small relative to the width of the intervals displayed in Figure 5.9, so these simulation results are sharp enough for our purposes.

One could very well argue that the unrestricted multivariate normal model has too many parameters to be estimated from a dataset of this size, and that the unnecessarily large number of nuisance parameters hinders us from making clear inferences about the parameters of interest. Indeed, the long tails exhibited in the marginal posteriors of Figure 5.9, particularly for the two contrasts involving μ_4, suggest that some of the nuisance parameters are very poorly estimated, and we might do well to simplify the model. One possible simplification is to reduce the number of free parameters by applying a priori constraints to Σ. For example, we could require Σ to satisfy the condition of *compound symmetry* (i.e. equal diagonal elements and equal off-diagonal elements). Simulation algorithms for incomplete multivariate normal data with constrained covariance structure are possible, but they are beyond the scope of this book. A slightly different approach would be to specify fixed, additive effects for the rows and columns of the data matrix, and define the parameters of interest to be contrasts among the column effects (Chapter 9).

Yet another possibility is to perform a simple bivariate analysis for each contrast, making inferences about $\delta_{j\kappa}$ using only the data in columns j and k. Under this bivariate approach, it is no longer possible to make joint inferences about the contrasts. Moreover, ignoring the data in columns other than j and k when making inferences about δ_{jk} may tend to introduce nonresponse biases;

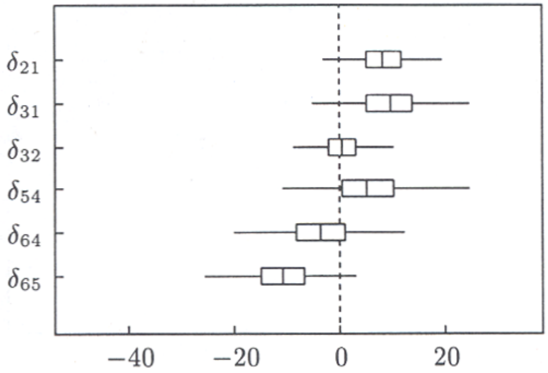


Figure 5.10. *Simulated posterior medians, quantiles and 951% equal-tailed intervals for six contrasts using a bivariate approach.*

the MAR assumption tends to be less plausible for the bivariate dataset than for the one with six variables. The decision whether to include additional variables in an analysis is not always an easy one, particularly for small datasets, and is an important topic worthy of further research.

Simulated posterior quantiles from a bivariate analysis are shown in Figure 5.10. For each contrast, data augmentation was applied to the bivariate dataset under the standard noninformative prior (5.18). Output analyses suggested that convergence to stationarity was rapid. For each contrast, 10 100 steps of a single Markov chain were simulated, beginning from the ML estimate. The first 100 values of the simulated contrast were discarded, and sample quantiles were calculated from the remaining 10 000. The distributions in Figure 5.10 are much narrower than those in Figure 5.9, and there is now a fair amount of evidence that the three contrasts δ_{21}, δ_{31}, and δ_{65} are nonzero.

CHAPTER 6

More On The Normal Model

6.1 Introduction

In the last chapter, we introduced EM and data augmentation algorithms for the multivariate normal model. In this chapter, we illustrate how to effectively apply these algorithms with more real-data examples, and discuss modifications to the algorithms that can help to increase their efficiency.

Sections 6.2 and 6.3 present two examples of analysis by multiple imputation. The first, which was previously analyzed in Chapter 5 by parameter simulation, is straightforward and illustrates some of the basic properties of multiple-imputation point and interval estimates. The second is more complicated, involving categorical variables and inestimable parameters. By working through this second example, the reader will come to understand some of the complications and subtle issues that often arise with real data, and learn strategies for effectively dealing with these issues.

Real data often do not conform to normality, and it is important to know whether the multiple-imputation procedures advocated in this book are robust to departures from the modeling assumptions. Section 6.4 presents a simulation experiment to demonstrate the robustness of multiple imputation in a realistic setting.

When rates of missing information are high, EM and data augmentation tend to converge slowly. Section 6.5 presents a new class of simulation algorithms, called monotone data augmentation, that tend to converge quickly under certain types of missingness.

6.2 Multiple imputation: example 1

6.2.1 Cholesterol levels of heart-attack patients

Recall the example introduced in Section 5.3.6 in which serum cholesterol levels for heart-attack patients were recorded 2 days (Y_1), 4 days (Y_2) and 14 days (Y_3) after attack. Nine of the n = 28 values of Y_3 were missing. In Section 5.4.3, we used data augmentation to simulate posterior distributions for three parameters of interest:

1. μ_3 the mean cholesterol level at 14 days;
2. $\delta_{13} = \mu_1 - \mu_3$, the average decrease in cholesterol level from day 2 to day 14; and
3. $\tau_{13} = 100(\mu_1 - \mu_3)/\mu_1$, the percentage decrease in cholesterol level from day 2 to day 14.

We now demonstrate how inferences for these same quantities can be conducted by multiple imputation.

6.2.2 Generating the imputations

Recall that proper multiple imputations are independent draws of Y_{mis} from the posterior predictive distribution of the missing data, $P(Y_{mis}|Y_{obs})$. The exploratory run of data augmentation revealed no discernible autocorrelations in scalar functions of θ beyond lag 10. Thus we can probably obtain acceptable imputations by (a) running data augmentation in a single chain starting from the MLE, and taking every tenth iterate of Y_{mis} as an imputation; or (b) running independent, parallel chains of ten iterations each starting from the MLE, and taking the final value of Y_{mis} from each chain as an imputation.

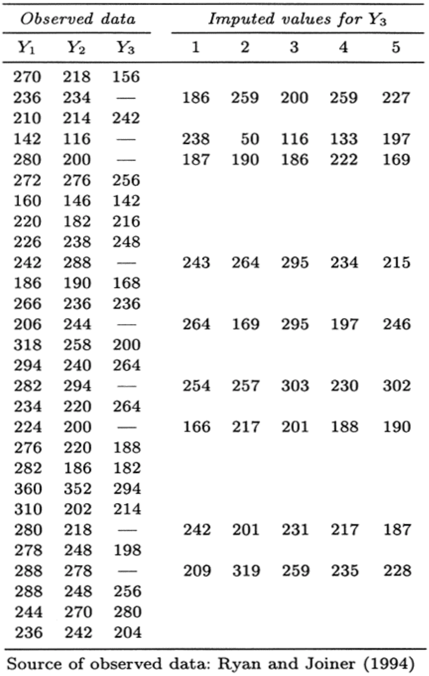
Because of the small size of this dataset, however, iterations are computationally inexpensive, and we can easily afford to increase the number of steps. To illustrate a conservative approach, we generated m=5 multiple imputations by simulating five independent chains of 50 steps each.

Independent starting values for the chains were obtained by running EM on independent bootstrap samples of size $n/2 = 14$ (Section 4.4.2). These starting values are probably overdispersed relative to the observed-data posterior $P(\theta|Y_{obs})$, so that in the unlikely event that stationarity has not been achieved by 50 steps, the resulting inferences will tend to be conservative. The $m = 5$ sets of imputed values for Y_3, rounded to integers, are displayed in Table 6.1.

6.2.3 Complete-data point and variance estimates

Multiple imputation r equires that for each estimand Q we specify a complete-data point estimate $\hat{Q}$ and a complete-data variance

Table 6.1. *Cholesterol levels for heart-attack patients measured 2, 4 and 14 days after attack, with m = 5 multiple imputations*



Observed data			*Imputed values for Y_3*				
Y_1	Y_2	Y_3	1	2	3	4	5
270	218	156					
236	234	—	186	259	200	259	227
210	214	242					
142	116	—	238	50	116	133	197
280	200	—	187	190	186	222	169
272	276	256					
160	146	142					
220	182	216					
226	238	248					
242	288	—	243	264	295	234	215
186	190	168					
266	236	236					
206	244	—	264	169	295	197	246
318	258	200					
294	240	264					
282	294	—	254	257	303	230	302
234	220	264					
224	200	—	166	217	201	188	190
276	220	188					
282	186	182					
360	352	294					
310	202	214					
280	218	—	242	201	231	217	187
278	248	198					
288	278	—	209	319	259	235	228
288	248	256					
244	270	280					
236	242	204					

Source of observed data: Ryan and Joiner (1994) estimate U. It also requires a sample size large enough for the approximation

$$\frac{\hat{Q}-Q}{\sqrt{U}} \sim N(0,1) \tag{6.1}$$

to work well with complete data. Let

$$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \; and \; S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_{ij}-\bar{y}_j\right)\left(y_{ik}-\bar{y}_k\right)$$

for j, k = 1, 2, 3 denote the complete-data sample means and covariances. For μ_3, the obvious complete-data estimates are $\hat{Q}=\bar{y}_3$ and $U{=}S_{33}/n$. For $\delta_{13}{=}\mu_1{-}\mu_3$, the obvious choices are

$$\begin{aligned}&\hat{Q}=\bar{y}_1-\bar{y}_3\\&U=\left(S_{11}-2S_{13}+S_{33}\right)/\,n.\end{aligned}$$

Asymptotic normality of $\bar{y}_3$ and $\bar{y}_1-\bar{y}_3$ is guaranteed by the Central Limit Theorem, and a sample of size n = 28 should be large enough for the normal approximations to work well.

For the nonlinear parameter τ_{13} = $100(\mu_1{-}\mu_3)/\mu_1$, a first-order Taylor expansion of the function $\left(\bar{y}_1-\bar{y}_3\right)/\,\bar{y}_1$ about (μ_1,μ_3),

$$\frac{\bar{y}_1-\bar{y}_3}{\bar{y}_1}-\frac{\mu_1-\mu_3}{\mu_1} \approx \frac{\mu_3}{\mu_1^2}\left(\bar{y}_1-\mu_1\right)-\frac{1}{\mu_1}\left(\bar{y}_3-\mu_3\right),$$

suggests that the complete-data point estimate

$$\hat{Q}=100\left(\bar{y}_1-\bar{y}_3\right)/\,\bar{y}_1$$

will be approximately unbiased for τ_{13}, with approximate variance

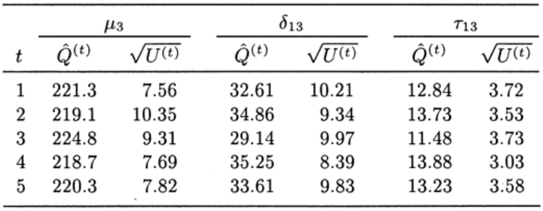
$$V(\hat{Q}) \approx \frac{100^2}{n}\left[\left(\frac{\mu_3^2}{\mu_1^4}\right)\sigma_{11} - 2\left(\frac{\mu_3}{\mu_1^3}\right)\sigma_{13} + \left(\frac{1}{\mu_1^2}\right)\sigma_{33}\right].$$

A reasonable complete-data variance estimate is thus

$$U = \frac{100^2}{n}\left[\left(\frac{\bar{y}_3^2}{\bar{y}_1^4}\right)S_{11} - 2\left(\frac{\bar{y}_3}{\bar{y}_1^3}\right)S_{13} + \left(\frac{1}{\bar{y}_1^2}\right)S_{33}\right].$$

A handy rule-of-thumb used by survey statisticians is that a ratio of sample means will be approximately unbiased and normally distributed if the coefficient of variation (the standard deviation divided by the mean) of the denominator is 10% or less (e.g. Cochran, 1977, p. 166). The observed values of Y_1 in Table 6.1 have a mean and standard deviation of 253.9 and 47.7, respectively, so the estimated coefficient of variation for $\bar{y}_1$ is $\left(47.7/\sqrt{28}\right)/253.9=0.036$, suggesting that the normal approximation should work well.

Table 6.2. *Complete-data point estimates and standard errors for μ_3, δ_{13} and τ_{13} from m=5 multiply-imputed datasets*



	μ_3		δ_{13}		τ_{13}	
t	$\hat{Q}^{(t)}$	$\sqrt{U^{(t)}}$	$\hat{Q}^{(t)}$	$\sqrt{U^{(t)}}$	$\hat{Q}^{(t)}$	$\sqrt{U^{(t)}}$
1	221.3	7.56	32.61	10.21	12.84	3.72
2	219.1	10.35	34.86	9.34	13.73	3.53
3	224.8	9.31	29.14	9.97	11.48	3.73
4	218.7	7.69	35.25	8.39	13.88	3.03
5	220.3	7.82	33.61	9.83	13.23	3.58

Following the notation of Section 4.3.2, let $\hat{Q}^{(t)}$ and $U^{(t)}$ denote the complete-data point and variance estimates from the tth imputed dataset. Point and variance estimates for μ_1, δ_{13} and τ_{13} over the five imputations are displayed in Table 6.2.

6.2.4 Combining the estimates

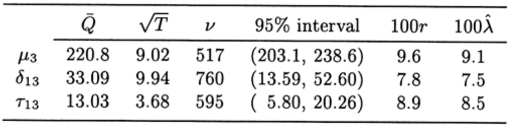
Combining the complete-data point and interval estimates is a straightforward application of the formulas in Section 4.3.2 for inference with a scalar estimand. The overall estimates $\overline{Q}$, standard errors $\sqrt{T}$, degrees of freedom v for the t-approximation and 95% interval estimates are displayed in Table 6.3. The values of v are large, suggesting that the total variance estimates T are stable even though they are based on only $m = 5$ imputations. The point and interval estimates in Table 6.3 differ somewhat from those obtained by parameter simulation in Section 5.4.3, but the differences are mild relative to the sizes of the standard errors.

Table 6.3 also displays two diagnostics described in Section 4.3.2: the relative increase in variance due to nonresponse r, and the estimated fraction of missing information $\hat{\lambda}$. Although 32% of the Y_3 values are missing, the estimated rates of missing information for μ_3, δ_{13} and τ_{13} are under 10%, due undoubtedly to the correlations between Y_3 and the two variables that are never missing.

6.2.5 *Alternative choices for the number of imputations*

For this analysis we chose m=5 imputations, because we knew that the fractions of missing information would not be severe. Recall that if the fraction of missing information for a parameter is λ, the relative efficiency of an estimate based on m imputations to one based on an infinite number is approximately $(1+\lambda/m)^{-1}$ (Section 4.3.1). From EM we learned that the worst fraction of missing information for this problem was about 47% (Section 5.3.6). Thus in the worst case, $m = 5$ would lead to a point estimate that is about $(1+0.47/5)^{-1} = 91\%$ as efficient as one with $m = \infty$. In fact, the estimated fractions of missing

Table 6.3. *Results of multiple-imputation inference for* μ_3, δ_{13} *and* τ_{13}



	$\bar{Q}$	$\sqrt{T}$	ν	95% interval	$100r$	$100\hat{\lambda}$
μ_3	220.8	9.02	517	(203.1, 238.6)	9.6	9.1
δ_{13}	33.09	9.94	760	(13.59, 52.60)	7.8	7.5
τ_{13}	13.03	3.68	595	(5.80, 20.26)	8.9	8.5

information for the parameters of interest were about 10%, so the estimates from m = 5 imputations appear to be about $(1+0.1/5)^{-1}$ = 98% efficient.

To those unaccustomed to multiple imputation, basing any conclusion on a Monte Carlo simulation with only m = 5 draws might seem risky. A critic might argue that with only five imputations, one or more ’bad’ (i.e. highly unusual) imputations could exert an undue influence on the results. To illustrate the effect of increasing the size of m, we generated an additional 95 imputations in the manner described above, for a total of 100 imputations. We then calculated point and interval estimates based on m – 3, 5, 10, 20, and 100. For m = 3 we used the first 3 imputations; for m = 5 we used the first 5 imputations; and so on. Finally, to get a rough idea of the amount of random variation in the estimates, we replicated the entire experiment, generating another 100 imputations from a different random-number generator seed and calculating another set of estimates for m = 3, 5, 10, 20, and 100.

The point and interval estimates for the various values of m are displayed graphically in Figure 6.1. For comparison, Figure 6.1 also displays the results of the two parameter-simulation runs of length 5000 described in Section 5.4.3. The multiple-imputation (MI) intervals for m = 3 and m = 5 appear to have more random variation than the parameter-simulation (PS) intervals. By m = 10, however, the MI intervals appear to remarkably stable, and there is little random variation (relative to the widths of the intervals) in any of the results for m = 10, 20 or 100.

The variability for m = 3 and m = 5 does not mean that these

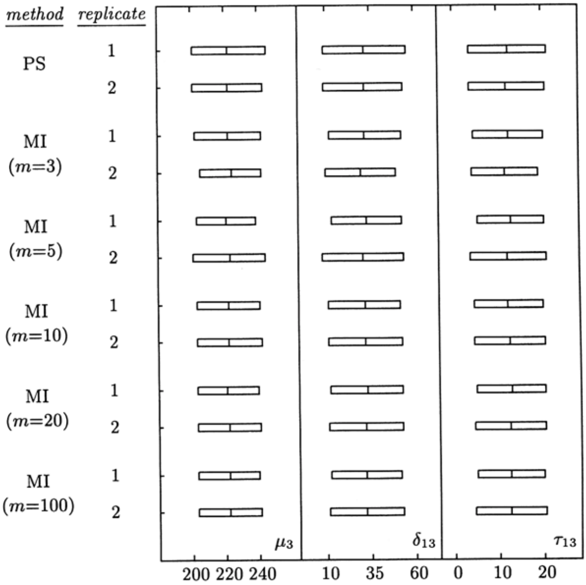


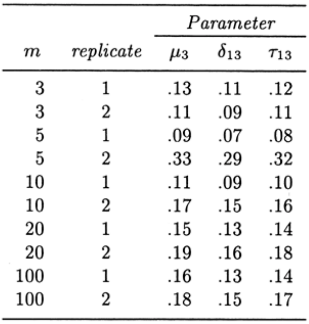
Figure 6.1. *Point and 95% interval estimates for μ_3, δ_{13} and τ_{13} from parameter simulation (PS) and multiple imputation (MI).*

intervals are unreliable. The intervals explicitly include simulation error as a component of uncertainty, and over repeated application they should still cover the true values of the parameters at least 95% of the time. To reduce random variation, one might consider increasing m, particularly if generating and storing imputations is not expensive. Based on Figure 6.1, however, there appears to be little reason to use more than $m = 10$ imputations for this problem.

Advantages of multiple imputation over parameter simulation

The PS estimates based on 5000 iterates of θ appear to be about as stable as MI estimates based on only m=10 imputations of Y_{mis}. Notice, however, that the latter required only one-tenth as much

Table 6.4. *Estimated fractions of missing information from m=3, 5, 10, 20 and 100 imputations*



		Parameter		
m	*replicate*	μ_3	δ_{13}	τ_{13}
3	1	.13	.11	.12
3	2	.11	.09	.11
5	1	.09	.07	.08
5	2	.33	.29	.32
10	1	.11	.09	.10
10	2	.17	.15	.16
20	1	.15	.13	.14
20	2	.19	.16	.18
100	1	.16	.13	.14
100	2	.18	.15	.17

computation (500 steps of data augmentation versus 5000) and 0.6% as much storage (10 × 9 = 90 locations to hold imputations of Y_{mis}, versus 5000 × 3 = 15 000 to hold values of μ_3, δ_{13} and τ_{13}).

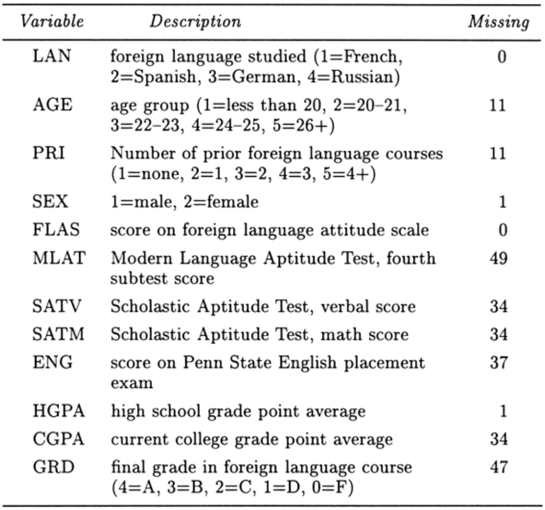
A further advantage of MI is that it provides an estimated fraction of missing information for each estimand. For small *m*, however, these estimates can be noisy. To illustrate, estimated fractions of missing information for μ_3, δ_{13} and τ_{13} based on $m = 3$, 5, 10, 20, and 100 imputations (both replicates) are shown in Table 6.4. For small *m*, the estimates vary substantially between replicates. This is to be expected, because they depend on the between-imputation components of variance which are estimated with only m_1 degrees of freedom. Recall that our initial estimates of A based on $m = 5$ imputations were all under 10% (Table 6.3); after increasing the value of *m* to 100, the estimates rose to 13-18%. Additional replications (not shown) demonstrate that even for $m = 100$, the estimates $\hat{\lambda}$ still have standard errors of approximately 0.02. Thus for small values of *m*, $\hat{\lambda}$ should be used only as a rough guide.

6.3 Multiple imputation: example 2

6.3.1 Predicting achievement in foreign language study

Raymond (1987) describes data that were collected to investigate the usefulness of a newly developed instrument, the Foreign Language Attitude Scale (FLAS), for predicting success in the study of foreign languages. In particular, the investigators wanted to determine whether the FLAS had substantial predictive ability beyond that already provided by other well-established instruments such as the Modern Language Aptitude Test (MLAT). Twelve variables were collected for a sample of n = 279 students enrolled in foreign language courses at The Pennsylvania State University in the early 1980s (Raymond and Roberts, 1983). Descriptions of the variables, along with the number of missing values for each one, appear in Table 6.5. The raw data, kindly provided by Dr. Mark Raymond, are reproduced in Appendix A.

Table 6.5. *Variables in foreign language achievement study, with number of missing values*



Variable	*Description*	*Missing*
LAN	foreign language studied (1=French, 2=Spanish, 3=German, 4=Russian)	0
AGE	age group (1=less than 20, 2=20–21, 3=22–23, 4=24–25, 5=26+)	11
PRI	Number of prior foreign language courses (1=none, 2=1, 3=2, 4=3, 5=4+)	11
SEX	1=male, 2=female	1
FLAS	score on foreign language attitude scale	0
MLAT	Modern Language Aptitude Test, fourth subtest score	49
SATV	Scholastic Aptitude Test, verbal score	34
SATM	Scholastic Aptitude Test, math score	34
ENG	score on Penn State English placement exam	37
HGPA	high school grade point average	1
CGPA	current college grade point average	34
GRD	final grade in foreign language course (4=A, 3=B, 2=C, 1=D, 0=F)	47

In this example, only 8% of all the values in the 279 × 12 data matrix are missing, and missingness rates per variable range from 0% to 18%. Only 62% of the cases (174 out of 279) have complete data for all twelve variables, however, so the case-deletion methods used by most statistical software packages would discard over one third of the entire dataset. Imputing for the missing values makes more efficient use of the available data.

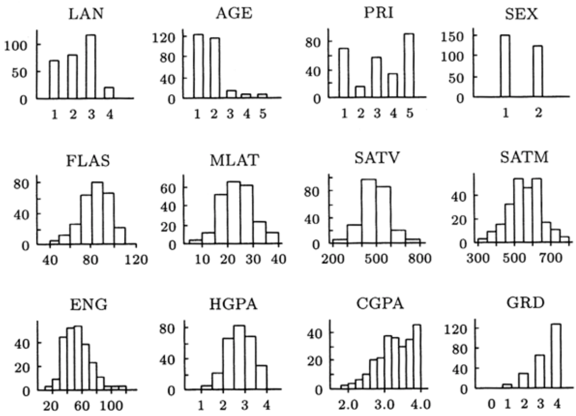


Figure 6.2. *Histograms of observed data for variables in foreign language achievement study.*

6.3.2 Applying the normal model

Histograms of the observed values for each variable are displayed in Figure 6.2. Although these data clearly do not follow a multivariate normal distribution, we will still use the normal model for imputation. For the dichotomous and ordinal variables, we will impute under an assumption of normality and round off the continuous imputes to the nearest category. Examination of Figure 6.2 suggests that this strategy might not work well for AGE, PRI or GRD, because these variables are far from being symmetric and unimodal.

To make the variables AGE, PRI and GRD less troublesome, we recoded them by collapsing some adjacent categories. (In Chapter 9, when we are able to explicitly model mixed continuous and categorical data, we will analyze these data again without recoding.) An overwhelming majority of students received final grades of A or B; very few received C or below; the data provide relatively little information to characterize the C-or-below group, so we recoded final grade as a simple dichotomy (A, B or below). Similarly, the three highest age groups had very few students in them, so age was collapsed to a dichotomy as well (less than 20, 20+). Prior experience was reduced from five categories to three. Histograms of the recoded versions of AGE, PRI, and GRD and the revised definitions of these variables appear in Figure 6.3 and Table 6.6, respectively.

Notice that the variable LAN is nominal and should not be handled as a normal variable; the four language groups have no intrinsic ordering. To address this issue, LAN was replaced by a set of three dummy variables to distinguish among the four language groups: $\text{LAN}_2 = 1$ if Spanish and 0 otherwise, $\text{LAN}_3 = 1$ if German and 0 otherwise, and $\text{LAN}_4 = 1$ if Russian and 0 otherwise. Including LAN_2, LAN_3 and LAN_4 effectively treats the eleven remaining variables as multivariate normal within each of the four language groups, with a separate mean vector for each group and a common covariance matrix. The multivariate normal model clearly misspecifies the marginal distribution of the dummy variables, but this misspecification is of no consequence because the dummies are completely observed and do not need to be imputed (Section 2.6.2).

Finally, it is important to remember that a normal distribution has support on the whole real line, but the continuous variables in this dataset have a limited range of possible values. For example, SAT scores may not exceed 800, and grade point averages may not exceed 4.0. Imputing under normality might occasionally result in an imputed value that is out of range. To handle this problem, we included a consistency check in our imputation routine. After performing the final I-step of data augmentation to create an imputation of Y_{mis}, each row of the imputed dataset was examined to see whether any of the imputed values were out of range; if so, the missing data for that row were re-drawn until the necessary constraints were satisfied. The final values of Y_{mis} created by this procedure approximate proper multiple imputations under a truncated multivariate normal model.

6.3.3 Exploring the observed-data likelihood and posterior

When LAN is replaced by three dummy variables, the dataset has p, = 14 variables. The EM algorithm applied to these 14 variables converged rapidly; the parameter estimates stabilized to four significant digits after only ten iterations. When EM converges so quickly, estimating the largest fraction of missing information from the iterations can be difficult, because the estimated elementwise rates of convergence (3.27) tend to become numerically unstable after only a few iterations. Moreover, the iterations at which instability begins vary from component to component. The multivariate normal model for 14 variables has 119 parameters. With so many parameters, it is not easy to estimate the fraction of missing information by visually inspecting the elementwise rates. In situations like this it is helpful to apply graphical techniques.

To estimate the worst fraction of missing information, we first calculated elementwise rates (3.27) for each of the 119 parameters over the first 20 iterations of EM. After trimming away any values outside the interval (0, 1), we formed boxplots of the remaining values for each parameter, displaying them side-by-side. Boxplots for 50 randomly selected elements of μ and Σ are shown in Figure 6.4. Although a large number of outliers are present, all of the boxplots tend to be centered around 0.4. The median of the values in Figure 6.4 is 0.42, so a reasonable estimate of the worst fraction of missing information is 42%.

Inestimability of parameters

The moderate rates of missing information and the rapid convergence of EM might lead one to believe that the observed-data

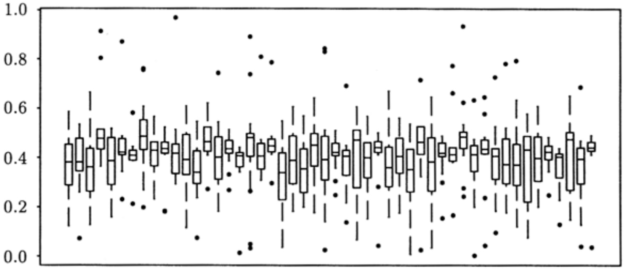


Figure 6.4. *Boxplots of estimated elementwise rates of convergence for 50 randomly selected parameters.*

likelihood function for this problem is well behaved. It turns out, however, that the likelihood is pathological. We performed a long exploratory run of data augmentation under the usual noninformative prior (5.18) and constructed time-series plots for selected elements of μ and Σ. For most parameters, the algorithm appeared to achieve stationarity very quickly. For a few parameters, however, the simulated values drifted into implausible regions of the parameter space. Time series plots for the means of the two variables with the highest rates of missingness, MLAT and GRD, are shown in Figure 6.5. Figure 6.5 (a) is typical of the plots for most parameters, with no discernible trends. Figure 6.5 (b), however, shows extreme long-range dependence. The mean of the dichotomous variable GRD is known to lie between 1 and 2, but by the 900th iteration the series has drifted above 2. This unusual behavior suggests that one or more components of θ are nearly or entirely inestimable from the observed data.

Additional runs of EM confirmed the presence of inestimable parameters. Using various simulated values of θ from the data-augmentation series as starting values, we re-ran EM and found that in each case it converged to a different stationary value. Moreover, when we evaluated the observed-data loglikelihood function at these stationary values, the loglikelihood was exactly the same in each case. Thus it appears that the stationary values are not distinct modes, but form a ridge of constant likelihood. The pathological behavior in Figure 6.5 (b) arises because the observed-data posterior distribution is not proper; although the I- and P-steps of data augmentation are both well defined, the algorithm is not converging to any stationary distribution (Section 3.5.2).

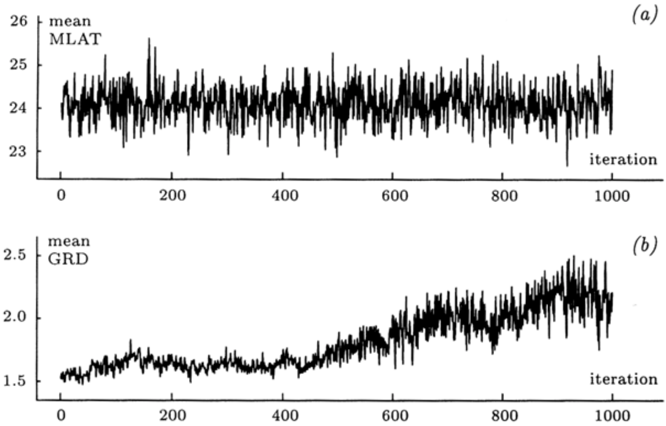


Figure 6.5. *Time series plots of (a) mean MLAT and (b) mean GRD over 1000 iterations of data augmentation.*

Table 6.7. *Cross-tabulation of LAN with GRD*

	LAN = 1	LAN = 2	LAN = 3	LAN = 4
GRD = 1	36	34	36	0
GRD = 2	27	31	68	0
GRD missing	4	13	10	20

With a little exploration it is easy to detect the source of difficulty. Figure 6.5 (b) suggests that the inestimable part of θ pertains to the distribution of GRD. A cross-tabulation of GRD with LAN, shown in Table 6.7, reveals that GRD is missing for all cases with LAN = 4. Because no values of GRD are available for any students enrolled in Russian courses, it is impossible to estimate the parameters of the conditional distribution of GRD given LAN= 4 from this dataset.

6.3.4 Overcoming the problem of inestimability

One way to solve the problem of inestimability is to simply exclude the Russian language group and the variable LAN_4

from the analysis. Because GRD is missing for all 20 of these cases, they contribute little or no information about the main question of scientific interest, which pertains to the quality of FLAS as a predictor of GRD. Another way to handle the problem is to introduce a small amount of information about the inestimable portions of θ through a mildly informative prior distribution. Although excluding the Russian language group is certainly reasonable, we will adopt the latter approach to illustrate the use of an informative prior distribution.

After centering and scaling the observed data for each variable to have mean 0 and variance 1, we applied the ridge prior described in Section 5.2.3 with τ=0, m=$\in$ and $\Lambda^{-1}=\in I$ for $\in$=3. This prior adds the equivalent of three degrees of freedom to the estimation of Σ and smooths the estimated correlation matrix toward 1. With a sample size of n = 279 the degree of smoothing is slight, and the effect on those portions of θ that are already well estimated is almost negligible. For portions of θ that are poorly estimated, however, this prior smooths the estimates toward a model of mutual independence among all variables. Inferences under this prior will thus tend to be conservative in the sense that we will be less likely to conclude that associations among variables are present when in fact they are not.

Under this prior, EM was found to converge reliably from a variety of starting values to a single posterior mode. The convergence was slower than before, requiring about 30 iterations, and the largest fraction of missing information was estimated at 92%. It may seem somewhat counterintuitive that the introduction of prior information appears to raise the worst fraction of missing information rather than lower it. This fraction, however, pertains only to those directions or functions of θ for which the function being maximized (i.e. the observed-data likelihood or posterior) is not flat. The elementwise rates estimate the largest eigenvalue of the asymptotic rate matrix that is less than one (Section 3.3.2). A ridge in the function produces one or more eigenvalues equal to one, and thus the inestimable functions of θ do not contribute to the estimated worst fraction of missing information when EM is used to maximize the likelihood. When an informative prior is introduced, however, the posterior is no longer precisely flat in any direction, and every function of θ then contributes to the estimated worst fraction of missing information.

Under the informative prior, data augmentation also appears to converge reliably. Starting at the mode, we ran a single chain for 1000 iterations and monitored a variety of functions of θ. Sample ACFs for two functions are shown in Figure 6.6. The mean of GRD, which behaved pathologically under the noninformative prior, now shows no appreciable dependence after lag 20. The worst linear function of θ, as estimated by the trajectory of EM in the vicinity of the posterior mode (Section 4.4.3), appears to achieve stationarity in about 25 steps.

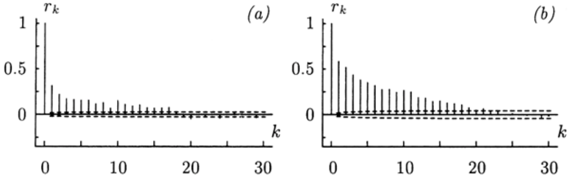


Figure 6.6. *Sample ACFs for (a) mean GRD and (b) the worst linear function of θ, estimated from 1000 iterations of data augmentation, with dashed lines indicating approximate critical values for testing* $\rho_k=\rho_{k+1}=\cdots=0$.

6.3.5 Analysis by multiple imputation

Following the preliminary run, we created $m = 20$ multiple imputations of the missing data by running 20 independent chains for 100 steps each. Starting values for the chains were obtained by finding posterior modes from independent bootstrap samples of 140 subjects each.

Inferences for logistic-regression coefficients

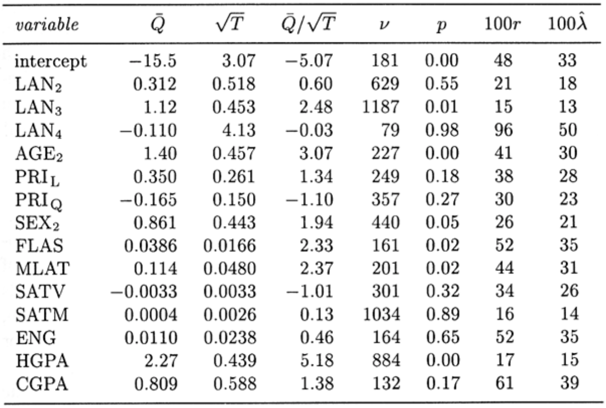
Because the response variable GRD was collapsed to a dichotomy, we decided to measure the predictive ability of

FLAS and the other variables by logistic regression (e.g. McCullagh and Nelder, 1989). Let π_i denote the probability of GRD = 2 for subject i. We examined the model

$$\log \frac{\pi_i}{1-\pi_i} = x_i^T \beta, \tag{6.2}$$

where x_i is a vector of covariates for subject i and β a vector of unknown coefficients. Covariates in x_i included a term for the intercept; three dummy indicators for language (LAN_2, LAN_3 and LAN_4); an indicator for age (AGE_2 = 1 if 20+ and 0 otherwise); an indicator for sex (SEX_2 = 1 if female and 0 otherwise); linear and quadratic contrasts for PRI (PRI_L = −1,0,1 and PRI_Q = 1,–2,1 for PRI = 1,2,3, respectively);

Table 6.8. *Multiple-imputation inferences for logistic-regression coefficients, full model*



variable	$\bar{Q}$	$\sqrt{T}$	$\bar{Q}/\sqrt{T}$	ν	p	$100r$	$100\hat{\lambda}$
intercept	−15.5	3.07	−5.07	181	0.00	48	33
LAN_2	0.312	0.518	0.60	629	0.55	21	18
LAN_3	1.12	0.453	2.48	1187	0.01	15	13
LAN_4	−0.110	4.13	−0.03	79	0.98	96	50
AGE_2	1.40	0.457	3.07	227	0.00	41	30
PRI_L	0.350	0.261	1.34	249	0.18	38	28
PRI_Q	−0.165	0.150	−1.10	357	0.27	30	23
SEX_2	0.861	0.443	1.94	440	0.05	26	21
FLAS	0.0386	0.0166	2.33	161	0.02	52	35
MLAT	0.114	0.0480	2.37	201	0.02	44	31
SATV	−0.0033	0.0033	−1.01	301	0.32	34	26
SATM	0.0004	0.0026	0.13	1034	0.89	16	14
ENG	0.0110	0.0238	0.46	164	0.65	52	35
HGPA	2.27	0.439	5.18	884	0.00	17	15
CGPA	0.809	0.588	1.38	132	0.17	61	39

and the variables FLAS, MLAT, SATV, SATM, ENG, HGPA and CGPA. For each of the 20 imputed datasets, we computed ML estimates and asymptotic standard errors for the elements of β, and then combined the 20 sets using the formulas for multiple-imputation inference for scalar estimands (Section 4.3.2).

The results of the analysis are summarized in Table 6.8. For each coefficient, Table 6.8 displays the point estimate $\overline{Q}$ and standard error $\sqrt{T}$, the *t*-statistic $\overline{Q}/\sqrt{T}$, the degrees of freedom v for the Student's *t*-approximation, and the p-value for testing the hypothesis $Q = 0$ against a two-sided alternative. Also shown are the relative increase in variance due to nonresponse r and the estimated fraction of missing information $\hat{\lambda}$. The p-value for FLAS (0.021) suggests that this variable is useful for predicting GRD. Increasing FLAS by ten points multiplies the odds $\pi_i/(1-\pi_i)$ by an estimated factor $e^{10\times 0.0386} = 1.47$ in other words, every ten-point increase in FLAS makes a student 47% more likely (on the odds scale) to receive a grade of A, if other covariates are held constant. The most powerful predictor of final grade appears to be high-school GPA; a one-unit increase in HGPA causes the predicted odds to be multiplied by $e^{2.27} = 9.68$. The only significant language effect is the coefficient of LAN$_3$, which distinguishes between the German and French groups; a student taking German appears to be about $e^{1.12} = 3.06$ times as likely to receive an A as a student taking French. Notice that LAN$_4$, which contrasts Russian with French, has a non-significant effect ($p = 0.979$) and a high fraction of missing information (50%). This is to be expected, because essentially all information about this parameter comes from the prior distribution which tends to pull the estimated coefficient toward zero.

Joint inferences for groups of coefficients

The inferences in Table 6.8 pertain to the logistic-regression coefficients individually. To make joint inferences about groups of coefficients, we need the methods for multidimensional estimands presented in Section 4.3.3. Of the three methods described there, we will demonstrate the procedure of Meng and Rubin (1992b) for combining likelihood-ratio test statistics.

With complete data, the loglikelihood function for the logistic model (6.2) may be written as

$$l\left(\beta \mid Y_{obs}, Y_{mis}\right) = \sum_{i=1}^{n} \left[z_i \log \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} + \left(1 - z_i\right) \log \frac{1}{1 + e^{x_i^T \beta}} \right],$$

where $z_i = 1$ if individual i has GRD = 2, and $z_i = 0$ otherwise (e.g. McCullagh and Nelder, 1989). Suppose we want to test whether the coefficients for a group of variables (say, LAN_2 and LAN_4) are simultaneously zero. The usual likelihood-ratio test with complete data requires us to fit (a) the full model with all variables, and (b) the reduced model with all variables except LAN_2 and LAN_4. Denote the ML estimates of β under the full and reduced models by $\hat{\beta}$ and $\hat{\beta}$ respectively. For notational convenience, we assume that $\hat{\beta}$ and $\tilde{\beta}$ are of the same length, with the elements of $\tilde{\beta}$ corresponding to the omitted variables set to zero. The likelihood-ratio test statistic is

$$d_L\left(\hat{\beta}, \tilde{\beta} \mid Y_{obs}, Y_{mis}\right) = 2\left[l\left(\hat{\beta} \mid Y_{obs}, Y_{mis}\right) - l\left(\hat{\beta} \mid Y_{obs}, Y_{mis}\right) \right],$$

which, under the reduced model, is approximately distributed as χ_2^2 because the reduced model differs from the full model by two parameters.

The method of Meng and Rubin (1992b) requires two passes through the imputed data. Let $\hat{\beta}^{(t)}$ and $\tilde{\beta}^{(t)}$ denote the ML estimates for the full and reduced models, respectively, fit to the tth imputed dataset. In the first pass, we calculate the likelihood-ratio statistic for each imputed dataset and find their average,

Table 6.9. *Multiple-imputation likelihood-ratio tests for eliminating groups of variables from the regression model*

variables omitted	D_3	k	ν_3	p	$100r_3$	$100\hat{\lambda}$
(a) LAN_2, LAN_4	−0.02	2	59	1.000	341	77
(b) SATV, SATM, ENG	0.40	3	941	0.750	30	23
(c) PRI_L, PRI_Q	1.62	2	461	0.200	36	26

$$\tilde{d}_L = \frac{1}{m} \sum_{t=1}^{m} d_L\left(\hat{\beta}^{(t)}, \tilde{\beta}^{(t)} \mid Y_{obs}, Y_{mis}^{(t)}\right).$$

In the second pass, we calculate the average of the likelihood-ratio test statistics with $\hat{\beta}^{(t)}$ and $\tilde{\beta}^{(t)}$ replaced by their averages,

$$\tilde{d}_L = \frac{1}{m} \sum_{t=1}^{m} d_L\left(m^{-1} \Sigma_{t=1}^{m} \hat{\beta}^{(t)}, m^{-1} \Sigma_{t=1}^{m} \tilde{\beta}^{(t)} \mid Y_{obs}, Y_{mis}^{(t)}\right).$$

The test statistic D_3 and p-value are then found by (4.44)-(4.46).

Using this technique, we tested three groups of variables and removed them from the model in turn after confirming that their p-values were high. The three groups were (a) the language indicators LAN_2 and LAN_4; (b) the test scores SATV, SATM and ENG; and (c) the linear and quadratic contrasts for PRI. Results from each test are shown in Table 6.9, including the test statistic D_3, the degrees of freedom k and v_3 for the F-approximation, the p-value, the relative increase in variance due to nonresponse r_3, and the fraction of missing information $\hat{\lambda}$ calculated as $\hat{\lambda} = r_3/(1-r_3)$. Notice that D_3 for omitting LAN_2 and LAN_4 is slightly less than zero. With complete data, a likelihood-ratio test statistic cannot be negative. With Meng and Rubin’s method, however, negative values do sometimes occur, particularly when the estimates of the coefficients in question are close to zero and their fractions of missing information are high. Multiple-imputation inferences for the coefficients of the final regression model are shown in Table 6.10.

6.4 A simulation study

We have claimed that it is often sensible to use a normal model to create multiple imputations even when the observed data are some

Table 6.10. *Multiple-imputation inferences for logistic-regression coefficients, final model*

variable	$\bar{Q}$	$\sqrt{T}$	$\bar{Q}/\sqrt{T}$	ν	p	$100r$	$100\hat{\lambda}$
intercept	−15.0	2.53	−5.91	160	0.00	53	35
LAN_3	0.874	0.401	2.18	235	0.03	40	29
AGE_2	1.30	0.434	3.01	197	0.00	28	32
SEX_2	0.891	0.405	2.20	398	0.03	28	22
FLAS	0.0351	0.0153	2.29	167	0.02	51	34
MLAT	0.0963	0.0399	2.41	269	0.02	36	27
HGPA	1.99	0.375	5.31	1417	0.00	13	12
CGPA	0.904	0.536	1.68	136	0.09	60	38

what nonnormal. A growing body of evidence supports this claim. The simulation results of Rubin and Schenker (1986), also reported by Rubin (1987, Chap. 4), demonstrate that for estimating the mean of a univariate population, imputations based on a normal model result in interval estimates with excellent repeated-sampling properties. Even for populations that are skewed or heavy-tailed, the actual coverage of multiple-imputation intervals is very close to the nominal coverage, except when the fraction of missing information is high (in excess of 50%). A recent simulation study in the context of a large national health survey produced encouraging results for a wide variety of linear and nonlinear estimators under plausible non-normal populations (Schafer et al., 1996). The study was designed to mimic the specific features of a health examination survey conducted by the U.S. National Center for Health Statistics, including a complex sampling plan with unequal selection probabilities and multiple phases of data collection. Results of that simulation, which involved a mixed model for continuous and categorical variables, will be discussed in Chapter 9. Here we present a miniature version of the simulation to convey the essential result: model-based multiple imputation tends to work well for a wide variety of estimands, and is robust to moderate departures from the data model.

6.4.1 Simulation procedures

Data for this simulation, provided by the National Center for Health Statistics (NCHS), were drawn from Phase 1 of the

Table 6.11. *Variables in the simulation study*

Variable	*Description*
AGE	age group (1=20–39, 2=40–59, 3=60+)
BMI	body mass index (kg/m^2)
HYP	hypertensive (1=no, 2=yes)
CHL	total serum cholesterol (mg/dL)

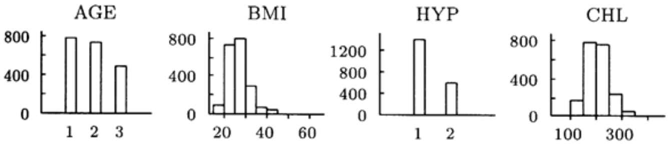


Figure 6.7. *Histograms of AGE, BMI, HYP and CHL in the population.*

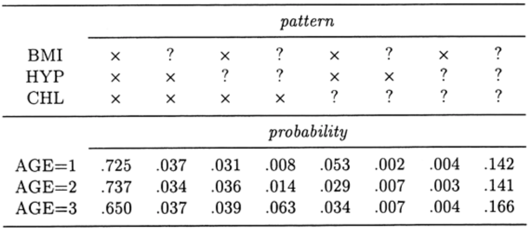
(NCHS, 1994). The data were collected by interviews and medical examinations in mobile examination centers. Because many of the sampled persons did not show up for examination, missingness rates for key exam variables exceeded 30%. To keep matters simple, this study is restricted to adult males (age 20+) and four variables. Definitions of the variables are given in Table 6.11.

An artificial population of 2000 subjects was created by drawing a simple random sample without replacement of all the adult males in the survey who had complete data for all four variables. Histograms for the variables in this population are shown in Figure 6.7. Because the survey used disproportionate sampling in certain racial, ethnic and age categories, and because we have omitted cases with missing data, these 2000 subjects are not representative of any population of substantive interest; the data and results presented here should not be regarded as estimates for any meaningful segment of the U.S. population. This study is meant only to illustrate the properties of model-based multiple imputation when applied to a population of real data that do not conform to simplistic modeling assumptions.

Sampling and response mechanism

From the population of 2000 subjects, simple random samples of size $n = 100$ were drawn without replacement. After a sample was drawn, a random pattern of missingness was imposed on BMI,

Table 6.12 *Probabilities for response patterns by AGE, with observed and missing variables denoted by* × *and* ?, *respectively*



	pattern							
BMI	×	?	×	?	×	?	×	?
HYP	×	×	?	?	×	×	?	?
CHL	×	×	×	×	?	?	?	?
	probability							
AGE=1	.725	.037	.031	.008	.053	.002	.004	.142
AGE=2	.737	.034	.036	.014	.029	.007	.003	.141
AGE=3	.650	.037	.039	.063	.034	.007	.004	.166

HYP and CHL for each sampled person according to his age. The probabilities for the $2^3 = 8$ possible response patterns by age were estimated from all adult males in the MANES III sample, and are shown in Table 6.12. Because the response probabilities depend only on AGE, which is always observed, this mechanism is ignorable. The mechanism creates missingness rates of approximately 20% for each of the three variables BMI, HYP and CHL over repetitions of the sampling procedure.

Imputation

After imposing a pattern of missingness, the 'missing' values were then imputed $m = 5$ times under a multivariate normal model. AGE was entered into the model as two dummy variables: $\text{AGE}_2 = 1$ for AGE = 2 and 0 otherwise; and $\text{AGE}_3 = 1$ for AGE = 3 and 0 otherwise. BMI, HYP and CHL were entered without recoding or transformation. The imputations were created by running five independent chains of data augmentation under the standard noninformative prior (5.18). Each chain was started at the ML estimate and allowed to run for 20 cycles. The final value of Y_{mis} from each chain was taken as an imputation, and the continuous imputes for HYP were rounded off to the nearest category.

6.4.2 Complete-data inferences

After imputing five times, five sets of complete-data point and variance estimates were calculated for a variety of scalar estimands, and the results were combined in the usual way (Section 4.3.2). Eighteen different estimands were examined, including population means, proportions, quantiles, a correlation coefficient and an odds ratio. Methods of complete-data inference for means and proportions are well known. If μ is the mean of a population, and $\bar{y}$ and S^2 are the sample mean and variance, respectively, from a simple random sample of size n, then the standard point and variance estimates are $\bar{y}$ and S^2/n. Similarly, if p is a population proportion and $\hat{p}$ is a sample proportion, the point and variance estimates are $\hat{p}$ and $\hat{p}(1-\hat{p})/n$. Complete-data inferences for quantiles, correlations and odds ratios are described below.

Quantiles

The following approximate method for quantiles was described by Woodruff (1952). Suppose that Q is the pth quantile of a distribution function F, and $\hat{Q}$ is an estimate of Q based on a simple random sample of size n. Then

$$Q_1 \le Q \le Q_2$$

will be true if and only if

$$F(Q_1) \le p \le F(Q_2),$$

because F and F^{-1} are strictly increasing. Rather than finding an interval estimate for Q directly, we instead construct an interval estimate for the proportion of the population that lies below Q, and then translate the endpoints of this interval into quantiles. For example, an approximate 95% interval for p ranges from

$$p_1 = p - 2\sqrt{\frac{p(1-p)}{n}} \text{ to } p_2 = p + 2\sqrt{\frac{p(1-p)}{n}}.$$

If we set Q_1 and Q_2 equal to the p_1ith and p_2ith sample quantiles, respectively, then the approximate 95% confidence interval for Q ranges from Q_1 to Q_2. This interval is not necessarily symmetric about $\hat{Q}$. It is well known, however that under mild smoothness conditions for F the sample quantiles are asymptotically normally distributed (e.g Serfling, 1980), and for large samples we can take $(Q_2 - Q_1)/4$ as an estimated standard deviation for $\hat{Q}$ (Francisco and Fuller, 1991).

Correlation coefficients

Suppose that r is a correlation coefficient from a simple random sample of n units, and ρ is the corresponding population value.
The familiar transformation due to Fisher (1921),

$$z(r) = \tanh^{-1}(r) = \tfrac{1}{2}\log\frac{1+r}{1-r},$$

makes $z(r)$ approximately normally distributed about $z(\rho)$ with variance $1/(n-3)$. This result is derived under an assumption of bivariate normality. An interval estimate for ρ can be calculated by first finding an interval for $z(\rho)$ using the normal approximation, and then applying the inverse transformation $z^{-1}(\cdot) = \tanh(\cdot)$ to the endpoints. Because $z(r) \approx r$ for values of r near zero (they agree to two decimal places for $|r|<0.24$), the approximation $V(r) \approx 1/(n-3)$ is also acceptable in the vicinity of $r = 0$.

Odds ratios

Suppose that Y_1 and Y_2 are two binary variables taking values 1 and 2. In a simple random sample of size n, let x_{ij} be the number of sample units for which $Y_1 = i$ and $Y_2 = j$, $i, j = 1, 2$. The population odds ratio, defined as

$$\omega = \frac{P(Y_1 = 1 \mid Y_2 = 1) / P(Y_1 = 2 \mid Y_2 = 1)}{P(Y_1 = 1 \mid Y_2 = 2) / P(Y_1 = 2 \mid Y_2 = 2)},$$

$$p_2 = p + 2\sqrt{\frac{p(1-p)}{n}}.$$

is estimated by $\hat{\omega} = (x_{11}x_{22})/(x_{12}x_{21})$. In large samples, the log odds ratio $\hat{\beta} = \log \hat{\omega}$ is approximately normally distributed about $\beta = \log\omega$, and a large-sample variance estimate for β is $x_{11}^{-1} + x_{12}^{-1} + x_{22}^{-1}$ (e.g. Agresti, 1990). An interval estimate for ω can be obtained by first finding an interval for β using the normal approximation, and then taking antilogs of the endpoints.

6.4.3 Results

The entire simulation procedure of drawing a sample, imposing patterns of missingness, creating five imputations and calculating point and interval estimates was carried out 1000 times. The results are summarized in Table 6.13. For each of the eighteen estimands, this table shows the true estimand Q (i.e. the population value), the multiple-imputation point estimates $\bar{Q}$, the endpoints of the nominal 95% interval estimates (low and high) and the estimated fraction of missing information $\bar{\lambda}$ averaged over 1000 iterations. In addition, the table reports the simulated actual coverage (cvg.), the number of intervals out of 1000 that covered the true estimand. The average simulated coverage across all eighteen estimands is 952.7, indicating that the procedure is well calibrated. Some of the

Table 6.13. *Summary of simulation results for eighteen estimands*

Estimand	Q	$\bar{Q}$	low	high	cvg.	$100\hat{\lambda}$
Mean BMI						
overall	26.6	26.6	25.5	27.7	956	25
AGE = 1	25.7	25.7	24.0	27.4	941	22
AGE = 2	27.7	27.7	25.9	29.5	942	22
AGE = 3	26.3	26.3	24.1	28.5	961	34
Mean CHL						
overall	206	207	197	216	956	22
AGE = 1	192	192	177	206	960	25
AGE = 2	219	220*	204	235	949	20
AGE = 3	210	211	191	230	956	24
Proportion HYP = 2						
overall	.294	.299*	.197	.402	959	22
AGE = 1	.107	.117*	.000	.235	951	26
AGE = 2	.323	.329*	.156	.502	941	20
AGE = 3	.545	.540	.304	.776	927	27
Percentiles						
BMI (50%)	26.0	26.1*	24.8	27.4	951	24
BMI (90%)	32.7	32.8*	30.1	35.5	960	19
CHL (50%)	202	204*	192	216	961	20
CHL (90%)	262	264*	241	287	960	19
Correlation						
BMI and CHL	.171	.174	−.064	.393	940	29
Odds ratio						
BMI > 27.8 by HYP	1.64	1.81	.614	5.47	977	27

* denotes a point estimate with statistically significant bias

multiple-imputation point estimates, those denoted by an asterisk, have a statistically significant bias; for these, the average of the 1000 values of $\overline{Q}$ was significantly different from Q at the 0.05 level as judged by an ordinary t-test. But the biases are minor when compared to the average width of the 95% interval estimates, and thus are of little consequence.

Multiple imputation performs well in this example even though the normality assumption of the imputation model is clearly violated: the distributions of BMI and CHL are skewed to the right, and CHL is binary. In practice, one would probably transform BMI

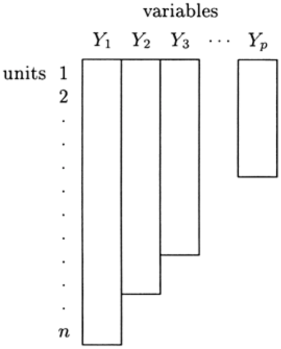


Figure 6.8. *Monotone missingness pattern.*

and CHL (e.g. to the log scale) before applying the normal model, in which case the performance should be even better.

6.5 Fast algorithms based on factored likelihoods

6.5.1 Monotone missingness patterns

This section presents a class of simulation algorithms for incomplete multivariate normal data which, in certain cases, will achieve stationarity more rapidly than ordinary data augmentation. These algorithms are based on the observation, first made by Li (1988), that we do not really need to fill in the entire set of missing data Y_{mis} at each I-step. The function of the I-step is to impute enough of the missing values to make the P-step into a tractable, complete-data posterior simulation. Under the multivariate normal model, however, the P-step can be made tractable by filling in only enough of the missing values to complete a *monotone pattern*.

The missingness pattern for a data matrix is said to be monotone if, whenever an element y_{ij} is missing, y_{ik} is also missing for all $k > j$ (Rubin, 1974; Little and Rubin, 1987). A monotone pattern is shown in Figure 6.8. Monotone patterns often arise in repeated-measures or longitudinal datasets, because if a subject drops out of the study in a given time period, then his or her data will typically be missing in all subsequent time periods. Sometimes a non-monotone dataset can be made monotone or nearly so by reordering the variables according to their missingness rates. Let n_j denote the number of rows of the data matrix for which Y_j is observed. If the pattern is monotone, then $n_p \leq n_p{-}1 \leq \cdots \leq n_1 = n$. We will assume that the rows of the monotone dataset have been sorted as in Figure 6.8, so that Y_j (and hence $Y_1,\ldots, Y_{j-1}$ as well) is observed for rows $1,\ldots,n_j$ and missing for rows $n_j + 1,\ldots,n$.

Factoring the observed-data likelihood

When the observed data Y_{obs} are monotone, the observed-data likelihood function can be expressed in a very convenient form. Let $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$, where ϕ_1 denotes the parameters of the marginal distribution of variable Y_1, ϕ_2, the parameters of the conditional distribution of Y_2 given Y_1, ϕ_3 the parameters of the conditional distribution of Y_3 given Y_1 and Y_2, and so on. In other words, ϕ_j contains the intercept, slopes and residual variance from the normal linear regression of Y_j on $\mathrm{Y}_1,\ldots,Y_{j-1}$. It is easy to show that $\phi = \phi(\theta)$ is a one-to-one function of the usual parameters $\theta = (\mu, \Sigma)$. Moreover, if no prior restrictions are imposed upon θ, then the components $\phi_1,\ldots,\phi_p$ are distinct in the sense that the parameter space of ϕ is the cross-product of the individual parameter spaces for $\phi_1,\ldots,\phi_{p.}$ Expressions for $\phi_1,\ldots,\phi_p$ in terms of $\theta = (\mu, \Sigma)$ can be found by partitioning μ and Σ and applying the formulas given in Section 5.2.4.

When Y_{obs} is monotone, the observed-data likelihood function for ϕ factors neatly into independent likelihoods for $\phi_1,\ldots,\phi_p$. To see this, notice that the joint density of the variables $Y_1,\ldots,Y_p$ can be factored as

$$P(Y_1,\ldots,Y_p \mid \phi) = P(Y_1 \mid \phi_1)P(Y_2 \mid Y_{1,}\phi_2) \\ \cdots P(Y_p \mid Y_1,\ldots,Y_{p-1},\phi_p),$$

which allows us to write the complete-data likelihood as

$$\begin{aligned} L(\phi \mid Y) &= \prod_{i=1}^{n} P(y_{i1},\ldots,y_{ip} \mid \phi) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{p} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j) \qquad (6.3) \\ &= \prod_{j=1}^{p}\prod_{i=1}^{n} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j). \end{aligned}$$

The inner product in (6.3),

$$\prod_{i=1}^{n} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j),$$

can also be written

$$\prod_{i=1}^{n_j} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j) \prod_{i=n_j+1}^{n} P(y_{ij} \mid y_{i1},\ldots,y_{i,j-1},\phi_j). \qquad (6.4)$$

The observed-data likelihood $L(\phi|Y_{obs})$ is by definition the integral of (6.3) over Y_{mis}. But notice that the first product in (6.4) does not involve Y_{mis}, because variable Y_j is observed in rows 1,..., n_j, whereas the second product integrates to unity because Y_j is missing in rows n_j + 1,..., n. It follows that

$$L(\phi \mid Y_{obs}) = \prod_{j=1}^{p} L(\phi_j \mid Y_{obs}), \qquad (6.5)$$

where

$$L(\phi_j \mid Y_{obs}) = \prod_{i=1}^{n_j} P(y_{ij} \mid y_{i1}, \ldots, y_{i,j-1}, \phi_j). \tag{6.6}$$

Under the multivariate normal model, (6.6) is simply the likelihood for the normal linear regression of Y_j on $Y_1, \ldots, Y_{j-1}$, based on the rows $1, \ldots, n_j$ of the data matrix. Thus the factorization (6.5) effectively reduces the problem of inference about ϕ to a sequence of complete-data regressions over subsets of the rows of the data matrix.

6.5.2 Computing alternative parameterizations

When the data are monotone, the observed-data likelihood has a convenient form when expressed in terms of $\phi = (\phi_1, \ldots, \phi_p)$. The parameters of the multivariate normal, however, are usually expressed in terms of $\theta = (\mu, \Sigma)$, a vector of means and a covariance matrix. To make use of the convenient form of the likelihood, we will need to switch back and forth between the two parameterizations.

A numerical procedure for computing $\phi = \phi(\theta)$ or $\theta = \phi^{-1}(\phi)$ can be formulated in terms of the sweep operator (Section 5.2.4). For convenience, we introduce a slight generalization of sweep which gives a compact notation to the process of sweeping a square submatrix of a larger matrix. Let G be a $p \times p$ symmetric matrix with elements g_{ij}, and let A be a subset of the p columns (and rows) of G. The generalized sweep operator SWP_A performs the usual sweep computations on the rows and columns of G in the set A but leaves the rest of G unchanged. Formally, $\text{SWP}_A[k]$ for some $k \in A$ operates on G by replacing it with another $p \times p$ matrix H,

$$H = SWP_A[k]G,$$

where the elements of H are given by

$$h_{kk} = -1 / g_{kk},$$

$$h_{jk} = h_{kj} = \begin{cases} g_{jk} / g_{kk} & j \in A, j \neq k \\ g_{jk} & j \notin A, \end{cases}$$

$$h_{jl} = h_{lj} = \begin{cases} g_{jl} - g_{jk}\, g_{kl} \; / \, g_{kk} & j \in A, l \in A, j \neq k, l \neq k, \\ g_{jl} & j \notin A \; or \; l \notin A. \end{cases}$$

This operation will be referred to as *sweeping submatrix A of G on position k*. Similarly, the corresponding reverse sweep operator RSW$_A$ applies the usual reverse-sweep computations to the rows and columns of G in set A, while leaving the rest of G unchanged. Formally, RSW$_A$[k] for some $k \in$ A operates on G by replacing it with another $p \times p$ matrix H,

$$H = RSW_A[k]G,$$

where the elements of H are given by

$$h_{kk} = -1 / g_{kk},$$

$$h_{jk} = h_{kj} = \begin{cases} -g_{jk} / g_{kk} & j \in A, j \neq k \\ g_{jk} & j \notin A, \end{cases}$$

$$h_{jl} = h_{lj} = \begin{cases} g_{jl} - g_{jk}\, g_{kl} \; / \, g_{kk} & j \in A, l \in A, j \neq k, l \neq k, \\ g_{jl} & j \notin A \; or \; l \notin A. \end{cases}$$

When the sweep or reverse sweep operators are written without a subscripting set, as in SWP[k] or RSW[k], it will be understood that the operation is being applied to the entire matrix.

We are now ready to give a compact notation to the process of computing ϕ from θ and vice-versa. Let

$$\phi_j = \left(\beta_j^T, \gamma_j\right)^T \tag{6.7}$$

where β_j is the $j \times 1$ vector of coefficients (including the intercept) from the linear regression of Y_j on Y_1, $Y_2,\ldots,Y_{j-1}$ and γ_j is the residual variance, so that

$$Y_j \mid Y_1,\ldots,Y_{j-1},\phi_j \sim N\left(\left(1, Y_1,\ldots,Y_{j-1}\right)\beta_j,\gamma_j\right), \qquad (6.8)$$

Let μ_j denote the jth element of μ and σ_{jk} the (j, k) element of Σ, and define

$$\theta_j = \left(\mu_j,\sigma_{1j},\sigma_{2j},\ldots,\sigma_{jj}\right)^T, \qquad (6.9)$$

so that $\theta = (\theta_1,\theta_2,\ldots,\theta_p)$. As in Section 5.2.4, let us express θ as a symmetric $(p + 1) \times (p + 1)$ matrix,

$$\theta = \begin{bmatrix} -1 & \mu^T \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} -1 \end{bmatrix} & \begin{bmatrix} \theta_1 \\ \end{bmatrix} & \begin{bmatrix} \\ \theta_2 \\ \end{bmatrix} & \begin{bmatrix} \\ \theta_3 \\ \\ \end{bmatrix} & \cdots \\ & & & & \\ & & & & \end{bmatrix}$$

where the lower portion of the matrix is not shown to avoid redundancy. Finally, let the row and column labels for this matrix run from 0 to p, so that θ_j appear in column j. To convert θ to

$$\phi = \begin{bmatrix} \begin{bmatrix} -1 \end{bmatrix} & \begin{bmatrix} \phi_1 \\ \end{bmatrix} & \begin{bmatrix} \\ \phi_2 \\ \end{bmatrix} & \begin{bmatrix} \\ \phi_3 \\ \\ \end{bmatrix} & \cdots \\ & & & & \\ & & & & \end{bmatrix}$$

note that sweeping θ on positions 1,2,..., j–1 produces a new matrix whose jth column is ϕ_j. Therefore, if we sweep the full θ matrix on positions 1, 2,..., p–1, then $\phi_{(p)}$ appears in the pth column. If we then reverse-sweep all but the last row and column on position p–1, then ϕ_{p-1} appears in column p–1. Reverse-sweeping all but the last two rows and columns on position p–2 makes ϕ_{p-2} appears in column p–2, and so on.

This procedure can be expressed very concisely in pseudocode. Let $A_j = \{0,1,...,j\}$ for $j = 1, 2,..., p$. The following two lines will overwrite a θ matrix, replacing it with $\phi = \phi(\theta)$.

```
for $j$:=1 to $p-1$ do $\theta$:=SWP[$j$]$\theta$

for $j$:=p–1 down to 1 do $\theta$:=RSW$_{Aj}$[$j$]$\theta$
```

The transformation from ϕ back to θ is simply a reversal of these steps. The following two lines will overwrite a ϕ matrix, replacing it with $\theta = \phi^{-1}(\phi)$.

```
for $j$:=1 to $p-1$ do $\phi$=SWP$_{Aj}$ [$j$] $\phi$

for $j$:=p–1 down to 1 do $\phi$:=RSW[$j$] $\phi$
```

6.5.3 Noniterative inference for monotone data

Maximum-likelihood estimation

When Y_{obs} has a monotone pattern, the factorization of the likelihood in terms of $\phi=(\phi_1,...,\phi_p)$,

$$L(\phi \mid Y_{obs}) = \prod_{j=1}^{p} L(\phi_j \mid Y_{obs}),$$

enables us to calculate ML estimates without iteration (Little and Rubin, 1987, Chapter 6). Because the parameters $\phi_1,\ldots,\phi_p$ are distinct, maximizing $L(\phi|Y_{obs})$ is equivalent to maximizing each factor $L(\phi_j|Y_{obs})$ separately for j=1,...,p. The ML estimate of ϕ is $\hat{\phi}$, where $\hat{\phi}_j$ is the maximizer of $L(\phi_j|Y_{obs})$.

The maximization of each factor $L(\phi_j|Y_{obs})$ is accomplished by ordinary least-squares regression of Y_j on $Y_1,\ldots, Y_{j-1}$, using rows 1,..., n_j of the data matrix. Let z_j denote the observed data in column j,

$$z_j = \left(y_{1j}, y_{2j}, \ldots, y_{n_j,j}\right)^T, \tag{6.10}$$

and X_j the upper-left n_j x (j–1) submatrix augmented by a column of ones,

$$X_j = \begin{bmatrix} 1 & y_{11} & y_{12} & \cdots & y_{1,j-1} \\ 1 & y_{21} & y_{22} & \cdots & y_{2,j-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n_j,1} & y_{n_j,2} & \cdots & y_{n_j,j-1} \end{bmatrix}. \tag{6.11}$$

By (6.8), the conditional distribution of z_j given X_j and ϕ_j is

$$z_j \mid X_j, \phi_j \sim N\left(X_j, \beta_j, \gamma_j I\right),$$

so the likelihood for ϕ_j is

$$L(\phi_j \mid Y_{obs}) \propto \gamma_j^{-n_j/2} \exp\left\{-\frac{1}{2\gamma_j}\left(z_j - X_j\beta_j\right)^T \left(z_j - X_j\beta_j\right)\right\}.$$

Using well-known properties of the normal linear regression model, the ML estimate of ϕ_j is given by

$$\hat{\beta}_j = \left(X_j^T X_j\right)^{-1} X_j^T z_j, \tag{6.12}$$

$$\hat{\gamma}_j = n_j^{-1}\, \hat{\in}_j^T \hat{\in}_j, \tag{6.13}$$

where $\hat{\in}_j = z_j - X_j \hat{\beta}_j$ (e.g. Draper and Smith, 1981). Notice that $\hat{\gamma}_j$, the ML estimate of the residual variance, is biased because its denominator is n_j rather than n_j–j. Calculating (6.12)-(6.13) for j = 1, 2,..., p yields $\hat{\phi}$, the ML estimate of ϕ. Because ML estimates are invariant under transformations of the parameter, the ML estimate for θ can be calculated as $\hat{\theta} = \phi^{-1}\left(\hat{\phi}\right)$.

Bayesian inference

Similarly, when Y_{obs} has a monotone pattern, we can also conduct Bayesian inferences without iteration provided that the prior distribution has a certain form. If we apply a prior density to ϕ that factors into independent densities,

$$\pi(\phi) = \pi_1(\phi_1)\pi_2(\phi_2)\cdots\pi_p\left(\phi_p\right), \tag{6.14}$$

then it is obvious that the posterior distribution $P(\theta|Y_{obs})$ will also factor into independent posteriors for ϕ_1,...,ϕ_p, a structure that Rubin (1987) calls *monotone distinct*. Bayesian inferences for ϕ can then be carried out as a sequence of independent inferences based on the posteriors

$$P\left(\phi_j \mid Y_{obs}\right) \propto L\left(\phi_j \mid Y_{obs}\right)\pi_j\left(\phi_j\right)$$

for j = 1,..., p. For example, we can simulate a value of ϕ from $P(\theta|Y_{obs})$ by drawing ϕ_j from $P(\theta_j|Y_{obs})$ independently for j =

1,...,J. A simulated value of θ from $P(\theta|Y_{obs})$ can then be obtained by applying the back-transformation $\theta = \phi^{-1}(\phi)$ to the simulated value of ϕ.

The noninformative prior most commonly used for multivariate normal data,

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}, \tag{6.15}$$

can be factored as in (6.14). To avoid confusion, let us refer to the density (6.15) as $\pi_\theta(\theta)$, and the corresponding density for ϕ induced by (6.15) as $\pi_\theta(\phi)$. The relationship between π_θ and π_ϕ is

$$\pi_\phi(\phi) = \pi_\theta\left(\phi^{-1}(\phi)\right)\|J\|^{-1}, \tag{6.16}$$

where $\theta=\phi_{-1}(\phi)$ is the inverse of the transformation $\phi=\phi(\theta)$, J is the Jacobian or first-derivative matrix of the transformation $\phi=\phi(\theta)$ and $\|J\|$ is the absolute value of the determinant of J. By a well known property of determinants, $|\Sigma|$ can be written as

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}|\,\left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|$$

for square submatrices Σ_{11} and Σ_{22}. But $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is the residual covariance matrix from the regression of the variables corresponding to Σ_{22} on the variables corresponding to Σ_{11} (Section 5.2.4). Taking $\Sigma_{22}=\sigma_{pp}$, the determinant of Σ becomes

$$|\Sigma| = |\Sigma_{11}| \gamma_p, \tag{6.17}$$

where Σ_{11} is Σ without the last row and column. Applying (6.17) recursively to Σ_{11} leads to

$$|\Sigma| = \prod_{j=1}^{p} \gamma_j. \tag{6.18}$$

To find $\pi_{\phi},(\phi)$ we also need to evaluate $\|J\|$. In Section 5.4.2, we derived the determinant of the Jacobian that arises when we condition on a subset of the variables $Y_1,\ldots, Y_p$. Suppose that we first transform θ to the intermediate parameter (ξ_{p-1},ϕ_p), where ξ_{p-1} represents the portions of μ and Σ pertaining to the marginal distribution of $Y_1,\ldots, Y_{p-1}$, andϕ_p pertains to the regression of Y_p on Y_{p-1} (5.28), the determinant of the Jacobian for going from θ to (ξ_{p-1},ϕ_p) is $|\Sigma_{11}|^{-1}$, where Σ_{11} is the covariance matrix for $Y_1,\ldots,Y_{p-1}$. But $|\Sigma_{11}|=$ $\gamma_1,\gamma_2\cdots\gamma_{p-1}$, so the determinant of the Jacobian of this intermediate transformation is $(\gamma_1,\gamma_2\cdots\gamma_{p-1})^{-1}$. If we then transform ξ_{p-1} to (ξ_{p-2},ϕ_{p-1}), where ξp_{-2} contains the portions of μ and Σ pertaining to $Y_1,\ldots,Y_{p-2}$ and ϕ_{p-1} pertains to the regression of Y_{p-1} on $Y_1,\ldots,Y_{p-2}$. the determinant of the Jacobian is $\gamma_1\gamma_2\cdots\gamma_{p-2}$. We can repeat this procedure until we have reached the final parameterization $\phi=(\phi_1,\ldots,\phi_p)$, and the determinant of the Jacobian for $\phi=\phi(\theta)$ will be the product of the determinants for each of the intermediate transformations. The result is

$$\|J\| = \gamma_1^{-(p-1)}\gamma_2^{-(p-2)}\cdots\gamma_{p-1}^{-1}. \tag{6.19}$$

Substituting (6.19) and (6.18) into (6.16) gives

$$\pi_{\phi}(\phi) \propto \prod_{j=1}^{p} \gamma_j^{-\left(\frac{p+1}{2} p+j\right)} \tag{6.20}$$

as the prior density for $\phi = \phi(\theta)$ induced by (6.15).

Now we show the posterior that results when this prior is combined with the observed-data likelihood from a monotone dataset. Consider the likelihood factor for ϕ_j,

$$L\left(\phi_j \mid Y_{obs}\right) \propto \gamma_j^{-n_j/2} \exp\left\{-\frac{1}{2\gamma_j}\left(z_j - X_j\beta_j\right)^T \left(z_j - X_j\beta_j\right)\right\}.$$

With some algebraic manipulation, it can be shown that

$$\left(z_j - X_j\beta_j\right)^T \left(z_j - X_j\beta_j\right) = \hat{\in}_j^T \hat{\in}_j + \left(\beta_j - \hat{\beta}_j\right)^T X^T X\left(\beta_j - \hat{\beta}_j\right),$$

where $\hat{\beta}_j = \left(X^T X_j\right)^{-1} X_j^T z_j$ and $\hat{\in}_j = X_j \hat{\beta}_j$. When L($\phi_j | Y_{obs}$) is combined with the factor in (6.20) involving ϕ_j,

$$\pi_j\left(\phi_j\right) \propto \gamma_j^{-\left(\frac{p+1}{2} - p+j\right)},$$

the resulting posterior can be written as

$$P\left(\phi_j \mid Y_{obs}\right) \propto \gamma_j^{-j/2} \exp\left\{-\frac{1}{2\gamma_j}\left(\beta_j - \hat{\beta}_j\right)^T X^T X\left(\beta_j - \hat{\beta}_j\right)\right\}$$

$$\times \gamma_j^{-\left(\left(n_j - p + j - 1\right)/2\right)-1} \exp\left\{-\frac{1}{2\gamma_j} \hat{\in}_j^T \hat{\in}_j\right\},$$

which is the product of a multivariate normal and a scaled inverted-chisquare density,

$$\beta_j \mid \gamma_j, Y_{obs} \sim N\left(\hat{\beta}_j, \gamma_j\left(X^T X\right)^{-1}\right), \tag{6.21}$$

$$\gamma_j \mid Y_{obs} \sim \hat{\in}_j^T \hat{\in}_j \, \chi^{-2}_{n_j - p + j - 1}. \tag{6.22}$$

6.5.4 Monotone data augmentation

Thus far we have discussed methods of inference that are appropriate when the observed data Y_{obs} are monotone. It often happens in practice that a dataset is not precisely monotone, but would become monotone if a relatively small portion of the missing data were filled in. This situation often arises with double sampling, where investigators attempt to measure certain variables for all units in a sample, and then measure additional variables for only a subsample. If there were no missing values except for those missing by design, then the data would be perfectly monotone; in practice, however, there is usually some additional unplanned missingness which makes the overall pattern deviate slightly from monotonicity. Near-monotonicity also results in many longitudinal or panel studies, in which variables are measured for individuals on multiple occasions. Subjects who drop out of the study at a particular occasion or wave usually do not reappear in subsequent waves, so that if the variables are ordered by wave the overall pattern is nearly monotone.

When this situation arises, we can exploit the near-monotone pattern to devise simulation algorithms that are computationally more efficient than the data augmentation procedures described in Chapter 5. These new procedures, which we call *monotone data augmentation*, differ from ordinary data augmentation in that they fill in only enough of the missing data at each I-step to complete a monotone pattern. Suppose that we partition the missing data as $Y_{mis} = (Y_{mis*}, Y_{mis**})$ where Y_{mis*} is some subset of the missing values which, if filled in, would result in (Y_{obs}, Y_{mis*}) having a monotone pattern. Monotone data augmentation proceeds in the following two steps.

1. I-step: Given the current simulated value $\theta^{(t)}$ of the parameter, draw a value from the conditional predictive distribution of Y_{mis*},

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis}{}^{*} \mid Y_{obs}, \theta^{(t)}\right). \qquad (6.23)$$

2. P-step: Conditioning on $Y_{mis*}^{(t+1)}$, draw a new value of θ from its posterior given the now-completed monotone pattern,

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis*}^{(t+1)}\right). \qquad (6.24)$$

In practice, the P-step will have to be carried out using the parameterization $\phi=(\phi_1,...,\phi_p)$ that corresponds to the monotone pattern of (Y_{obs}, Y_{mis*}). That is, we will have to draw

$$\phi^{(t+1)} = (\phi_1^{(t+1)},...,\phi_p^{(t+1)})$$

by drawing

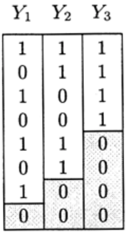
$$\phi_j^{(t+1)} \sim P\left(\phi_j \mid Y_{obs}, Y_{mis*}^{(t+1)}\right)$$

independently for $j = 1, 2,..., p$, and then calculate

$$\theta^{(t+1)} = \phi^{-1}\left(\phi^{(t+1)}\right)$$

using the procedures for numerical transformation described earlier in this section.

Monotone data augmentation has two computational advantages over ordinary data augmentation. First, it requires fewer random number draws per iteration, i.e. it is typically faster to fill in Y_{mis*} than the full Y_{mis}. Second, it will achieve approximate stationarity in fewer iterations. Liu, Wong and Kong (1994) show that `collapsing' the data augmentation by drawing only a subset of the unknown quantities at each iteration leads to faster convergence



Y_1	Y_2	Y_3
1	1	1
0	1	1
1	0	1
0	0	1
1	1	0
0	1	0
1	0	0
0	0	0

Figure 6.9. *Possible missingness patterns for a three-variable dataset, with observed and missing variables denoted by 1 and 0, respectively.*

and smaller autocorrelations between successive iterates. With ordinary data augmentation, convergence is governed by the amount of information contained in Y_{mis} relative to Y_{obs} (Section 3.5.3). With monotone data augmentation, however, convergence is governed by the amount of information in Y_{mis*} relative to Y_{obs}. When Y_{obs} is not far from monotone, Y_{mis*} is relatively small; the distribution $P(\theta|Y_{obs},Y_{mis*})$ is then nearly independent of Y_{mis*}, and only a few steps of monotone data augmentation will be needed to achieve approximate stationarity. In the extreme case where Y_{obs} is precisely monotone, Y_{mis*} is empty and the algorithm reaches stationarity in one step.

Monotone data augmentation was first proposed by Li (1988) who demonstrated its use in simple bivariate examples. The algorithm presented here, which assumes multivariate normal data and the customary noninformative prior

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)},$$

has also been described by Liu (1993).

Choosing the monotone pattern to be completed

To identify a Y_{mis*} it helps to group the rows of the data matrix by their patterns of missingness. For example, the possible patterns of missingness for a three-variable dataset are shown in Figure 6.9. The missing values in the unshaded region constitute Y_{mis*} and need to be filled in at every I-step; missing data in the shaded region constitute Y_{mis**} and do not need to be filled in.

In most cases, of course, there is no unique set of missing data Y_{mis*} that will complete a monotone pattern. By simply reordering the columns Y_1, Y_2,..., Y_p of the data matrix, we can identify alternative sets of missing values that are candidates for Y_{mis*}. For computational efficiency, it is advantageous to choose Y_{mis*} to be ’small’ in two senses. First, the actual number of missing values contained in Y_{mis*}. should be small, to reduce the number of random variates that need to be drawn at each I-step. Second, Y_{mis*} should contain as little information as possible about the unknown parameters, to reduce the number of steps required to achieve approximate stationarity. These two objectives may sometimes conflict. In a normal dataset, for example, there may be a tradeoff between filling in a large number of relatively noninfluential observations and filling in a smaller number with high leverage. Finding a set Y_{mis*} to maximize the efficiency of the algorithm is a difficult problem, as it involves questions about the convergence of Markov chain Monte Carlo algorithms that are not easy to answer at present. Moreover, finding such an optimal set may itself require substantial computation, offsetting the potential gains of a more efficient algorithm.

To choose Y_{mis*}, we suggest the naive approach of simply ordering the columns of Y according to their fractions of missing observations. That is, choose Y_1 to be the variable with the fewest missing values, Y_2 the variable with the second fewest, and so on. This approach is attractive because it is computationally trivial. Moreover, it has the feature that if Y_{obs} is already monotone, it will find the monotone pattern and identify Y_{mis*} to be empty.

6.5.5 Implementation of the algorithm

In discussing how to implement monotone data augmentation for the multivariate normal model, we will need to build on the bookkeeping notation of Chapter 5. Suppose that the rows of the data matrix have been grouped together according to their patterns of missingness as shown in Figure 6.10. Index the missingness patterns by $s = 1, 2,..., S$. Let s_j denote the last pattern for which variable Y_j may need to be filled in to complete the overall monotone pattern, so that

$$S = s_1 \geq s_2 \geq \cdots \geq s_p.$$

Following Section 5.3.1, let

$$r_{sj} = \begin{cases} 1 \text{ if } Y_j \text{ is observed in pattern } s, \\ 0 \text{ if } Y_j \text{ is missing in pattern } s. \end{cases}$$

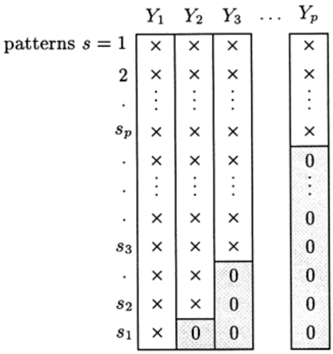


Figure 6.10. *Arrangement of missingness patterns for monotone data augmentation, with 0 denoting a variable that is missing and × denoting a variable that is either observed or missing.*

Let $O(s)$ and $M(s)$ denote the column labels corresponding to variables that are observed and missing, respectively, in pattern s,

$$O(s) = \{j : r_{sj} = 1\}$$

$$M(s) = \{j : r_{sj} = \mathrm{o}\}$$

Also, let $M^*(s)$ denote the subset of $M(s)$ that must be filled in to complete the monotone pattern, and let $M^{**}(s)$ be the remainder of $M(s)$,

$$M^*(s) = \{j : r_{sj} = 0 \text{ and } s_j \geq s\}$$

$$M^{**}(s) = \{j : r_{sj} = 0 \text{ and } s_j < s\}$$

For any s, $M^*(s)$ lists the columns with missing values in the unshaded region of Figure 6.10, and $M^{**}(s)$ lists the columns in the shaded region. Finally, let $I(s)$ denote the subset of $\{1, 2,\ldots, n\}$ corresponding to the rows of the data matrix Y in pattern s.

The I- and P-steps

The I-step for monotone data augmentation is nearly identical to the I-step for ordinary data augmentation; the only difference is that rather than imputing all the missing values Y_{mis}, we need only impute the portion Y_{mis^*} to complete the monotone pattern. Consequently, the pseudocode for the I-step shown in Figure 5.6 can be used for monotone data augmentation with only one modification: replace every occurrence of $M(s)$ with the potentially smaller set $M^*(s)$. The four lines of code in Figure 5.6 preceded by the character 'C' are not needed and may be removed.

The P-step, however, is computationally more complicated than the P-step for ordinary data augmentation, because the posterior distributions of $\phi_1,\phi_2,\ldots,\phi_p$ depend on different sets of sufficient statistics. The posterior of ϕ_j, given by (6.21)-(6.22), depends on $\hat{\beta}_j\left(X_j^T X_j\right)^{-1}$, and $\hat{\in}_j^T\hat{\in}_j$, which are obtained from the regression of Y_j on $Y_1,\ldots,Y_{j-1}$ over the rows of the data matrix in missingness patterns $s = 1,\ldots, s_j$. To perform this regression, we need to accumulate sums of squares and cross-products for variables $Y_1,\ldots,Y_j$ and patterns $s = 1,\ldots, s_j$.

As in Section 5.3.3, define $T(s)$ to be the $(p + 1) \times (p + 1)$ matrix of complete-data sufficient statistics from missingness pattern s,

$$T(s) = \begin{bmatrix} n_s & \Sigma y_{i1} & \Sigma y_{i2} & \cdots & \Sigma y_{ip} \\ & \Sigma y_{i1}^2 & \Sigma y_{i1} y_{i2} & \cdots & \Sigma y_{i1} y_{ip} \\ & & \Sigma y_{i2}^2 & \cdots & \Sigma y_{i2} y_{ip} \\ & & & \ddots & \vdots \\ & & & & \Sigma y_{ip}^2 \end{bmatrix},$$

where all sums are taken over i$\in I(s)$, and $n_s = \sum_{i\in I(s)}$ is the sample size in pattern s. For simplicity, we will number the rows and columns of $T(s)$ from 0 to p rather than from 1 to p + 1. Let T_{mis} (s) and T_{obs} (s) be matrices of the same size as $T(s)$ with elements defined as follows: the (j, k)th element of T_{mis} (s) is equal to the (j, k)th element of T(s) if $j\in M(s)$ or $k\in M(s)$, and zero otherwise; and $T_{obs}(s)$ = T(s) − $T_{mis}(s)$. Notice that $T_{mis}(s)$ contains the sufficient statistics that depend on Y_{mis}, whereas $T_{obs}(s)$ contains the sufficient statistics that are functions only of Y_{obs}. Finally, let $T_{mis*}(s)$ be a matrix identical to $T_{mis}(s)$, but with the following exception: the rows and columns corresponding to variables that are not needed to complete the monotone pattern are set to zero. That is, set the (j, k)th element of $T_{mis*}(s)$ equal to zero if $j\in M^{**}(s)$ or $k\in M^{**}(s)$, otherwise set it equal to the (j, k)th element of $T_{mis}(s)$. Thus T_{mis*} contains the sufficient statistics that depend on Y_{mis*} but not on Y_{mis**}.

Suppose that the unknown values in Y_{mis*} have been filled in by an I-step so that $T_{mis*}(s)$ can be calculated. If we let

$$T_j = \sum_{s=1}^{s_j} T_{obs}(s) + \sum_{s=1}^{s_j} T_{mis*}(s),$$

then

$$T_j = \begin{bmatrix} X_j^T X_j & X_j^T z_j & 0 \\ z_j^T X_j & z_j^T z_j & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where z_j and X_j, given by (6.10) and (6.11), are the response vector and covariate matrix needed for the regression of Y_j on $Y_1,\ldots, Y_{j-1}$. If this matrix is swept on positions 0, 1,..., $j-1$, the result is

$$\begin{bmatrix} -\left(X_j^T X_j\right)^{-1} & \left(X_j^T X_j\right)^{-1} X_j^T z_j & 0 \\ z_j^T X_j \left(X_j^T X_j\right)^{-1} & z_j^T z_j - z_j^T X_j \left(X_j^T X_j\right)^{-1} X_j^T z_j & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

But notice that

$$\left(X_j^T X_j\right)^{-1} X_j^T z_j = \hat{\beta}_j$$

is the vector of estimated coefficients from ordinary least-squares regression of z_j on X_j. Moreover, it is straightforward to show that

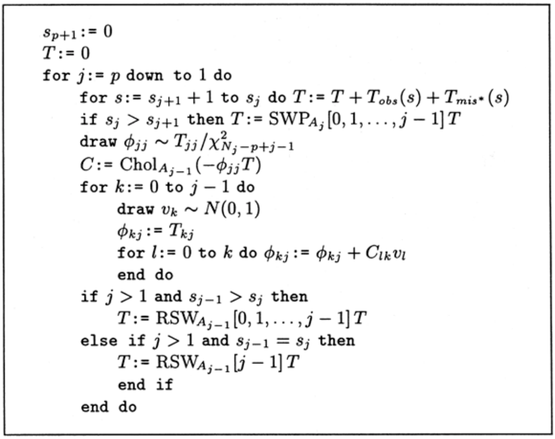
$$z_j^T z_j - z_j^T X_j \left(X_j^T X_j\right)^{-1} X_j^T z_j = \hat{\in}_j^T \hat{\in}_j,$$

where $\hat{\in}_j = z_j - X_j \hat{\beta}_j$ is the vector of estimated residuals. The quantities needed to describe the posterior distribution of ϕ_j, given the observed data Y_{obs} and imputed data in Y_{mis*} can thus be obtained by sweeping the matrix T_j. Note that all of the elements of T_j in rows and columns $j + 1,\ldots, p$ are zero before and after sweeping. Superfluous arithmetic can be avoided by applying the generalized sweep operator (Section 6.5.2) to sweep only the nonzero portions of T. The regression computations become

$$SWP_{A_j}[0,\ldots,j-1]T_j = \begin{bmatrix} -\left(X_j^T X_j\right)^{-1} & \hat{\beta}_j & 0 \\ \hat{\beta}_j^T & \hat{\in}_j^T \hat{\in}_j & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $A_j = \{0, 1, ..., j\}$.

An implementation of the P-step is shown in Figure 6.11. The components of ϕ are simulated in the reverse order $\phi_1, \phi_{p-1}, ..., \phi_1$ and placed in a $(p + 1) \times (p + 1)$ matrix as shown in Section 6.5.2. This implementation requires two matrix workspaces of the same size as ϕ : T, in which the sufficient statistics T_j are accumulated and swept; and C, which holds the Cholesky factors required for simulating the vectors of regression coefficients β_j. In addition, a vector workspace $v = (v_0, v_1, ..., v_{p-1})$ is needed for temporary storage of normal random variates. The quantity N_j, which appears in the degrees of freedom of the chisquare random variate, is



```
$s_{p+1} := 0$
$T := 0$
for $j := p$ down to 1 do
    for $s := s_{j+1} + 1$ to $s_j$ do $T := T + T_{obs}(s) + T_{mis^*}(s)$
    if $s_j > s_{j+1}$ then $T := \text{SWP}_{A_j}[0, 1, \ldots, j-1]\,T$
    draw $\phi_{jj} \sim T_{jj} / \chi^2_{N_j - p + j - 1}$
    $C := \text{Chol}_{A_{j-1}}(-\phi_{jj} T)$
    for $k := 0$ to $j - 1$ do
        draw $v_k \sim N(0, 1)$
        $\phi_{kj} := T_{kj}$
        for $l := 0$ to $k$ do $\phi_{kj} := \phi_{kj} + C_{lk} v_l$
        end do
    if $j > 1$ and $s_{j-1} > s_j$ then
        $T := \text{RSW}_{A_{j-1}}[0, 1, \ldots, j-1]\,T$
    else if $j > 1$ and $s_{j-1} = s_j$ then
        $T := \text{RSW}_{A_{j-1}}[j-1]\,T$
        end if
    end do
```

Figure 6.11. *P-step for monotone data augmentation.*

$$N_j = \sum_{s=1}^{s_j} n_s,$$

the total number of rows of the data matrix Y for which variable Y_j is either observed or imputed.

The algorithm of Figure 6.11 operates as follows. After the elements of T are initialized to zero, the sufficient statistics for

$Y_1,\dots, Y_p$ are accumulated in T over missingness patterns $1,\dots, s_p$. The matrix T is swept on positions $0,\dots, p-1$, producing statistics from the regression of Y_p on $Y_1,\dots, Y_{p-1}$. A random value of $\phi_p = (\gamma_p,\beta_p)$ is then drawn from its posterior, distribution. If additional rows of Y will enter into the next regression, i.e. if $s_{p-1} > s_p$ then T is reverse-swept on positions $0,\dots, p-1$ to prepare for the accumulation of sufficient statistics over these additional rows. Otherwise, T is reverse-swept only on position $p-1$, yielding the results from the regression of Y_{p-1} on $\mathrm{Y}_1,\dots, Y_{p-2}$. Continuing in this fashion, the algorithm draws $\phi_{p-1},\phi_{p-2},\dots,\phi_1$. Upon completion, the resulting value of ϕ should be transformed to the θ-scale (Section 6.5.2) to prepare for the next I-step.

The accumulation of sufficient statistics in line 4 of this algorithm may be rewritten as

$$T := T + \sum_{s=s_{j+1}+1}^{s_j} T_{obs}(s) + \sum_{s=s_{j+1}+1}^{s_j} T_{mis*}(s). \tag{6.25}$$

The first sum on the right-hand side of (6.25) depends only on the observed data Y_{obs} and does not need to be recalculated at each P-step. Calculating

$$B_j = \sum_{s=s_{j+1}+1}^{s_j} T_{obs}(s)$$

once at the outset of the program and storing it for future iterations can substantially reduce the amount of computation required at each P-step. Notice that we do not need to calculate and store B_j for every $j = 1, 2,\dots, p$, but only for those values of j for which $s_{j+1} < s_j$. The second sum on the right-hand side of (6.25) depends on Y_{mis*}, the missing values imputed at the I-step, so these terms will need to be recalculated at each P-step.

6.5.6 Uses and extensions

Like ordinary data augmentation, monotone data augmentation enables us to (a) simulate values of θ from the observed-data posterior $P(\theta|Y_{obs})$, and (b) create proper multiple imputations of Y_{mis}. The output stream is a sequence

$$\left(Y_{mis*}^{(1)},\theta^{(1)}\right),\left(Y_{mis*}^{(2)},\theta^{(2)}\right),\ldots,\left(Y_{mis*}^{(t)},\theta^{(t)}\right),\ldots$$

with $P(Y_{mis*,}\ \theta|Y_{obs})$ as its stationary distribution. After a sufficient burn-in period, successive values of θ,

$$\theta^{(t)},\theta^{(t+1)},\theta^{(t+2)},\ldots,$$

constitute a dependent sample from $P(\theta|Y_{obs})$ and may be summarized using any of the methods described in Chapter 4. Iterates of Y_{mis*} that are sufficiently far apart in the output stream, say

$$Y_{mis*}^{(t)},Y_{mis*}^{(t+k)},Y_{mis*}^{(t+2k)},\ldots$$

for some large value of k, can be taken as proper multiple imputations of Y_{mis*}.

In many applications, we will want proper multiple imputations of all the missing data in Y_{mis}, not just the missing data Y_{mis*}. needed to complete a monotone pattern. To obtain m proper multiple imputations of Y_{mis}, we should first generate m values of θ that are approximately independent, say

$$\theta(t),\theta(t+k),\ldots,\theta(t+mk)$$

and then draw a value of Y_{mis} given each one,

$$\begin{aligned} Y_{mis}^{(1)} &\sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right), \\ Y_{mis}^{(2)} &\sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t+k)}\right), \\ &\vdots \\ Y_{mis}^{(m)} &\sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t+mk)}\right), \end{aligned}$$

using the I-step for ordinary data augmentation described in Chapter 5. Of course, to obtain independent values of θ we do not necessarily need to subsample every kth value from a single chain of monotone data augmentation; we can also run m independent chains of length k from a common starting value, or better still, from m independent starting values drawn from an overdispersed starting distribution (Section 4.4.2).

Alternative priors

The monotone data augmentation algorithm described above uses the customary noninformative prior distribution

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}.$$

It is occasionally helpful to use other priors. For example, in sparse-data situations where some aspects of the covariance structure are poorly estimated, we may want to apply the ridge prior described in Section 5.2.3. A strategy for monotone data augmentation under an arbitrary inverted-Wishart prior distribution for Σ is outlined by Liu (1993). Liu's algorithm uses a clever factorization of the posterior distribution under monotone data, derived using an extension of the Bartlett decomposition (Section 5.4.2).

6.5.7 Example

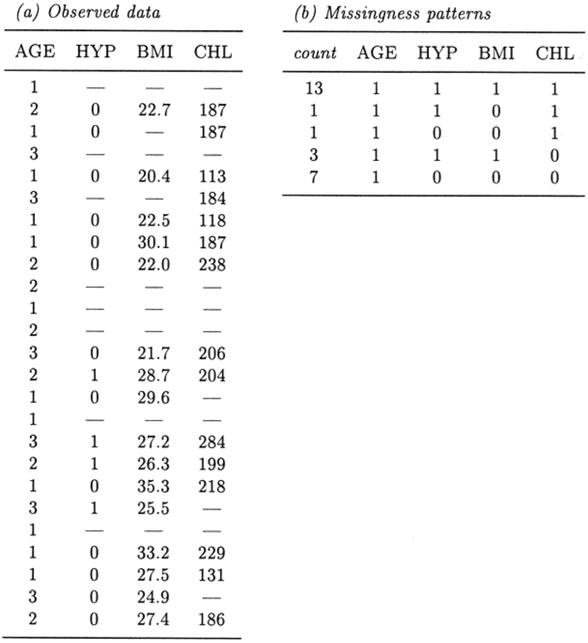
Section 6.4 presented a small simulation study designed to mimic the types of data and missingness found in a national health examination survey. The response mechanism shown in Table 6.12, which was estimated from an actual survey, tends to produce samples that are nearly monotone. The most common missingness pattern, which occurs for about 70% of sampled individuals, has all four survey variables (AGE, BMI, HYP, CHL) observed. The next most common pattern, which occurs about 15% of the time, has AGE observed and the other three variables missing. If AGE is placed in the first column of the data matrix, then at least 85% of the sampled individuals will tend to conform to a monotone pattern. This is precisely the type of situation in which monotone data augmentation should outperform ordinary data augmentation.

To illustrate, a simple random sample of n = 25 individuals was drawn from the study population, and a random pattern of missingness was imposed on the sample according to the estimated mechanism. The simulated data and missingness patterns are shown in Table 6.14. Overall there are 27 missing values, but only three of them (one value of HYP and two of BMI) are needed to complete a monotone pattern.

As in the simulation study, we replaced AGE by two dummy indicators (AGE$_2$ = 1 for AGE = 2 and 0 otherwise; AGE$_3$ = 1 for AGE = 3 and 0 otherwise) and modeled the resulting five-variable dataset as multivariate normal. An exploratory run of the EM algorithm revealed that the worst fraction of missing information, as estimated from the elementwise rates of convergence, is about 66%. Runs of data augmentation and monotone data augmentation under the customary noninformative prior verified that monotone data augmentation does indeed converge faster. Sample autocorrelations for two functions of the parameter θ, calculated over 5000 iterations of each algorithm, are displayed in Figure 6.12. Figure 6.12 (a) shows ACFs for the correlation between BMI and CHL, and Figure 6.12 (b) shows ACFs for the worst linear function of θ, which was estimated from the trajectory of EM. With respect to these two parameters, data augmentation appears to be approximately stationary by lag k = 4, whereas monotone data augmentation seems nearly stationary at lag k = 1.

For a dataset of this size, iterations of either algorithm can be executed so quickly on modern computers that the advantage of monotone data augmentation is of little practical importance.

Table 6.14. *Sample data from a health examination survey with simulated Pattern of missingness (1=observed, 0=missing)*



(a) Observed data

AGE	HYP	BMI	CHL
1	—	—	—
2	0	22.7	187
1	0	—	187
3	—	—	—
1	0	20.4	113
3	—	—	184
1	0	22.5	118
1	0	30.1	187
2	0	22.0	238
2	—	—	—
1	—	—	—
2	—	—	—
3	0	21.7	206
2	1	28.7	204
1	0	29.6	—
1	—	—	—
3	1	27.2	284
2	1	26.3	199
1	0	35.3	218
3	1	25.5	—
1	—	—	—
1	0	33.2	229
1	0	27.5	131
3	0	24.9	—
2	0	27.4	186

(b) Missingness patterns

count	AGE	HYP	BMI	CHL
13	1	1	1	1
1	1	1	0	1
1	1	0	0	1
3	1	1	1	0
7	1	0	0	0

In a large database, however, a four-fold reduction in the time required to produce a given number of multiple imputations can be a substantial improvement. Moreover, the gains tend to become even more dramatic as the rates of missing information increase. In studies that employ double sampling or matrix sampling, it is not uncommon for the rates of missing information for some parameters to be 90% or more. These high rates of missingness, due primarily to data that are missing by design, can make the convergence of ordinary data augmentation painfully slow. It is easy to envision scenarios where exploiting a near-monotone pattern that

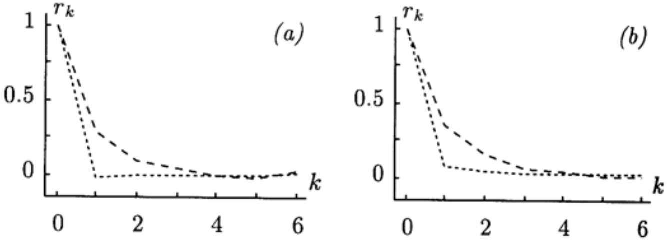


Figure 6.12. *Sample ACFs of series from ordinary data augmentation (dashed line) and monotone data augmentation (dotted line) for (a) the correlation between BMI and CHL, and (b) the worst linear function of the parameter.*

arises by design can reduce the computations by one or more orders of magnitude.

CHAPTER 7

Methods For Categorical Data

7.1 Introduction

The past three decades have seen enormous growth in the theory and application of models for categorical data. Categorical-data techniques such as logistic regression and loglinear modeling are now commonplace in the social and biomedical sciences and nearly every other major area of statistical application. For the most part, however, principled methods for handling missing values in categorical data analysis have not been readily available.

We have already demonstrated that, under certain circumstances, categorical variables can be handled quite reasonably by applying the multivariate normal distribution (Sections 6.3 and 6.4). In other situations, however, it is desirable to use a model specifically designed for categorical data. This chapter develops techniques for parameter simulation and multiple imputation for incomplete categorical data under the saturated multinomial model. The saturated multinomial is more general than the multivariate normal in the sense that it allows for three-way and higher associations among the variables; the multivariate normal captures simple (two-way) associations only. When maintaining higher-order associations among continuous variables is a priority, it may even be worthwhile to categorize them and apply the methods of this chapter rather than normal-based methods, even though the categorization may result in a slight loss of information.

The generality of the saturated multinomial model can also be a drawback, however, because in many applications (particularly as the number of variables grows) some of the higher-order associations may be poorly estimated. In these situations, it often helps to simplify the model by selectively removing some of these complex associations. Elimination of higher-order associations will be discussed within the framework of loglinear modeling, which will be covered in Chapter 8.

Section 7.2 lays the groundwork for our categorical-data methods by reviewing fundamental properties of two multivariate distributions, the multinomial and the Dirichlet. Basic EM and data augmentation algorithms for the saturated multinomial model are developed in Section 7.3. Section 7.4 introduces a class of algorithms that tends to be more efficient when the missing values fall in a pattern that is nearly monotone.

7.2 The multinomial model and Dirichlet prior

7.2.1 The multinomial distribution

Let Y_1, Y_2,..., Y_p denote a set of categorical variables. For notational convenience, we will suppose that the levels of each variable are coded as positive integers, so that

$$Y_j \text{ takes possible values } 1, 2, \ldots, d_j$$

for j = 1, 2,..., p. Throughout this chapter, we will regard the levels 1, 2,..., d_j as nominal or unordered categories; we do not consider models that explicitly account for ordering, e.g. the models for ordinal variables discussed by Agresti (1984) and Clogg and Shihadeh (1994). Incomplete ordinal data can sometimes be handled, at least approximately, by pretending that they are normally distributed and applying the methods of Chapters 5-6. Alternatively, one can disregard the order of the levels and apply the methods described here. Disregarding the order results in some loss of information and may lead to models that are more complex (i.e. having more parameters) than necessary to describe the essential relationships among the variables. For developing models that are parsimonious and scientifically meaningful, it is usually desirable to retain the ordering of the levels, if possible. On the other hand, if the immediate goal is to create plausible multiple imputations of missing data for future analyses, then disregarding the order and applying the methods of this chapter may be a perfectly reasonable approach.

If values of Y_1, Y_2,..., Y_p are recorded for a sample of n units, then the complete data can be expressed as an $n \times p$ data matrix Y. If the sample units are independent and identically distributed (iid), then without loss of information we can reduce Y to a contingency table with D cells, where $D = \prod_{j=1}^{p} d_j$ is the number of distinct combinations of the levels of Y_1, Y_2,..., Y_p. In practice, logical constraints among the variables may render some of these combinations impossible. For example, if Y_1 represents age (1=0-9 years, 2=10-19 years,...) and Y_2 represents marital status (1=never married, 2=currently married,...) then under most circumstances ($Y_i = 1$, $Y_2 = 2$) should be regarded as an impossible event. Cells of the contingency table that are necessarily empty due to logical constraints are called *structural zeroes* (e.g. Agresti, 1990). Structural zeroes present only minor complications, most of which are notational. For now, we will proceed as if there are no structural zeroes.

Let us index the cells of the contingency table by the single subscript d = 1, 2,..., D. Let x_d be the number of sample units that fall into cell d, and let

$$x = \left(x_1, x_2, ..., x_D\right)$$

denote the entire set of cell frequencies or counts. If the sample units are iid and the sample size $n = \sum_{d=1}^{D} x_d$ is regarded as fixed, then x has a multinomial distribution. We will write

$$x \mid \theta \sim M(n, \theta)$$

to indicate that x is multinomial with index n and parameter

$$\theta = \left(\theta_1, \theta_2, ..., \theta_D\right),$$

where θ_d is the probability that a unit falls into cell d. The probability distribution for x is given by

$$P(x \mid \theta) = \frac{n!}{x_1!\, x_2! \cdots x_D!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_D^{x_D} \tag{7.1}$$

for $\sum_{d=1}^{D} x_d = n$ and 0 otherwise. Because the total sample size n is fixed, one of the elements of x is redundant; we can replace X_D by $\sum_{d=1}^{D-1} x_d$ and regard (7.1) as the probability distribution for $(x_1,\ldots, x_{D-1})$

Notice that the cell probabilities must satisfy $\sum_{d=1}^{D} \theta_d = 1$, so the multinomial model has only $D - 1$ free parameters; θ_D can be replaced by $1 - \sum_{d=1}^{D-1} \theta_d$. Alternatively, we can regard the full vector θ = $(\theta_1, \ldots, \theta_D)$ as the unknown parameter, with the understanding that it must lie within the simplex

$$\Theta = \left\{ \theta : \theta_d \geq 0 \text{ for all } d \text{ and } \sum_{\mathrm{d}=1}^{\mathrm{D}} \theta_d = 1 \right\}, \tag{7.2}$$

a $(D - 1)$-dimensional subset of D-dimensional space. When $D = 3$, for example, Θ is the region encompassed by the triangle with vertices (1, 0, 0), (0, 1, 0) and (0, 0, 1).

The simplex Θ is the natural parameter space for the multinomial, i.e. the set of all possible values of θ for which (7.1) is a valid probability model. Throughout this chapter, we allow θ to lie anywhere in Θ. Such a model is said to be *saturated*, because it includes the maximum number of free parameters $(D - 1)$. The saturated model is very general; it allows for any kind of relationships to exist among the variables Y_1, $Y_2,\ldots$, Y_p. In many applications, however, such generality is undesirable because the information contained in the observed data may not be sufficient to estimate so many parameters adequately. Moreover, when the goal is to develop a model that is scientifically meaningful, models that are more parsimonious (i.e. having fewer parameters) than the saturated model may be easier to interpret. In Chapter 8, we will show how to reduce the number of free parameters by imposing loglinear constraints on the elements of θ.

When the multinomial vector x has only $D = 2$ cells, x_2 and θ_2 can be replaced by $n - x_1$, and $1 - \theta_1$, respectively, and (7.1) reduces to a binomial distribution,

$$P(x \mid n) = \frac{n!}{x_1!\left(n - x_1\right)!}\theta_1^{x_1}\left(1 - \theta_1\right)^{n - x_1}.$$

In this special case, we will sometimes use the notation

$$x_1 \mid \theta_1 \sim B\left(n, \theta_1\right)$$

as an alternative to

$$\left(x_1, n - x_1\right) \mid \theta \sim M\left(n, \left(\theta_1, n - \theta_1\right)\right).$$

The first two moments of the multinomial distribution are given by

$$\begin{aligned} E\left(x_d \mid \theta\right) &= n\theta_d, \\ V\left(x_d \mid \theta\right) &= n\theta_d\left(1 - \theta_d\right), \\ \mathrm{Cov}\left(x_d, x_{d'} \mid \theta\right) &= -n\theta_d\theta_{d'}, d' \neq d \end{aligned}$$

Further properties of the multinomial distribution can be found in texts on discrete data (e.g. Bishop, Fienberg and Holland, 1975; Agresti, 1990).

Maximum-likelihood estimation

The likelihood function for the multinomial parameter is

$$L(\theta \mid Y) \propto \prod_{d=1}^{D} \theta_d^{x_d} I_\Theta(\theta), \tag{7.3}$$

where $I_\Theta(\theta)$ is an indicator function equal to 1 if $\theta \in \Theta$ and 0 otherwise. Notice that we have written $L(\theta|Y)$ rather than $L(\theta|x)$. We are allowed to do this because all relevant information about θ in the data matrix Y is captured in the contingency table x; that is, we can reconstruct Y from x except for the order of the sample units, which under the iid assumption is statistically irrelevant. The loglikelihood is

$$l(\theta \mid Y) = \sum_{d=1}^{D} x_d \log \theta_d, \tag{7.4}$$

defined over the simplex Θ. The multinomial is a regular exponential family distribution whose sufficient statistics are simply the cell counts $x = (x_1,..., X_D)$. Therefore, complete-data ML estimates can be obtained simply by equating each observed cell count x_d to its expectation $E(x_d|\theta) = n\theta_d$, leading to the well-known result that the ML estimates for the cell probabilities are the observed proportions

$$\hat{\theta}_d = \frac{x_d}{n}, d = 1,..., D. \tag{7.5}$$

7.2.2 Collapsing and partitioning the multinomial

The multinomial distribution has two convenient properties that enable us to factor the probability distribution $P(x|\theta)$ and the likelihood $L(\theta|Y)$. Suppose that we collapse two cells of the contingency table, say x_1 and x_2, adding the frequencies together to produce a new table $x^* = (z, x_3, ..., x_D)$ where $z = x_1 + x_2$. Then (a) the distribution of x^* is multinomial,

$$x^* \mid \theta \sim M(n, \theta^*), \tag{7.6}$$

where $\theta^* = (\xi, \theta_3, \ldots, \theta_D)$ and $\xi = \theta_1 + \theta_2$; and (b) the conditional distribution of (x_1, x_2) given z is also multinomial,

$$(x_1, x_2) \mid z, \theta \sim M\left(z, (\theta_1 / \xi, \theta_2 / \xi)\right). \tag{7.7}$$

Property (a) is derived by summing the multinomial probabilities for all x-vectors consistent with $x_1 + x_2 = z$,

$$\begin{aligned}
P(x^* \mid \theta) &= \sum_{j=0}^{z} P(x_1 = j, x_2 = z - j, x_3, \ldots, x_D) \\
&= \sum_{j=0}^{z} \frac{n!}{j!(z-j)!\, x_3! \cdots x_D!} \theta_1^j \theta_2^{z-j} \theta_3^{x_3} \cdots \theta_D^{x_D} \\
&= \frac{n!}{z!\, x_3! \cdots x_D!} \theta_3^{x_3} \cdots \theta_D^{x_D} \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \theta_1^j \theta_2^{z-j},
\end{aligned}$$

and noting that

$$\sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \theta_1^j \theta_2^{z-j} = (\theta_1 + \theta_2)^z$$

by the Binomial Theorem. Property (b) can be deduced as follows.

Notice that if we repeatedly apply Property (a) to collapse the table down to $x_1 + x_2 = z$ and $x_3 + \cdots + x_D = n - z$, we obtain

$$(z, n - z) \mid \theta \sim M\left(n, (\xi, 1 - \xi)\right).$$

Moreover, if we collapse $x_3 + \cdots + x_D = n - z$ but leave x_1 and x_2 intact, then

$$(x_1, x_2, n - z) \mid \theta \sim M\left(n, (\theta_1, \theta_2, 1 - \xi)\right).$$

The conditional distribution of (x_1, x_2) given z is by definition

$$P(x_1, x_2 \mid z, \theta) = \frac{P(x_1, x_2, n - z \mid \theta)}{P(z, n - z \mid \theta)} \tag{7.8}$$

for $x_1 + x_2 = z$ and 0 otherwise. Substituting expressions for the numerator and denominator, the right-hand side of (7.8) becomes

$$\left[\frac{n!}{x_1!\,x_2!(n-z)!}\theta_1^{x_1}\theta_2^{x_2}(1-\xi)^{n-z}\right]\left[\frac{n!}{z!(n-z)!}\xi^z(1-\xi)^{n-z}\right]^{-1}$$

which reduces to

$$P(x_1, x_2 \mid z, \theta) = \frac{z!}{x_1!\,x_2!}\left(\frac{\theta_1}{\xi}\right)^{x_1}\left(\frac{\theta_2}{\xi}\right)^{x_2},$$

the desired result.

We have stated these results in terms of collapsing just two cells ($x_1 + x_2 = z$), but they extend to arbitrary types of collapsing. Suppose that we partition the cell numbers $\{1, 2, ..., D\}$ into subsets $A_1, A_2, ..., A_K$ that are mutually exclusive and collectively exhaustive. Denote the part of x corresponding to A_k by

$$x_{(k)} = \{x_d : d \in A_k\}.$$

The collection $\{x_{(1)}, x_{(2)}, ..., x_{(K)}\}$ of these parts will be called the *partitioned table*, and $x_{(k)}$ will be called the *kth part of* x. Denote the total frequency for the *k*th part by

$$z_k = \sum_{d \in A_k} x_d.$$

The collection $z = (z_1, z_2, ..., z_K)$ of these total frequencies will be called the *collapsed table*. Denote the probability that a sample unit falls into the *k*th part by

$$\xi_k = \sum_{d \in A_k} \theta_d, \tag{7.9}$$

and the conditional probability that a sample unit falls into cell *d* given that it falls into the *k*th part by

$$\phi_{kd} = \theta_d / \xi_k \text{ for all } d \in A_k. \tag{7.10}$$

Denote the collection of all such conditional probabilities for the kth part by

$$\phi_k = \{\phi_{kd} : d \in A_k\}.$$

Notice that ϕ_k is simply the kth part of θ, rescaled so that its elements sum to one. Under these conditions, it can be shown that (a) the marginal distribution of the collapsed table is multinomial,

$$z \mid \theta \sim M(n, \xi), \tag{7.11}$$

where $\xi = (\xi_1, \xi_2, \ldots, \xi_K)$; and (b) the conditional distribution of the partitioned table given the collapsed table is a set of independent multinomials,

$$\begin{aligned} x_{(1)} \mid z, \theta &\sim M(z_1, \phi_1), \\ x_{(2)} \mid z, \theta &\sim M(z_2, \phi_2), \\ &\vdots \\ x_{(K)} \mid z, \theta &\sim M(z_K, \phi_K). \end{aligned} \tag{7.12}$$

A set of independent multinomial distributions over a partitioned contingency table is often called a *product multinomial*. For any collapsing scheme, we can thus factor the multinomial distribution into a multinomial for the frequencies in the collapsed table, whose parameters are obtained by summing or collapsing θ in the same manner that x was collapsed, and a product multinomial for the conditional distribution of the partitioned table given the collapsed table, whose parameters are obtained by partitioning θ and rescaling each part to sum to one.

Factoring the likelihood

It is easy to see that the parameters for the collapsed table and the partitioned table, which we denote collectively by

$$\psi = \left(\xi, \phi_1, ..., \phi_K\right),$$

are a one-to-one function of θ = $(\theta_1,...,\theta_d)$; the forward transformation $\psi = \psi(\theta)$ is defined by (7.9)-(7.10), and the back transformation θ=$\psi^{-1}(\psi)$ is

$$\theta_d = \xi_k \phi_{kd} \text{ for all } d \in A_k, \tag{7.13}$$

$k = 1, 2,..., K$. Moreover, the parameters for the collapsed table and each part of the partitioned table are mutually distinct; any values of $\xi,\phi_1,...,\phi_K$ in their respective simplexes will produce a value of θ in its simplex Θ. It follows that the likelihood function for ψ can be factored into a sequence of independent multinomial likelihoods,

$$L(\psi \mid x) = L(\xi \mid z) L\left(\phi_1 \mid x_{(1)}\right) \cdots L\left(\phi_K \mid x_{(K)}\right).$$

Likelihood-based inferences about each part of ψ can be carried out independently, and the results can then be combined to produce a valid overall inference. For example, ML estimates for each part $\xi,\phi_1,...,\phi_K$ can be calculated independently; they are

$$\hat{\xi}_k = \frac{z_k}{n} \quad \text{and } \hat{\phi}_{kd} = \frac{x_d}{z_k} \text{ for all } d \in A_k. \quad \text{for all } d \in \mathrm{A}_k.$$

Applying the back transformation $\theta = \psi^{-1}(\psi)$ to these values gives $\hat{\theta}_d = x_d/n$, the ML estimates for θ. Bayesian inferences for each part can also proceed independently, provided that the prior distribution applied to ψ factors into independent priors for $\xi, \phi_1, \ldots, \phi_K$.

Non-multinomial sampling

This factorization of the multinomial likelihood has important implications for statistical inference. In many datasets, the distribution of one or more categorical variables is not random but fixed by design. Common examples of this include (a) treatment indicators in randomized experiments and (b) variables used to define strata in sample surveys. When the distribution of one or more variables is fixed by design, the cell frequencies $x = (x_1, x_2, \ldots, X_D)$ are not multinomial; rather, they follow a product-multinomial distribution. If we erroneously apply a multinomial model, however, we can still obtain valid likelihood-based or Bayesian inferences about the parameters of the nonfixed portion of the model. This result holds for incomplete data, provided that the missing values are confined to variables that are not fixed (Section 2.6.2). In addition, the multinomial likelihood may lead to valid conditional inferences in situations where the total sample size n is random (e.g. Poisson sampling) (Bishop, Fienberg and Holland, 1975; Agresti, 1990). Although we will speak almost exclusively of the multinomial model throughout this chapter and the next, the reader should be aware that the methods presented here can be reasonably applied in many non-multinomial situations.

7.2.3 The Dirichlet distribution

The simplest way to conduct Bayesian inference with a multinomial model is to choose a parametric family of prior distributions whose density has the same functional form as the likelihood (7.3). Suppose that $\theta = (\theta_1, \ldots, \theta_D)$ is a vector of random variables with the property that $\theta_d \geq 0$ for $d = 1, \ldots, D$ and $\sum_{d=1}^{D} \theta_d = 1$. Then θ is said to have a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_D)$ if its density is

$$P(\theta \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_D)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1} \qquad (7.14)$$

over the simplex Θ, where $\alpha_0 = \sum_{d=1}^{D} \alpha_d$ and $\mathbf{\Gamma}(\cdot)$ denotes the gamma function. As a shorthand for (7.14), we will write

$$\theta \mid \alpha \sim D(\alpha).$$

The right-hand side of (7.14) is a valid probability density provided that $\alpha_d >$ for 1, . . . , D.

When the Dirichlet is used as a prior distribution for the parameters of the multinomial, we will typically omit the normalizing constant and write the prior density as

$$\pi(\theta) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1}, \qquad (7.15)$$

where $\alpha_1,\dots,\alpha_D$ are understood to be user-specified hyperparameters. Although this appears to be a joint density for D random variables, we must remember that one of the elements of θ is redundant. In taking expectations, for example, we would replace θ_D by $1 - \sum_{d=1}^{D-1} \theta_d$ and integrate with respect to $\theta_1,\dots,\theta_{D-1}$ In the special case of $D = 2$, $\theta_2 = 1 - \theta_1$ and the Dirichlet reduces to a beta distribution for θ_1. In this special case we may write

$$\theta_1 \mid \alpha \sim Beta(\alpha_1, \alpha_2)$$

as an alternative to

$$(\theta_1, \theta_2) \mid \alpha \sim D(\alpha).$$

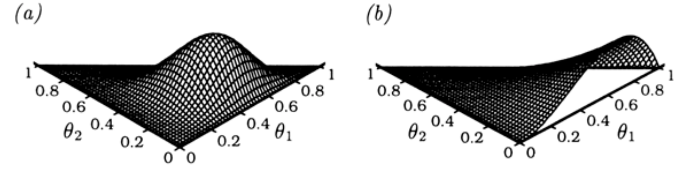


Figure 7.1. *Dirichlet densities for (a) α = (5,3,4) and (b) α = (3,1,2), plotted as functions of θ_1 and θ_2.*

Properties of the Dirichlet distribution

Here we state without proof some basic properties of the Dirichlet distribution. For a more detailed treatment, see Wilks (1962). The first two moments are given by

$$E(\theta_d) = \frac{\alpha_d}{\alpha_0},$$

$$V(\theta_d) = \frac{\alpha_d(\alpha_0 - \alpha_d)}{\alpha_0^2(\alpha_0 + 1)},$$

$$\mathrm{Cov}(\alpha_d, \alpha_{d'}) = -\frac{\alpha_d \alpha_{d'}}{\alpha_0^2(\alpha_0 + 1)}, d' \neq d.$$

If the means α_d/α_0 are held constant but α_0 is allowed to increase, then the variances and covariances are of order $O(\alpha_0^{-1})$. For this reason, α_0 may be regarded as a precision parameter; as it increases, the distribution becomes more tightly concentrated about the mean.

The mode of the Dirichlet can be found by noting that its density is equivalent to the likelihood function from a multinomial contingency table $x = (x_1,\ldots,X_D)$ with $x_d = \alpha_d - 1$, d = 1,…,D. This function is maximized at $\theta_d = x_d / \sum_{d'=1}^{D} x_{d'}$ provided that every x_d is nonnegative. Therefore, the mode of the Dirichlet density occurs at

$$\theta_d = \frac{\alpha_d - 1}{\alpha_0 - D} \quad d = 1,\ldots,D, \tag{7.16}$$

provided that every $\alpha_d \geq 1$.

Two examples of the Dirichlet density for $D = 3$ are shown in Figure 7.1. Because one of the elements of $\theta = (\theta_1,\theta_2,\theta_3)$ is redundant, the densities are plotted as functions of θ_1 and θ_2 over the triangular region $\theta_1 \geq 0$, $\theta_2 \geq 0$, $\theta_1 + \theta_2 \leq 1$. Figure 7.1 (a) shows the density for $\alpha = (5,3,4)$, and Figure 7.1 (b) shows the density for $\alpha = (3,1,2)$. Notice that in (a) the mode lies in the interior of the parameter space Θ, whereas in (b) the mode lies on the boundary. It is true in general that if every $\alpha_d > 1$, then the density has a unique mode in the interior of Θ. If every $\alpha_d = 1$, then the Dirichlet density is uniform over Θ. If one or more of the parameters ad is equal to one but none are less than one, then the density is bounded and the mode occurs on the boundary. Finally, if $\alpha_d < 1$ for any d then the density function becomes infinite on the boundary. These properties suggest that if $\theta \sim D(\alpha)$ represents the current state of knowledge about θ, and if one or more elements of a are less than or equal to one, then the mode may not be a sensible point estimate for θ; a better estimate would be the mean.

Relationship to the gamma distribution

An important relationship exists between the Dirichlet distribution and the gamma distribution. A random variable v is said to have a standard gamma distribution with parameter $a > 0$ if its density is

$$P(v \mid a) = \frac{1}{\Gamma(a)} v^{a-1} e^{-v}$$

for $v > 0$, and we write

$$v \mid a \sim G(a).$$

The gamma distribution is usually presented as a two-parameter family, with one parameter determining the shape and the other determining the scale. The standard gamma distribution is obtained by setting the usual scale parameter to one. The mean and variance of the standard gamma are both equal to a. The standard gamma also has the following reproductive property: if $v_1 \sim G(a_1)$ and $v_2 \sim G(a_2)$ are independent, then $v_1+v_2 \sim G(a_1+a_2)$.

The Dirichlet distribution can be obtained from the standard gamma as follows. Suppose that v_1, v_2,..., v_D are independent standard gamma variates with parameters $\alpha_1,\alpha_2,\ldots,\alpha_D$, respectively. If we take

$$\theta_d = \frac{v_d}{\sum_{d'=1}^{D} v_{d'}}, \quad d = 1,2,\ldots,D,$$

then $\theta = (\theta_1,\theta_2, \ldots,\theta_D)$ will have a Dirichlet distribution with parameter $\alpha = (\alpha_1,\alpha_2, \ldots,\alpha_D)$. This property enables us to simulate a Dirichlet random vector using a standard gamma variate generator. Methods for efficient generation of gamma random variates are reviewed by Kennedy and Gentle (1980).

Limitations of the Dirichlet prior

From a purely conceptual standpoint, the Dirichlet distribution is not the most attractive prior for cross-classified contingency tables. One of its drawbacks is that it treats the cells of the table in an unordered fashion, ignoring its cross-classified structure. We have adopted the Dirichlet prior mainly for computational convenience, because with complete data it leads to posterior distributions that are easily summarized. If the parameters of the data model are not well estimated by the data, and it becomes apparent that the choice of prior has a substantial impact on the results, then one should be wary of drawing firm conclusions from an analysis under a Dirichlet prior or, for that matter, any other type of prior.

7.2.4 Bayesian inference

It is easy to see what happens when a Dirichlet prior is applied to the parameters of the multinomial. Suppose that a contingency table $x = (x_1,\ldots,X_D)$ has a multinomial distribution with parameter $\theta = (\theta_1,\ldots,\theta_D)$, and the prior distribution for θ is Dirichlet with hyperparameter $\alpha = (\alpha_1, \ldots,\alpha_D)$,

$$x \mid \theta \sim M(n, \theta), \tag{7.17}$$

$$\theta \sim D(\alpha). \tag{7.18}$$

Multiplying the Dirichlet density (7.15) by the multinomial likelihood (7.3) produces

$$P(\theta \mid Y) \propto \theta_1^{\alpha_1 + x_1 - 1} \theta_2^{\alpha_2 + x_2 - 1} \cdots \theta_D^{\alpha_D + x_D - 1}, \tag{7.19}$$

which is a Dirichlet density with parameters

$$\begin{aligned} \alpha' &= \left(\alpha_1', \alpha_2', \ldots, \alpha_D'\right) \\ &= \left(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_D + x_D\right) \\ &= \alpha + x. \end{aligned} \tag{7.20}$$

The posterior distribution of θ under (7.17)-(7.18) is thus

$$\theta \mid Y \sim D(\alpha').$$

The posterior mean is

$$E(\theta \mid Y) = \left(\frac{\alpha_1'}{\alpha_0'}, \frac{\alpha_2'}{\alpha_0'}, \ldots, \frac{\alpha_D'}{\alpha_0'}\right)$$

where $\alpha_0' = \sum_{d=1}^{D} \left(\alpha_d + x_d\right) = \alpha_0 + n,$ and the posterior mode is

$$\text{mode}(\theta \mid Y) = \left(\frac{\alpha_1' - 1}{\alpha_0' - D}, \frac{\alpha_2' - 1}{\alpha_0' - D}, \ldots, \frac{\alpha_D' - 1}{\alpha_0' - D}\right)$$

provided that every $\alpha_d' \geq 1$.

The Dirichlet prior (7.17) is a proper probability distribution if $\alpha_1,\alpha_2,...,\alpha_D$ are all positive. Notice, however, that for (7.19) to be proper, we only need the updated hyperparameters $\alpha'_d = \alpha_d+x_d$ to be positive. This means that we can adopt an improper prior density function such as

$$\pi(\theta) \propto \theta_1^{-1}\theta_2^{-1}\cdots\theta_D^{-1}, \tag{7.21}$$

which is the limiting form of the $D(\alpha)$ density as a approaches (0, 0,..., 0), and still obtain a proper posterior if $\alpha_d+x_d > 0$ for every d. In a slight abuse of terminology, we will call (7-21) the Dirichlet density with α = (0, 0,..., 0); it should be understood that this is not a density per se, but it leads to a proper Dirichlet posterior if there are no empty cells (i.e. if every $x_d \geq 1$).

7.2.5 Choosing the prior hyperparameters

Because of the rule (7.20) for updating the hyperparameters α = ($\alpha_1,\ldots,\alpha_D$), it is tempting to think of these as imaginary prior counts in the cells of the contingency table. This notion is certainly correct in a relative sense; increasing α_d by one has the same inferential effect as observing one additional sample unit in cell d. In an absolute sense, however, we hesitate to interpret α_d as the number of prior observations in cell d, because it is not necessarily true that $\alpha_d = 0$ represents no prior observations in cell d.

Noninformative priors

When little prior information is available about 0, it may be sensible to take $\alpha_1,\alpha_2,\ldots,\alpha_D$ equal to a common value that is, to set $\alpha = (c,c,...,c)$ for some constant c. However, there is no unique choice for c that clearly represents a state of ignorance about θ. Most statisticians would agree that without strong prior information, the ML estimate

$$\hat{\theta} = \left(\frac{x_1}{n}, \frac{x_2}{n}, \ldots, \frac{x_D}{n} \right) \tag{7.22}$$

is a reasonable point estimate for θ. This is particularly true if $\hat{\theta}$ lies in the interior of the parameter space, i.e. if there are no empty cells. Notice that (7.22) is the posterior mean of θ under the improper prior with $c = 0$, assuming that there are no empty cells. But it is also the posterior mode under the uniform prior with $c = 1$. From the standpoint of estimating θ, one could thus argue that either (or neither!) of these priors is noninformative. Moreover, the Jeffreys invariance principle for choosing a noninformative prior (e.g. Box and Tiao, 1992) leads to the choice $c = 1/2$. Therefore, it seems reasonable to regard the whole range of values of c between zero and one as potentially noninformative.

With certain techniques or algorithms, there may be a natural choice for a noninformative prior. For example, with a mode-finding algorithm such as EM, the uniform prior ($c = 1$) will cause the procedure to converge to an ML estimate. In a general-purpose implementation of EM, therefore, it would be natural to adopt $c = 1$ as a default noninformative prior. In other situations, however, the choice is less clear. In data augmentation, for example, the parameters of interest are typically estimated by their simulated-posterior means. Under the prior with $c = 0$, the posterior mean coincides with the ML estimate (at least in the complete-data case) for parameters that are linear functions of $\theta_1,\ldots,\theta_D$, but not for nonlinear parameters (e.g. odds ratios). Unlike ML estimates, posterior means are not invariant under nonlinear transformations. Therefore, we cannot really claim that $c = 0$ is a good default prior for a general-purpose data augmentation routine. The $c = 0$ prior is also unattractive because it is improper; the existence of a proper posterior under this prior is not guaranteed.

If the sample size is large relative to the number of parameters being estimated, the choice of prior will tend to have little impact on the final inferences. For the examples in this book, we will adopt the Jeffreys prior ($c = 1/2$) as a default noninformative prior for simulations where the sample size is large. This choice is admittedly somewhat arbitrary. If there is any doubt that the influence of the prior is minimal, one should always conduct a sensitivity analysis, applying a variety of alternative priors to see how the resulting inferences change. If the results vary dramatically over a range of plausible priors, then the only scientifically justifiable conclusion may be that no firm conclusions are possible.

Sparse tables and flattening priors

When the sample size n is not much larger than the number of cells D, a substantial number of cells may contain no observations. A table $x = (x_1,...,x_D)$ in which a high proportion of the frequencies x_d are zero is said to be *sparse*. It is well known that when common models for discrete data (e.g. loglinear or logistic models) are fit to sparse tables, the empty cells can lead to inestimable parameters and/or ML estimates on the boundary. For this reason, it has often been suggested that a small positive number such as 1/2 should be added to every cell of a sparse table prior to model fitting. The use of such a number, called a *flattening constant*, is reviewed by Clogg et al. (1991).

The effect of a flattening constant is to smooth the estimate of θ toward a uniform table in which all cell probabilities are equal. When $x = (x_1,..., x_D)$ represents a cross-classification by discrete variables $Y_1,Y_2,...,Y_p$, a uniform table has no relationships whatsoever among the variables. Adding a constant $\epsilon > 0$ to every cell thus tends to be conservative, in the sense that it makes us less likely to conclude that relationships among the variables exist when in fact they do not.

A prior distribution that smooths parameter estimates toward a uniform table will be called a *flattening prior*. Flattening priors can be helpful for ensuring that the mode of θ is unique and lies in the interior of the parameter space. For mode-finding algorithms, the prior with $\alpha = (c,c,...,c)$ for some $c > 1$ is flattening; it adds the equivalent of $\epsilon = c-1$

prior observations to every cell. Values of c less than one are not recommended for mode-finding algorithms because they are 'anti-flattening,' pushing the estimate of θ away from a uniform table. For simulations in which the results are summarized by posterior means, any prior with $c > 0$ has a flattening effect on the elements of θ, adding the equivalent of $\in = c$ observations to every cell relative to the ML estimate. For odds ratios and other nonlinear parameters, however, the effect of these priors when c is near zero may hardly be flattening. For such parameters, priors with c close to zero may place too much mass near the boundary, causing inferences about nonlinear parameters to be unstable when the table is sparse. In sparse-data situations, it is always advisable to apply a variety of reasonable alternative priors and see how the results change.

When using a flattening prior, care should be taken not to oversmooth the data. Adding $\in$ imaginary counts to every cell introduces information equivalent to $D\in$ prior observations. In a very sparse table, adding, say, 1/2 to every cell may result in an effective prior sample size comparable to or greater than the actual sample size. In the absence of strong prior beliefs about θ, it is probably unwise to add prior information that amounts to more than about 10-20% of the actual sample size, so that the integrity of the observed data is not seriously compromised. If inferences about the parameters of interest cannot be stabilized by these modest amounts of prior information, then the model is probably too complex to be supported by the observed data. In such situations, it would be wise to simplify the model by eliminating unnecessary variables or by imposing loglinear constraints (Chapter 8).

Data-dependent priors

One obvious potential drawback of flattening priors is that when they are applied to cross-classified contingency tables, they smooth the data toward a model in which each variable Y_j has a uniform distribution over its levels $1,2,\ldots,d_j$. In many contexts, it is more desirable to smooth toward a model of mutual independence among the variables but to leave the marginal distributions of the variables unaffected. This can be achieved by making the prior data-dependent.

Suppose that one of the variables (say Y_1) represents the response of greatest interest, and the other variables are potential predictors of Y_1. Clogg et al. (1991) advocate a strategy in which prior observations are divided among cells of the contingency table in such a way that the marginal distribution of Y_1 in the observed data is preserved. For example, suppose that Y_1 is dichotomous, with $Y_1 = 1$ and $Y_1 = 2$ observed for 30% and 70% of the sample units, respectively. After an appropriate total number of prior observations n_0 has been chosen, 30% of this total can be allocated to cells of the table corresponding to $Y_1 = 1$, with the remaining 70% going to cells corresponding to $Y_1 = 2$. This strategy, which has an empirical Bayes flavor, smooths the estimates of θ toward a null model in which none of the predictors has any effect on Y_1, but it does not affect the overall distribution of Y_1 itself.

This strategy can be extended to formulate a prior that simultaneously preserves the marginal distributions of all the variables in the dataset (Fienberg and Holland, 1970, 1973). Suppose that cell d of a frequency table corresponds to the event $Y_1 = y_1$, $Y_2 = y_2, \ldots, Y_p = y_p$. If $Y_1, Y_2, \ldots, Y_p$ are mutually independent, then the probability associated with this cell is

$$\theta_d = P(Y_1 = y_1)P(Y_2 = y_2)\cdots P(Y_p = y_p). \tag{7.23}$$

The probabilities on the right-hand side of (7.23) can be estimated by the observed proportions in the sample. Substituting these estimates into (7.23), and multiplying the resulting estimate of θ_d by the desired total number of prior observations n_0, gives the number of prior observations to be allocated to cell d. For mode-finding algorithms, the hyperparameter associated with cell d would be

$$\alpha_d = 1 + n_0 \prod_{j=1}^{p} \hat{P}(Y_j = y_j) \tag{7.24}$$

where $\hat{P}(Y_j = y_j)$ is the observed proportion of sample units for which $Y_j = y_j$. For simulations,

$$\alpha_d = n_0 \prod_{j=1}^{p} \hat{P}(Y_j = y_j) \tag{7.25}$$

is a more natural choice, at least when we are concerned with linear functions of the elements of θ. These data-dependent priors can be thought of as discrete-data versions of the ridge prior for the parameters of the multivariate normal (Section 5.2.3), which also smooths toward a model of mutual independence among variables.

If the marginal distribution of each Y_j is not far from uniform (i.e. if the levels 1, 2,..., d_j occur with roughly the same frequency), then these data-dependent priors will have nearly the same effect as flattening priors. If some levels are relatively much rarer than others, however, then flattening priors may exert undue influence on these rarer categories, inflating their probabilities and distorting the inferences about certain functions of θ. When this is the case, data-dependent priors can be an attractive alternative to flattening priors, particularly when the data are sparse.

7.2.6 Collapsing and partitioning the Dirichlet

A Dirichlet random vector can be collapsed and partitioned in a manner analogous to that already described for the multinomial (Section 7.2.2), and the resulting vectors will have Dirichlet distributions. Let us first consider what happens when we collapse two elements. Suppose that $\theta = (\theta_1,\ldots,\theta_D)$ has a Dirichlet distribution with parameter $\alpha = (\alpha_1,\ldots,\alpha_D)$. If we form a new vector $\theta^* = (\xi,\theta_3,\ldots,\theta_D)$, where $\xi = \theta_1+\theta_2$, then (a) the distribution of θ^* is Dirichlet with parameter $\alpha^* = (\beta,\alpha_3, \ldots,\alpha_D)$, where $\beta = \alpha_1+\alpha_2$; and (b) the conditional distribution of $(\theta_1/\xi,\theta_2/\xi)$ given ξ is Dirichlet with parameter (α_1,α_2). Proofs of these properties are given by Wilks (1962);

they can also be justified by appealing to the relationship between the Dirichlet and the standard gamma distribution (Section 7.2.3).

More generally, suppose that $\theta = (\theta_1,\ldots,\theta_D)$ represents the cell probabilities for a multinomial vector $x = (x_1,\ldots,x_D)$, and we apply the transformation described in Section 7.2.2 to θ, transforming it into the cell probabilities for the collapsed and partitioned versions of x. That is, suppose that $A_1, A_2,\ldots, A_K$ are mutually exclusive and collectively exhaustive subsets of $\{1, 2,\ldots, D\}$ let

$$x_{(k)} = \left\{x_d; d \in A_k\right\}$$

be the kth part of x; and let

$$z_k = \sum_{d \in A_k} x_d$$

be the total frequency for the kth part. The cell probabilities for the collapsed table $z = (z_1, z_2,\ldots, z_k)$ are $\xi = (\xi_1,\xi_2,\ldots,\xi_K)$, where

$$\xi_k = \sum_{d \in A_k} \theta_d,$$

and the conditional probability of falling into cell d given that we are already in the kth part of the table is

$$\phi_{kd} = \theta_d / \xi_k \text{ for all } d \in A_k.$$

If θ has a Dirichlet distribution with parameter $\alpha = (\alpha_1,\ldots,\alpha_D)$, then it can be shown that the distribution of ξ is Dirichlet,

$$\xi \mid \alpha \sim D(\beta),$$

where the parameters $\beta = (\beta_1,\ldots,\beta_K)$ are obtained by summing the elements of α in the same way the elements of θ were summed to obtain ξ,

$$\beta_k = \sum_{d\in A_k} \alpha_d,$$

Moreover, if $\phi_k = \{\phi_{kd} : d\in A_k\}$ is the set of conditional probabilities for the kth part of x, then the conditional distribution of $\phi = (\phi_1, \phi_2, \ldots, \phi_K)$ given ξ is a set of K independent Dirichlet distributions,

$$\begin{aligned} \phi_1 \mid \xi, \alpha &\sim D\left(\alpha_{(1)}\right), \\ \phi_2 \mid \xi, \alpha &\sim D\left(\alpha_{(2)}\right), \\ &\vdots \\ \phi_K \mid \xi, \alpha &\sim D\left(\alpha_{(K)}\right), \end{aligned} \tag{7.26}$$

where $\alpha(k) = \{\alpha_d : d\in A_k\}$ denotes the kth part of α.

These properties imply that if a Dirichlet prior is applied to the parameter θ of a multinomial contingency table x, then the prior distribution of $\psi = (\xi, \phi)$ which is a one-to-one function of θ can be factored into independent Dirichlet distributions for $\xi, \phi_1, \ldots, \phi_K$. This ability of the Dirichlet distribution to be collapsed and partitioned makes it a very attractive prior for use in simulation algorithms, and provides the basis for a monotone data augmentation routine to be described in Section 7.4.

7.3 Basic algorithms for the saturated model

7.3.1 Characterizing an incomplete categorical dataset

This section presents EM and data augmentation algorithms for incomplete categorical datasets under the saturated multinomial model, which imposes no restrictions on the types of relationships that may exist among the variables Y_1, $Y_2,...,Y_p$. These algorithms are conceptually simple, but the notation needed to describe them in a general setting is somewhat unwieldy. To characterize the information contained in an incomplete multivariate categorical dataset, we must extend our notation for contingency tables in several ways.

First, we must account for the fact that the complete-data contingency table $x = (x_1, x_2,...,X_D)$ is actually a cross-classification by the levels of Y_1, Y_2,..., Y_p, and as such can be regarded as a p-dimensional array. Suppose that variable Y_j takes possible values 1, 2,...,d_j. Let x_y, where $y = (y_1, y_2,..., y_p)$, be the total number of units in the sample for which the event $Y_1 = y_1$, $Y_2 = y_2$,..., $Y_p = y_p$ occurs, and let θ_y be the probability of this event for any unit. Here we are using y to represent a generic realization of $(Y_1, Y_2,..., Y_p)$ for a single unit, i.e. a possible row of the $n \times p$ data matrix Y. We will denote the set of all possible values of y by Y. Assuming for the moment that there are no structural zeroes, Y is the Cartesian cross-product of the sets for $\{1, 2,...,d_j\}$ for $j = 1, 2,..., p$. When a cell count or probability appears with the vector subscript $y = (y_1, y_2,..., y_p)$ it should be interpreted as an element of an array with dimensions $d_1 \times d_2 \times \cdots \times d_p$, but when it appears with the scalar subscript d it should be interpreted as the dth element of a vector of length $D = \prod_{j=1}^{p} d_j$. Depending on the context, we will sometimes think of the tables x and θ as vectors,

$$x = (x_1, x_2, ..., x_D), \qquad \theta = (\theta_1, \theta_2, ..., \theta_D),$$

and at other times as p-dimensional arrays,

$$x = \left\{ x_y : y \in Y \right\}, \quad \theta \left\{ \theta_y : y \in Y \right\}.$$

The distinction between the two forms is simply a matter of notational convenience, because it is always possible to turn an array into a vector by assigning a linear ordering to its cells.

Now we must extend the notation to allow for missing data. Let us assume that observations have been grouped according to their missingness patterns. Index the missingness patterns that appear in the dataset by $s = 1, 2, \ldots, S$ and define a set of binary response indicators

$$r_{sj} = \begin{cases} 1 \text{ if } Y_j \text{ is observed in pattern } s, \\ 0 \text{ if } Y_j \text{ is missing in pattern } s. \end{cases}$$

Let $x_y^{(s)}$ denote the number of sample units within missingness pattern s for which $(Y_1, Y_2, \ldots, Y_p) = y$, and let

$$x^{(s)} = \left\{ x_y^{(s)} : y \in Y \right\}$$

denote the full set of these counts for pattern s. If any variables are missing in pattern s, then $x^{(s)}$ is not observed; rather, we observe the counts for a lower dimensional table in which the sample units have been cross-classified only by the observed variables. Let O_s and M_s be functions that extract from $y = (y_1, y_2, \ldots, y_p)$ the elements corresponding to the variables that are observed and missing, respectively, in pattern s,

$$O_s(y) = \left\{ y_j : r_{sj} = 1 \right\},$$
$$M_s(y) = \left\{ y_j : r_{sj} = 0 \right\}.$$

Also, let O_s and M_s be, respectively, the sets of all possible values of $O_s(y)$ and $M_s(y)$. For example, suppose that in a dataset with $p = 4$ variables, missingness pattern s has Y_1 and

Y_4 observed but Y_2 and Y_3 missing; then $O_s(y) = (y_1,y_4)$, $M_s(y) = (y_2,y_3)$,

$$O_s = \{(y_1, y_4) : y_1 = 1, 2, ..., d_1;\ y_4 = 1, 2, ..., d_4\},$$
$$M_s = \{(y_2, y_3) : y_2 = 1, 2, ..., d_2;\ y_3 = 1, 2, ..., d_3\}.$$

When the units within missingness pattern s are cross-classified only by their observed variables, the result is a table with counts that we shall denote by

$$z_{O_s(y)}^{(s)} = \sum_{M_s(y)\in M_s} x_y^{(s)} \text{ for all } O_s(y) \in O_s. \tag{7.27}$$

The marginal probability that an observation falls within cell $O_s(y)$ of this table will be called

$$\beta_{O_s(y)} = \sum_{M_s(y)\in M_s} \theta_y. \tag{7.28}$$

Observed-data likelihood

When $x = (x_1, x_2,..., X_D)$ has a multinomial distribution with parameter $\theta = (\theta_1, \theta_2,...,\theta_D)$, then the complete-data loglikelihood function for θ is

$$l(\theta \mid Y) = \sum_{d=1}^{D} x_d \log \theta_d$$

over the simplex Θ (Section 7.2.1). Equivalently, viewing x and θ as p-dimensional arrays, we can write the loglikelihood as

$$l(\theta \mid Y) = \sum_{y\in Y} x_y \log \theta_y. \tag{7.29}$$

When some data are missing, the observed-data loglikelihood can be calculated as follows. For any missingness pattern s, the observed data are summarized by the table

$$z^{(s)} = \left\{ z^{(s)}_{O_s(y)} : O_s(y) \in \mathcal{O}_s \right\}. \tag{7.30}$$

Notice that $z^{(s)}$ is a collapsed version of the unobserved $x^{(s)}$. By our rules for collapsing multinomial tables (Section 7.2.2), it follows that the contribution of $z^{(s)}$ to the observed-data loglikelihood is equivalent to that of a multinomial distribution with index

$$n_s = \sum_{y \in \mathcal{Y}} x_y^{(s)}$$

and parameter

$$\beta^{(s)} = \left\{ \beta_{O_s(y)} : O_s(y) \in \mathcal{O}_s \right\}. \tag{7.31}$$

That is, the contribution of $z^{(s)}$ to the observed-data loglikelihood is

$$\sum_{O_s(y) \in \mathcal{O}_s} z^{(s)}_{O_s(y)} \log \beta_{O_s(y)}.$$

The observed-data loglikelihood is the sum of these contributions for missingness patterns $s = 1, 2, \ldots, S$,

$$l(\theta \mid Y_{obs}) = \sum_{s=1}^{S} \sum_{O_s(y) \in \mathcal{O}_s} z^{(s)}_{O_s(y)} \log \beta_{O_s(y)}. \tag{7.32}$$

Despite the concise appearance of (7.32), it is a rather complicated function of the individual elements of θ. Evaluating $l(\theta|Y_{obs})$ at specific numerical values of θ is not difficult, but calculating analytic expressions for its first two derivatives can be tedious. For this reason, it is inconvenient to maximize $l(\theta|Y_{obs})$ by gradient methods. The EM algorithm is straightforward, however, because it involves only the repeated maximization of the complete-data loglikelihood (7.29).

7.3.2 The EM algorithm

EM for the saturated multinomial model was first described by Chen and Fienberg (1974) in the special case of $p = 2$ variables, and extended by Fuchs (1982) to arbitrary p. A description also appears in Chapter 9 of Little and Rubin (1987). The algorithm, which was already presented in Section 3.2.2 for two binary variables, is simple and intuitive. For each missingness pattern $s = 1,\ldots, S$, we allocate the counts in the observed table $z^{(s)}$ to the cells of the full p-way table $x^{(s)}$. This allocation is carried out in the proportions implied by the current estimate of θ. When the allocation is complete, the proportions in the resulting table x = $x^{(1)}+x^{(2)}+\cdots x^{(S)}$ provide the updated estimate of θ.

Before running EM, the observed data for each missingness pattern should first be cross-classified according to the observed variables; that is, the data should be reduced to $z^{(1)},\ldots, z^{(S)}$. Notice that $z^{(1)},\ldots, z^{(S)}$ can be regarded as arrays of varying dimensions; the number of dimensions for $z^{(s)}$ is equal to the number of variables observed in pattern s. When implementing EM on a computer, however, storing $z^{(1)},\ldots, z^{(S)}$ as multidimensional arrays tends to be cumbersome and inefficient. As the number of variables p grows, the number of arrays S can increase very rapidly. Moreover, these arrays can be very sparse; many of them may contain only a few or perhaps even just one observation each. A general-purpose computer program should be efficient in its use of memory, and the data structures it creates should have predictable size and shape. A more efficient way to store and manipulate the counts in $z^{(1)},\ldots, z^{(S)}$ is outlined in Appendix B.

The E- and M-steps

The complete-data loglikelihood (7.29) is a linear function of the elements of x = $\{x_y : y \in Y\}$, the unobserved p-dimensional table that cross-classifies all sample units by their values of Y_1, Y_2,..., Y_p. To perform the E-step, we must find the expectation of each count x_y given the observed data and an assumed value for θ. Notice that x can be expressed as $x = \sum_{s=1}^{S} x^{(s)}$, the sum of individual tables for missingness patterns 1,..., S. Moreover, the observed data $z^{(s)}$ for pattern s is a collapsed version of $x^{(s)}$, and by our rules for collapsing and partitioning (Section 7.2.2) it follows that the conditional distribution of $x^{(s)}$ given $z^{(s)}$ is product-multinomial. Let

$$x_{O_s(y)}^{(s)} = \left\{ x_y^{(s)} : M_s(y) \in \mathcal{M}_s \right\} \qquad (7.33)$$

denote the portion of $x^{(s)}$ that is obtained by fixing $O_s(y)$ at a specific value but varying $M_s(y)$ over $\mathcal{M}_s$; that is, $x_{O_s(y)}^{(s)}$ is simply the set of all cell counts in $x^{(s)}$ that contribute to the observed count $z_{O_s(y)}^{(s)}$. By the partitioning rules, $x_{O_s(y)}^{(s)}$ has, given $z_{O_s(y)}^{(s)}$, a multinomial distribution with index $z_{O_s(y)}^{(s)}$ and parameters

$$\gamma O_s(y) = \left\{ \theta_y / \beta_{O_s(y)} : M_s(y) \in \mathcal{M}_s \right\}; \qquad (7.34)$$

that is,

$$x_{O_s(y)}^{(s)} \Big| z_{O_s(y)}^{(s)}, \theta \sim M\left(z_{O_s(y)}^{(s)}, \gamma_{O_s(y)} \right). \qquad (7.35)$$

Notice that (7.34) is simply the portion of θ corresponding to $x_{O_s(y)}^{(s)}$, rescaled so that its elements sum to one. It follows that the conditional expectation of an element of $x^{(s)}$ is

$$E\left(x_y^{(s)} \Big| z^{(s)}, \theta \right) = z_{O_s(y)}^{(s)} \theta_y / \beta_{O_s(y)}. \qquad (7.36)$$

The E-step consists of calculating (7.36) for every s = 1,...,S and summing the results,

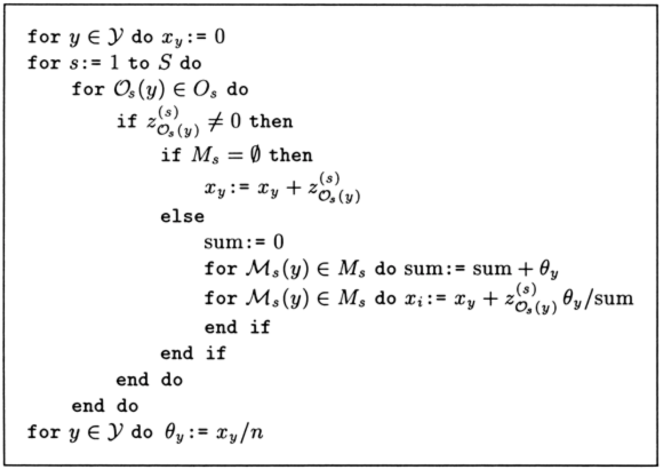
$$E\left(x_y \mid Y_{obs},\theta\right) = \sum_{s=1}^{S} z^{(s)}_{O_s(y)}\theta_y / \beta_{O_s(y)}. \tag{7.37}$$

Once the E-step has been completed, the M-step is trivial. The complete-data loglikelihood (7.29) is maximized at $\theta_y = x_y/n$, so the M-step is simply to

$$\text{estimate } \theta_y \text{ by } E\left(x_y \mid Y_{obs},\theta\right)/n \tag{7.38}$$

for all $y \in Y$

A pseudocode implementation of the E- and M-steps is shown in Figure 7.2. Given the observed counts $z^{(1)},\ldots, z^{(S)}$ and the current value of θ, this code overwrites θ with its updated value. A temporary workspace x of the same size as θ is required for accumulating the expected sufficient statistics. The algorithm cycles through the missingness patterns and checks to see whether the current pattern s has any missing variables (i.e. if $M_s(y)$ is nonempty). If not, then the observed counts for pattern s are added into the



```
for $y \in \mathcal{Y}$ do $x_y$ := 0
for $s$ := 1 to $S$ do
    for $\mathcal{O}_s(y) \in O_s$ do
        if $z^{(s)}_{\mathcal{O}_s(y)} \neq 0$ then
            if $M_s = \emptyset$ then
                $x_y$ := $x_y + z^{(s)}_{\mathcal{O}_s(y)}$
            else
                sum := 0
                for $\mathcal{M}_s(y) \in M_s$ do sum := sum + $\theta_y$
                for $\mathcal{M}_s(y) \in M_s$ do $x_i$ := $x_y + z^{(s)}_{\mathcal{O}_s(y)}\,\theta_y$/sum
                end if
            end if
        end do
    end do
for $y \in \mathcal{Y}$ do $\theta_y$ := $x_y/n$
```

Figure 7.2. *Single iteration of EM for the saturated multinomial model.*

elements of x; otherwise, the expectations (7.36) are calculated and added into x. After this is done for $s = 1, 2, \ldots, S$, the resulting elements of x are divided by n, which yields the updated value of θ.

Starting values and posterior modes

If the starting value of θ lies on the boundary of the parameter space Θ, i.e. if some of its elements are zero, then an inconsistency could arise in the initial E-step. It could happen that a nonzero count appears in one of the cells of the observed-data tables $z^{(1)}, \ldots, z^{(S)}$ for which the probability implied by the starting value of θ is zero. If this occurs, then the algorithm may halt due to attempted division by zero. To prevent such inconsistencies from arising, a starting value should be chosen in the interior of the parameter space. A good default starting value is a uniform table, in which all the elements of θ are equal.

The algorithm in Figure 7.2 calculates an ML estimate, but with a slight modification it can also be used to find a posterior mode under a Dirichlet prior. The E-step remains the same, but the M-step must be altered to maximize the complete-data posterior density rather than the complete-data likelihood. If the prior distribution of θ is Dirichlet with hyperparameter $\alpha = \{\alpha_y : y \in Y\}$, then the last line in Figure 7.2 should be changed to

$$\text{for } y \in Y \text{ do } \theta_y := \left(x_y + \alpha_y - 1\right) / \left(n + \alpha_0 - D\right), \tag{7.39}$$

where α_0 and D is the total number of cells in θ. Taking $\alpha_y = 1$ for all $y \in Y$ results in a uniform prior, under which the posterior mode and the ML estimate coincide. Notice that if any $\alpha_y < 1$ and the corresponding cell count x_y is zero, then (7.39) will produce a negative estimate for θ_y. For computing posterior modes, priors with $\alpha_y < 1$ are not recommended.

Random zeroes and structural zeroes

When cells of the observed-data tables $z^{(1)},\ldots, z^{(S)}$ are empty not because the events corresponding to those cells are impossible but merely as an artifact of chance, the cells are said to contain random zeroes. Random zeroes in $z^{(1)},\ldots, z^{(S)}$ may have two undesirable effects. First, they may produce an ML estimate on the boundary of Θ. Such an estimate is conceptually unattractive, because it implies that some events in the discrete sample space have zero probability even though they have not been deemed impossible on a priori grounds. Second, random zeroes may render certain functions of θ inestimable, in which case the ML estimate will not be unique; the observed-data likelihood will be maximized along a ridge, and EM will converge to different stationary values depending on the starting value (Fuchs, 1982).

When random zeroes result in inestimable parameters or ML estimates on the boundary, the algorithm in Figure 7.2 does not experience any numerical difficulty; it still converges reliably from any starting value in the interior of Θ. The value to which it converges, however, may be a poor estimate for certain functions of θ. When this happens, it is often helpful to apply a Dirichlet prior distribution in which all the hyperparameters are greater than one, e.g. a flattening prior with $\alpha = (c, c,\ldots, c)$ for some $c > 1$, which adds the equivalent of $c - 1$ prior observations to each cell. Another good choice is a data-dependent prior that smooths the estimate toward a null model of independence (Section 7.2.5).

A cell that is empty because the corresponding event is logically impossible is said to contain a structural zero. Structural zeroes are qualitatively different from random zeroes and should not be handled in the same way. Because the probabilities associated with structural zeroes are known to be zero a priori, those cells should be omitted from the estimation procedure. In the algorithm of Figure 7.2, structural zeroes can be handled by providing a starting value for θ in which the elements corresponding to structural zeroes have been set to zero. If the initial value of θ_y is zero, then the first E-step will not allocate any portion of the observed counts in $z^{(1)},\ldots,z^{(S)}$ to cell y, and the resulting expectation $E(x_y|Y_{obs},\theta)$ will be zero. To ensure that the estimate of θ_y remains zero for all subsequent iterations, the last line of the algorithm should be revised to

$$\mathit{for}\; y \in Y^* \;\; \mathit{do}\; \theta_y := \left(x_y + \alpha_y - 1\right) / \left(n + \alpha_0^* - D^*\right), \tag{7.40}$$

where Y^* is the set of all possible values of y excluding the structural zeroes, $\alpha_0^* = \sum_{y \in Y^*} \alpha_y$ is the sum of the prior hyperparameters and D^* is the number of elements in Y^* (i.e. the total number of cells excluding structural zeroes).

Observed-data loglikelihood

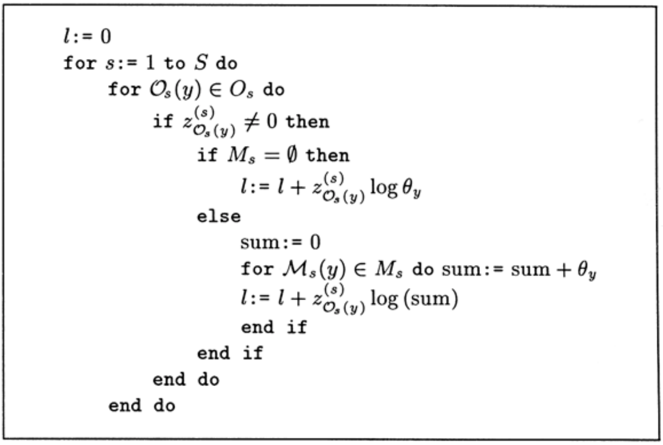
The observed-data loglikelihood function $l(\theta|Y_{obs})$, given by (7.32), and the observed-data log-posterior density

$$\log P\left(\theta \mid Y_{obs}\right) = l\left(\theta \mid Y_{obs}\right) + \log \pi(\theta),$$

are not difficult to calculate for specific values of θ. Evaluating the loglikelihood or log-posterior density can be helpful for monitoring the progress of EM and data augmentation (Sections 3.3.4 and 4.4.3). Pseudocode for evaluating $l(\theta|Y_{obs})$ is shown in Figure 7.4. The loglikelihood at the current value of θ is calculated and stored in l. Notice that this code is very similar to the E-step and could easily be woven into EM.

7.3.3 Data augmentation

Data augmentation for the saturated multinomial model is quite similar to the EM algorithm described above. Recall that in data augmentation, we alternately draw from the predictive distribution of the missing data given the observed data and the parameters (the I-step) and from the complete-data posterior distribution of the parameters (the P-step). The observed data consist of the tables $z^{(s)}$ for missingness patterns $s = 1,\ldots, S$, and the missing data consist of the information needed to expand each $z^{(s)}$ into a full p-dimensional table $x^{(s)}$. The predictive distribution of $x^{(s)}$ given $z^{(s)}$ and θ is the product multinomial given by (7.33)-(7.35). Therefore, the I-step consists of drawing each $x^{(s)}$ from its product multinomial distribution and summing them to obtain a simulated complete-data table $x = x^{(1)} + x^{(2)} + \cdots + x^{(S)}$. Under the Dirichlet prior $\theta \sim D(\alpha)$, the P-step is then just a simulation of θ from its complete-data posterior $D(\alpha + x)$.



```
$l$:= 0
for $s$:= 1 to $S$ do
    for $\mathcal{O}_s(y) \in O_s$ do
        if $z^{(s)}_{\mathcal{O}_s(y)} \neq 0$ then
            if $M_s = \emptyset$ then
                $l$:= $l + z^{(s)}_{\mathcal{O}_s(y)} \log \theta_y$
            else
                sum:= 0
                for $\mathcal{M}_s(y) \in M_s$ do sum:= sum $+ \theta_y$
                $l$:= $l + z^{(s)}_{\mathcal{O}_s(y)} \log$ (sum)
                end if
            end if
        end do
    end do
```

Figure 7.3. *Evaluation of the observed-data loglikelihood function.*

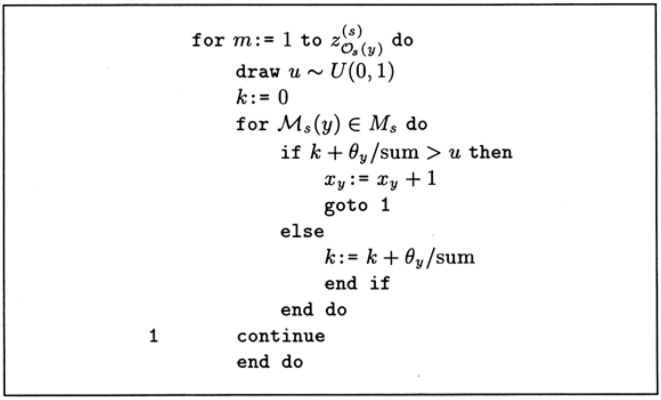
In the pseudocode of Figure 7.2, the line

$$\text{for } M_s(y) \in M_s \text{ do } x_y := x_y + z^{(s)}_{O_s(y)} \theta_y / \text{sum} \qquad (7.41)$$

allocates an observed count $z^{(s)}_{O_s(y)}$ to the cells of the complete-data table in fixed proportions determined by the current value of θ. To convert this E-step into an I-step, the proportional allocation must be replaced by a random allocation; that is, we must replace (7.41) by a routine that will draw

$$x^{(s)}_{O_s(y)} \sim M\left(z^{(s)}_{O_s(y)}, \gamma_{O_s(y)}\right)$$

and add the result into x. One method for simulating the multinomial counts, called *table sampling*, is to compare standard uniform U (0, 1) random variates to cumulative sums of the probabilities in $\gamma_{O_s(y)}$. Pseudocode for table sampling is shown in Figure 7.4. Substituting this code for (7.41) will change the E-step into an I-step. Table sampling can be slow if the counts in the observed-data tables $z^{(s)}$ are large. A more efficient method for simulating multinomial draws in that situation, which relies on a Poisson variate generator, is described by Brown and Bromberg (1984).



```
          for $m$:= 1 to $z^{(s)}_{O_s(y)}$ do
              draw $u \sim U(0,1)$
              $k$:= 0
              for $\mathcal{M}_s(y) \in M_s$ do
                  if $k + \theta_y/\text{sum} > u$ then
                      $x_y$:= $x_y + 1$
                      goto 1
                  else
                      $k$:= $k + \theta_y/\text{sum}$
                      end if
                  end do
    1         continue
          end do
```

Figure 7.4. *Table sampling for the data augmentation I-step.*

To complete the conversion of the EM algorithm to data augmentation, the M-step (the final line of Figure 7.2) must be changed to a P-step; that is, the estimation of θ from the complete-data table x must be replaced by a random draw of θ from the Dirichlet posterior $D(\alpha+x)$. The Dirichlet is easily simulated using standard gamma variates (Section 7.2.3). If any structural zeroes are present, those cells should be omitted from the P-step and their probabilities should be set to zero. If random zeroes occur in $z^{(1)},..., z^{(S)}$ and the improper Dirichlet prior with $\alpha = (0,0, ,0)$ is being used, then depending on the pattern of the zeroes the P-step could be undefined, because some elements of a + x could be zero. For this reason, the prior $\alpha = (0, 0,..., 0)$ should be avoided whenever random zeroes are present. A proper prior, e.g. a flattening prior with $\alpha = (c, c,..., c)$ for some positive value of c, should be used instead.

Imputation of unit-level missing data

The I- and P-steps of the data augmentation algorithm described above operate on the sufficient statistics stored in the workspace x. After enough steps have been taken to achieve approximate stationarity, x will contain a simulated draw from the posterior predictive distribution of the complete-data contingency table $P(x|Y_{obs})$. If the algorithm is being used for multiple imputation, however, it may be necessary at the end of the simulation run to impute the missing values at the unit level, i.e. to fill in the missing elements Y_{mis} of the $n \times p$ data matrix Y.

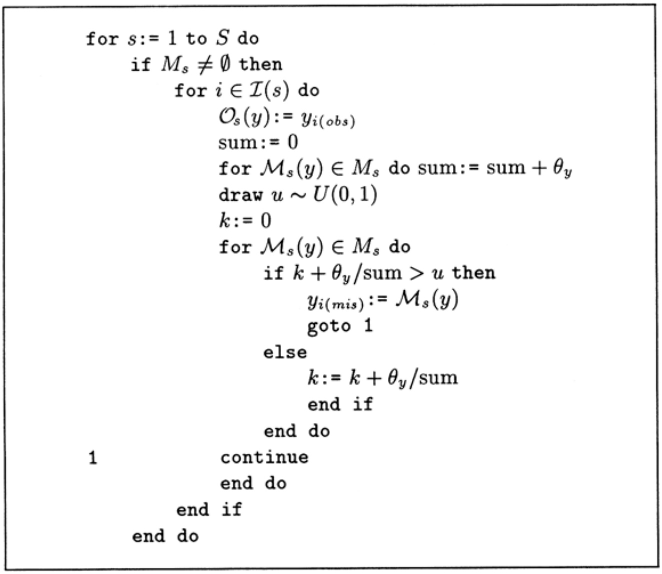
Figure 7.5 shows pseudocode for a modified I-step that imputes the missing elements of Y. Executing this code once at the end of a sufficiently long data augmentation run will result in a proper imputation of Y_{mis} i.e. a simulated draw from $P(\theta|Y_{obs},Y_{mis})$. In Figure 7.5, $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing portions, respectively, of the ith row of the data matrix Y, and $I(s)$ denotes the rows of Y that exhibit missingness pattern s. The vector workspace $y = (y_1, y_2,..., y_p)$ serves as a counter, indexing the cells of the p-dimensional contingency table. For any row i in missingness pattern s, the subvector $O_s(y)$ of y is first set equal to the observed data in $y_{i(obs)}$, so that the remaining portion $M_s(y)$ indexes all the cells of the contingency table into which observation i might fall. The missing values in $y_{i(mis)}$ are then drawn simultaneously by table sampling, comparing a single uniform variate u to the set of probabilities derived from θ that describe the conditional distribution of $y_{i(mis)}$ given $y_{i(obs)}$.

7.3.4 Example: victimization status from the National Crime Survey

Recall the data of Table 3.3 from the National Crime Survey, in which households were classified according to whether they had been victimized by crime in two six-month periods. In the sample of 756 households, 38 had victimization status missing for the first period, 42 had status missing for the second period and 115 had status missing for both periods. Using the EM algorithm and likelihood-ratio tests, we found very strong evidence that victimization status on the two occasions was related; the p-value for testing the hypothesis of independence was essentially zero. Moreover, we found fairly strong evidence that the rates of victimization in the two periods were not equal; the p-value for testing the hypothesis of marginal homogeneity/symmetry was 0.06 (Section 3.2.4).

Analysis by parameter simulation

Tests of independence and marginal homogeneity/symmetry can also be readily carried out by parameter simulation. To test a hypothesis by parameter simulation, we first select a function of the



```
for $s$:= 1 to $S$ do
     if $M_s \neq \emptyset$ then
          for $i \in \mathcal{I}(s)$ do
               $\mathcal{O}_s(y)$:= $y_{i(obs)}$
               sum:= 0
               for $\mathcal{M}_s(y) \in M_s$ do sum:= sum + $\theta_y$
               draw $u \sim U(0,1)$
               $k$:= 0
               for $\mathcal{M}_s(y) \in M_s$ do
                    if $k + \theta_y/\text{sum} > u$ then
                         $y_{i(mis)}$:= $\mathcal{M}_s(y)$
                         goto 1
                    else
                         $k$:= $k + \theta_y/\text{sum}$
                         end if
                    end do
1              continue
               end do
          end if
     end do
```

Figure 7.5. *I-step for imputing missing values at the unit level.*

cell probabilities θ that measures the degree to which θ departs from the null hypothesis, and simulate the posterior distribution of this quantity given the observed data. For independence, a natural quantity to examine is the odds ratio

$$\omega = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}}, \tag{7.42}$$

where θ_{ij} denotes the probability of ($Y_1 = i$, $Y_2 = j$) for $i, j = 1$, 2. The proportion of simulated values of ω that are less than or equal to one can be taken as an approximate one-sided p-value for testing the hypothesis of independence ($\omega = 1$) against the alternative that households victimized in the first period were more likely to be victimized in the second period ($\omega > 1$). For marginal homogeneity/symmetry, we can examine the difference in victimization rates between the second period ($\theta_{+2} = \theta_{12} + \theta_{22}$) and the first period ($\theta_{+2} = \theta_{12} + \theta_{22}$).

$$\begin{aligned}\delta &= \theta_{+2} - \theta_{2+} \\ &= \theta_{12} - \theta_{21}.\end{aligned} \tag{7.43}$$

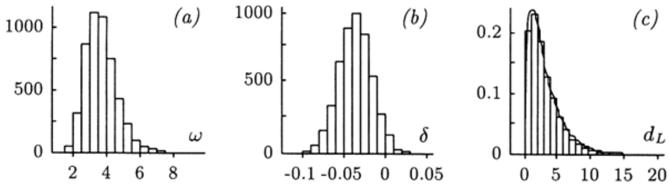


Figure 7.6. *Histograms of (a)* $\omega = (\theta_{11}\theta_{22})/(\theta_{12}\theta_{21})$ *and (b)* $\delta = \theta_{12}-\theta_{21}$ *over 5000 iterations of data augmentation, and of (c) the likelihood- ratio statistic* d_L *with the* χ_3^2 *density superimposed.*

The proportion of simulated values of δ that fall above zero is an approximate one-sided p-value for testing the hypothesis of no change ($\delta = 0$) against the alternative that the victimization rate has dropped ($\delta < 0$).

One interesting question is whether the 115 households for which both variables are missing should be included in the simulations. From an inferential standpoint it does not matter; under the ignorability assumption these sample units contribute nothing to the likelihood function for θ, so likelihood-based or Bayesian inferences for θ will be the same whether these units are included or not. From a computational standpoint, however, it is slightly better to omit them, because their presence needlessly increases the fractions of missing information and slows the convergence of data augmentation. In this particular example, the difference is barely noticeable. Without these 115 cases, the worst fraction of missing information as estimated from the iterations of EM (Section 3.3.4) is about 13%. Including these cases, it rises to 26%. Either way, data augmentation converges very quickly; in preliminary runs under the Jeffreys prior (all $\alpha = 1/2$), the autocorrelations in scalar functions of θ essentially died out after lag 2 or 3 even when the 115 cases were included.

Starting from the ML estimate of θ, we simulated 5000 steps of data augmentation under the Jeffreys prior following a burn-in period of 100 steps. Histograms of the simulated values of ω and δ are shown in Figure 7.6 (a) and (b), respectively. All 5000 values of w were greater than one, so the simulated p-value for the test of independence is zero. Of the 5000 values of δ, 164 fell above zero, so the simulated p-value for testing $\delta = 0$ against the one-sided alternative $\delta < 0$ is 164/5000 = 0.033; the p-value against the two-sided alternative is $2 \times 0.033 = 0.066$.

Notice that these simulated p-values agree closely with those from the likelihood-ratio tests performed in Chapter 3. Because of the large sample size and the small number of parameters in this example, Bayesian and likelihood-based inferences are essentially identical. Further evidence that the large-sample properties are working well is provided by the posterior distribution of the likelihood-ratio statistic. The quantity

$$d_L = 2\left[l\left(\hat{\theta} \mid Y_{obs}\right) - l\left(\theta \mid Y_{obs}\right)\right],$$

where $\hat{\theta}$ is the ML estimate, has (when regarded as a function of θ) a posterior distribution that is asymptotically chisquare with three degrees of freedom, because the multinomial model for this example has three free parameters. A histogram of the 5000 simulated values of d_L is shown in Figure 7.6 (c) with the χ_3^2 density function superimposed; the two are nearly indistinguishable.

By averaging the 5000 iterates of ω and δ, we obtain simulated posterior means

$$\hat{E}\left(\omega \mid Y_{obs}\right) = 3.67 \text{ and } \hat{E}\left(\delta \mid Y_{obs}\right) = -0.036.$$

Comparing these to the ML estimates obtained in Section 3.2.2,

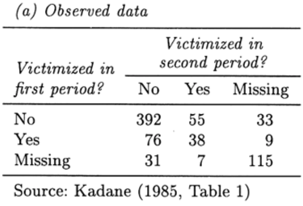
$$\hat{\omega} = 3.57 \text{ and } \hat{\delta} = -0.037,$$

we find that the agreement is close. Simulated 95% posterior intervals for ω and δ based on sample quantiles of the 5000 iterates are (2.20, 5.77) and (-0.076, 0.001), respectively.

Analysis by multiple imputation

In this example, it is also straightforward to conduct inferences by multiple imputation. We generated a set of m = 10 imputations by running ten independent chains of data augmentation for 100 steps, starting each chain from the ML estimate. To speed convergence, the 115 households for which both variables were missing were omitted from the sample. At the final I-step of each chain, however, these households were restored to the sample so that their missing data could be imputed. Because these households contribute nothing to the observed-data likelihood, inferences will be essentially the same whether they are included or not. We decided to include them in the final I-steps so that the variation among the imputed datasets would more accurately reflect the real levels of missing-data uncertainty. The observed data and ten imputations of the complete-data table are shown in Table 7.1.

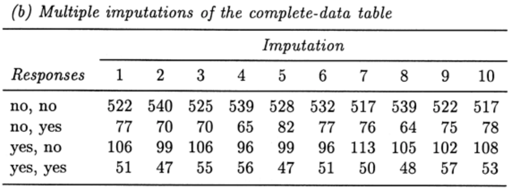
Table 7.1. *Victimization status for households in the National Crime Survey, with m = 10 multiple imputations*



(a) Observed data

Victimized in first period?	*Victimized in second period?* No	Yes	Missing
No	392	55	33
Yes	76	38	9
Missing	31	7	115

Source: Kadane (1985, Table 1)



(b) Multiple imputations of the complete-data table

	Imputation									
Responses	1	2	3	4	5	6	7	8	9	10
no, no	522	540	525	539	528	532	517	539	522	517
no, yes	77	70	70	65	82	77	76	64	75	78
yes, no	106	99	106	96	99	96	113	105	102	108
yes, yes	51	47	55	56	47	51	50	48	57	53

As before, let us make inferences about the odds ratio $\omega = (\theta_{11}\theta_{22})/(\theta_{12}\theta_{21})$ and the difference $\delta = \theta_{12} - \theta_{21}$. The standard method for obtaining a point estimate and confidence interval for an odds ratio with complete data is given in Section 6.4.2. The obvious complete-data estimate of the difference δ is $\hat{\delta} = \hat{\theta}_{12} - \hat{\theta}_{21}$, where $\hat{\theta}_{12} = x_{12}/n$ and $\hat{\theta}_{21} = x_{21}/n$. In large samples $\hat{\delta}$ will be approximately normally distributed, and a consistent estimate of its variance is

$$\hat{V}\left(\hat{\delta}\right) = \frac{1}{n}\left[\hat{\theta}_{12}\left(1 - \hat{\theta}_{12}\right) + \hat{\theta}_{21}\left(1 - \hat{\theta}_{21}\right) + 2\hat{\theta}_{12}\hat{\theta}_{21}\right]$$

by elementary properties of the multinomial distribution (Section 7.2.1). Given these complete-data methods and the ten imputations in Table 7.1 (b), multiple-imputation point and interval estimates were obtained by Rubin's method for scalar estimands (Section 4.3.2). The resulting point estimates for ω and δ are 3.60 and –0.039, respectively, which agree closely with the ML estimates and the simulated posterior means. The resulting 95% interval estimates are (2.15, 6.04) and (-0.079, 0.001), which also agree well with the intervals obtained through parameter simulation. Estimated fractions of missing information for ω and δ are 35% and 26%, respectively.

7.3.5 Example: Protective Services Project for Older Persons

Fuchs (1982) analyzed data from the Protective Services Project for Older Persons, a longitudinal study designed to measure the impact of enriched social casework services on the well-being of elderly clients (Blenkner et al., 1971). For 101 clients in the study, six dichotomous variables were recorded:

Variable	*Levels*	*Code*
Group membership	1 = experimental, 2 = control	*G*
Age	1 = under 75, 2 = 75+	*A*
Sex	1 = male, 2 = female	*S*
Survival status	1 = deceased, 2 = survived	*D*
Physical status	1 = poor, 2 = good	*P*
Mental status	1 = poor, 2 = good	*M*

For an additional 63 clients, values of physical and/or mental status were missing. The observed dataset, including complete and incomplete cases, is shown in Table 7.2.

Results from this project generated considerable controversy in the social work literature. Some (Fischer, 1973) argued that the enriched services seemed to be detrimental to the clients, because the mortality rate for the experimental group was actually higher than for the control group. Classifying the subjects by only G and D, both of which are observed for the entire sample, we obtain the marginal frequencies displayed in Table 7.3. The test for independence in this table, based on the well-known Pearson X^2 statistic, yields $X^2 = 5.03$ with one degree of freedom; the approximate p-value is 0.025, which provides fairly strong evidence that G and P are related. The estimated odds ratio is 2.04, suggesting that subjects in the experimental group were about twice as likely (on the odds scale) to die than subjects in the control group.

If subjects had been assigned to treatments in a random fashion, then Table 7.3 would indeed provide evidence that the services given to the experimental group were detrimental. If we examine the relationships between G and the other variables, however, we find that the treatment assignments were not random. Subjects in the experimental group tended to be older, and also tended to have poorer physical and mental status, than subjects in the

Table 7.2. *Data from the Protective Services Project for Older Persons*

			Male				Female			
			< 75		≥ 75		< 75		≥ 75	
Mental	*Physical*	*Survival*	E†	C†	E	C	E	C	E	C
		(a) Fully categorized								
Poor	Poor	Deceased	0	2	5	3	0	0	2	1
		Survived	1	0	0	0	0	0	0	1
	Good	Deceased	0	0	2	2	1	1	1	0
		Survived	0	2	2	0	0	0	0	0
Good	Poor	Deceased	0	0	3	1	0	0	1	2
		Survived	3	1	1	2	0	1	1	0
	Good	Deceased	1	1	4	6	2	0	0	2
		Survived	5	10	6	8	3	5	2	4
		(b) Missing physical status								
Poor	Missing	Deceased	0	0	0	0	0	0	0	0
		Survived	0	0	1	0	0	0	0	0
Good		Deceased	0	0	0	0	0	0	0	0
		Survived	0	0	0	0	0	0	0	0
		(c) Missing mental status								
Missing	Poor	Deceased	2	0	5	3	1	1	2	0
		Survived	1	1	0	3	0	0	0	1
	Good	Deceased	1	0	0	0	0	0	1	2
		Survived	1	3	2	1	1	1	0	0
		(d) Missing both physical and mental status								
Missing	Missing	Deceased	0	1	2	2	1	0	3	1
		Survived	2	8	1	2	1	1	2	2

†E denotes experimental; C denotes control. Source: Fuchs (1982)

Table 7.3 *Classification of subjects by G and D*

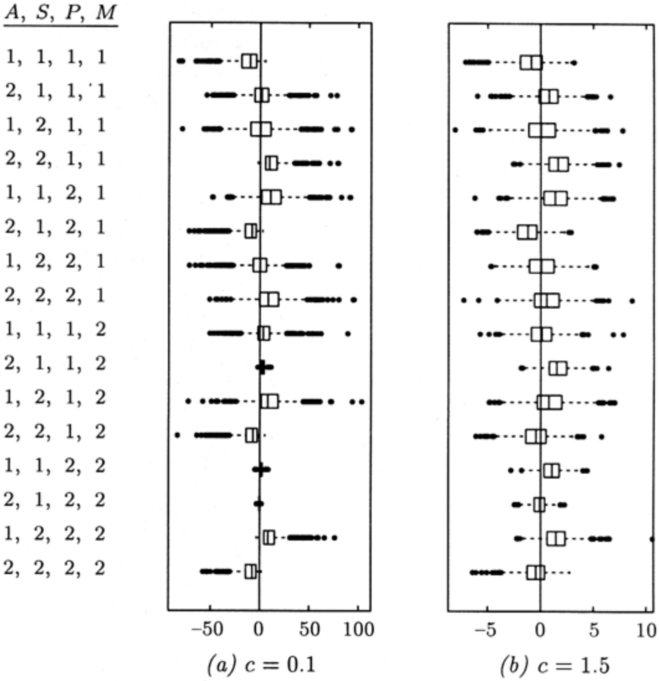
	Survived?	
Group	No	Yes
Experimental	40	36
Control	31	57

control group. It appears that the investigators tended to give the enriched services to clients who appeared to have the greatest need for them. The marginal association between *G* and *D* could thus be due, at least in part, to the fact that the subjects in the experimental group were simply more prone to die than the subjects in the control group, regardless of any services they received. Rather than examining the marginal association between *G* and *D*, we ought to focus on their conditional associations given the covariates A, S, P and M, to see whether G and D are still related after the possibly confounding effects of these covariates have been removed. That is, we should examine the odds ratios for G and D within the sixteen 2 × 2 tables that correspond to the unique combinations of the levels of A, S, P and M.

The complete-data contingency table has $2^6 = 64$ cells; with a sample size of $n = 164$, this results in an average of only 2.6 observations per cell. As noted by Fuchs (1982), the ML estimate of θ under the saturated model is not unique due to the pattern of random zeroes in the observed-data tables. Moreover, the suprema of the likelihood function lie on the boundary of the parameter space. To make EM converge to a unique mode in the interior, a Dirichlet prior was applied with $\alpha=(c,c, ...,c)$ for $c = 1.1$ which adds the equivalent of 6.4 prior observations and spreads them uniformly across the 64 cells. Then, taking this mode as a starting value, single chains of data augmentation were simulated under two alternative priors: $c = 0.1$ and $c = 1.5$. Each chain was run for 1000 steps following a burn-in period of 200 steps.

Boxplots of the simulated GD odds ratios for each of the sixteen ASPM combinations are shown in Figure 7.7. The odds ratios are plotted on the natural log scale, with positive values indicating a positive association between enriched services ($G = 1$) and death ($D = 1$). Under the $c = 0.1$ prior, the simulated odds ratios show enormous variability; this prior assigns high probability to regions of the parameter space near the boundary, where odds ratios can approach 0 or $+\infty$. Under the stronger prior $c = 1.5$ the situation has improved, but the range of the simulated odds ratios is still implausibly wide. Notice that under either prior, all of the boxplots straddle the null value of zero, and there is no overwhelming tendency for the boxplots to be centered either to the left or to the right of zero. Thus there seems to be no strong evidence against the null hypothesis that G and D are unrelated.

To further sharpen the posterior distributions, we could increase the value of c even more. But this does not seem appropriate,



(a) $c = 0.1$

(b) $c = 1.5$

Figure 7.7. *Boxplots of simulated log-odds ratios from 1000 iterations of data augmentation under two flattening priors.*

because with c = 1.5 we have already added the equivalent of 1.5 × 64 = 96 prior observations with respect to estimation of the elements of θ. It appears that modest amounts of prior information are not sufficient to stabilize the inference; the observed data are simply too sparse to support the estimation of separate odds ratios within each cell of the ASPM classification. We will deal with this problem of sparseness in Chapter 8 by fitting a simpler model that assumes a common odds ratio for all sixteen levels of ASPM.

7.4 Fast algorithms for near-monotone patterns

7.4.1 Factoring the likelihood and prior density

In Chapter 6 we introduced a class of algorithms called monotone data augmentation. Monotone data augmentation is similar to ordinary data augmentation except that in each I-step we impute only enough of the missing values to complete a monotone pattern. The advantage of monotone data augmentation is that it tends to converge very quickly when the observed data are nearly monotone. In this section we present monotone data augmentation for the saturated multinomial model.

Monotone data augmentation is feasible when the prior and likelihood for the complete data factor neatly into independent pieces corresponding to the marginal distribution of Y_1, the conditional distribution for Y_2 given Y_1, the conditional distribution for Y_3 given Y_1 and Y_2, and so on. Let us first consider the likelihood. Until now we have been describing the data by a single multinomial distribution for the complete-data contingency table x, but we can equivalently characterize this model as a sequence of product-multinomials. Suppose we write

$$\begin{aligned} P\left(Y_1,\ldots,Y_p \mid \theta\right) = {} & P\left(Y_1 \mid \phi_1\right) P\left(Y_2 \mid Y_1, \phi_2\right) \\ & \cdots P\left(Y_p \mid Y_1,\ldots,Y_{p-1},\phi_p\right), \end{aligned} \tag{7.44}$$

where ϕ_j denotes the parameters governing the conditional distribution of Y_j given $(Y_1,\ldots, Y_{j-1})$. Each of the factors of the right-hand side of (7.44) corresponds to a product-multinomial distribution on a collapsed version of x.

To be more precise we need some additional notation. Suppose that $y = (y_1, y_2,\ldots, y_p)$ is a generic realization of $(Y_1, Y_2,\ldots, Y_p)$ for a single unit. Let F_j be a function that extracts from y the first j elements,

$$F_j(y) = \left(y_1,\ldots,y_j\right),$$

and let L_j extract the last $p - j$ elements,

$$L_j(y) = \left(y_{j+1}, \ldots, y_p\right).$$

Let F_j and L_j, respectively, be the sets over which $F_j(y)$ and $L_j(y)$ are allowed to vary; that is, F_j will be the Cartesian cross-product of the sets $\{1, 2, \ldots, d_k\}$ for $k = 1, \ldots, j$, and L_j the cross-product for $k = j + 1, \ldots, p$. We will write the probability of the event Y_1, y_1, $Y_2 = y_2, \ldots, Y_j = y_j$ as

$$\xi_{F_j(y)} = \sum_{L_j(y) \in L_j} \theta_y,$$

and the full set of parameters governing the marginal distribution of $(Y_1, Y_2, \ldots, Y_j)$ as

$$\xi_j = \left\{\xi_{F_j(y)} : F_j(y) \in F_j\right\}.$$

The conditional probability of the event $Y_j = y_j$ given that $Y_1 = y_1$, $Y_2 = y_2, \ldots, Y_{j-1} = y_{j-1}$ will be

$$\phi_{F_j(y)} = \xi_{F_j(y)} / \xi_{F_{j-1}(y)}, \tag{7.45}$$

and the full set of parameters governing the conditional distribution of Y_j given $(Y_1, Y_2, \ldots, Y_{j-1})$ is

$$\phi_j = \left\{\phi_{F_j(y)} : F_j(y) \in F_j\right\}.$$

Suppose we collapse the p-dimensional contingency table x on its last $p - j$ dimensions, producing a table that cross-classifies the units by $(Y_1, Y_2, \ldots, Y_j)$. Denote a frequency in this table by

$$z_{F_j(y)} = \sum_{L_j(y) \in L_j} x_y,$$

and the entire j-dimensional table by

$$z_j = \left\{ z_{F_j(y)} : F_j(y) \in F_j \right\}.$$

By the rules for collapsing and partitioning (Section 7.2.2), z_j has a multinomial distribution with index n and parameter ξ_j. Moreover, the conditional distribution of z_j given z_{j-1} is a product-multinomial whose parameters are contained in ϕ_j. More specifically, suppose we partition z_j into a set of $d_1 \times d_2 \times \cdots d_{j-1}$ vectors, each of length d_j. Denote one of these vectors by

$$z_{j;F_{j-1}(y)} = \left\{ z_{F_j(y)} : y_j = 1, 2, ..., d_j \right\},$$

which is simply the portion of z_j obtained by fixing $(y_1,..., y_{j-1})$ at a specific value but letting y_j vary over $\{1, 2,..., d_j\}$. The table z_j is then the collection of these vectors,

$$z_j = \left\{ z_{j;F_{j-1}(y)} : F_{j-1}(y) \in F_{j-1} \right\}.$$

If we partition ϕ_j in the same fashion, as

$$\phi_j = \left\{ \phi_{j;F_{j-1}(y)} : F_{j-1}(y) \in F_{j-1} \right\}$$

where

$$\phi_{j;F_{j-1}(y)} = \left\{ \phi_{F_j(y)} : y_j = 1, 2, ..., d_j \right\},$$

then the conditional distribution of z_j given z_{j-1} is

$$z_{j;F_{j-1}(y)} \mid z_{j-1}, \phi_j \sim M\left(z_{F_{j-1}(y)} \phi_{j;F_{j-1}(y)} \right) \qquad (7.46)$$

independently for all $F_{j-1}(y) \in F_{j-1}$.

By these properties, it follows that the multinomial likelihood function for any ξ_j can be factored as

$$L(\xi_j \mid z_j) = L(\xi_{j-1} \mid z_{j-1}) L(\phi_j \mid z_j),$$

the product of a multinomial likelihood for ξ_{j-1} whose sufficient statistics are contained in z_{j-1} and a productmultinomial likelihood for ϕ_j whose sufficient statistics are contained in z_j. Applying this factorization recursively, first to $\xi_p = \theta$, then to ξ_{p-1}, and so on down to ξ_2, we obtain

$$L(\phi \mid Y) = \prod_{j=1}^{p} L(\phi_j \mid z_j),$$

where each factor $L(\phi_j|z_j)$ is a product-multinomial likelihood. The full set of parameters $\phi = (\phi_1,\phi_2, ...,\phi_p)$ forms a one-to-one transformation of θ, and it follows from (7.45) that the back-transformation is

$$\theta_y = \phi_{F_1(y)} \phi_{F_2(y)} \cdots \phi_{F_p(y)}.$$

Factoring the prior

Just as the likelihood function factors into independent pieces for $\phi_1,\phi_2, ...,\phi_p$, the density function for ϕ induced by the ordinary Dirichlet prior on θ also factors into a product of independent densities. Suppose that a priori θ has a Dirichlet distribution,

$$\theta \sim D(\alpha), \tag{7.48}$$

where the hyperparameters are regarded as an array with the same dimensions as θ,

$$\alpha = \left\{ \alpha_y : y \in Y \right\}.$$

By the collapsing rules for the Dirichlet discussed in Section 7.2.6, the distribution for ξ_j implied by (7.48) is also Dirichlet. The parameters of this distribution, which we shall call

$$\beta_j = \left\{ \beta_{F_j(y)} : F_j(y) \in F_j \right\},$$

are obtained by summing the elements of a in the same way the elements of θ were summed to produce ξ_j,

$$\beta_{F_j(y)} = \sum_{L_j(y) \in L_j} \alpha_y.$$

Moreover, by the results of Section 7.2.6, the conditional distribution of ϕ_j given ξ_{j-1} for any j is a product of independent Dirichlet distributions. That is, if we partition the j-dimensional table β_j in precisely the same manner as we partitioned ϕ_j, as

$$\beta_j = \left\{ \beta_{j;F_{j-1}(y)} : F_{j-1}(y) \in F_{j-1} \right\}$$

where

$$\beta_{j;F_{j-1}(y)} = \left\{ \beta_{F_j(y)} : y_j = 1, 2, \ldots, d_j \right\},$$

the conditional distribution of ϕ_j given ξ_{j-1} is

$$\phi_{j;F_{j-1}(y)} \mid \xi_{j-1} \sim D\left(\beta_{j;F_{j-1}(y)} \right) \tag{7.49}$$

independently for all $F_{j-1}(y) \in F_{j-1}$.

Now from (7.45) it is clear that ξ_j is a one-to-one function of $(\phi_1, \ldots, \phi_j)$ for any j. The prior density for $\phi = (\phi_1, \ldots, \phi j)$ can thus be written

$$\begin{aligned}\pi(\phi) &= \pi_1(\phi_1)\prod_{j=2}^{p}\pi_j(\phi_j \mid \phi_1,\ldots,\phi_{j-1}) \\ &= \pi_1(\phi_1)\prod_{j=2}^{p}\pi_j(\phi_j \mid \xi_{j-1}).\end{aligned} \tag{7.50}$$

But notice that ξ_{j-1} does not appear on the right-hand side of (7.49); thus ϕ_j is independent of ξ_{j-1}, and (7.50) becomes

$$\pi(\phi) = \prod_{j=1}^{p}\pi_j(\phi_j), \tag{7.51}$$

where each of the terms $\pi_j(\phi_j)$ is a product of independent Dirichlet densities whose parameters are contained in β_j.

7.4.2 Monotone data augmentation

By the factorizations described above, it immediately follows that complete-data Bayesian inferences under the saturated multinomial model and Dirichlet prior,

$$\begin{gathered} x \mid \theta \sim M(n,\theta), \\ \theta \sim D(\alpha), \end{gathered}$$

can be carried out as a sequence of independent Bayesian inferences for $\phi_1, \phi_2, \ldots, \phi_p$,

$$P(\phi \mid Y) = \prod_{j=1}^{p} P(\phi_j \mid z_j).$$

By combining (7.46) with (7.49), we see that the complete-data posterior distribution for any term ϕ_j is

$$\phi_{j;F_{j-1}(y)} \mid z_j \sim D\left(\beta_{j;F_{j-1}(y)} + z_{j;F_{j-1}(y)}\right) \tag{7.52}$$

independently for all $F_{j-1}(y) \in F_{j-1}$.

This factorization of the posterior applies not only when the data are complete; more generally, it holds whenever the observed data form a monotone pattern as described in Section 6.5. Suppose that the observed data are monotone in the sense that if Y_j is missing for a unit, then $Y_{j+1}, \ldots, Y_p$ are missing as well (Figure 6.8). By essentially the same argument as was given in Section 6.5.1, the observed-data likelihood for ϕ given Y_{obs} can be factored as

$$L(\phi \mid Y_{obs}) = \prod_{j=1}^{p} L\left(\phi_j \mid z_j^*\right),$$

where z_j^* is the contingency table that cross classifies all the units for which Y_j is observed by their values of $Y_1, \ldots, Y_j$. If we denote a cell of this table by $z_{F_j(y)}^*$ and let

$$z_{j;F_{j-1}(y)}^* = \left\{ z_{j;F_j(y)}^* : y_j = 1, 2, \ldots, d_j \right\}$$

be a subvector within this table, $L\left(\phi_j \mid z_j^*\right)$ will be the likelihood that arises from the product-multinomial distribution

$$z_{j:F_{j-1}(y)}^* \mid z_{j-1}^*, \phi_j \sim M\left(z_{F_{j-1}(y)}, \phi_{j;F_{j-1}(y)} \right)$$

for all $F_{j-1}(y) \in F_{j-1}$. Combining this new likelihood with the prior (7.49) leads to the observed-data posterior

$$P(\phi \mid Y_{obs}) = \prod_{j=1}^{p} P\left(\phi_j \mid z_j^*\right), \tag{7.53}$$

where $P\left(\phi_j \mid z_j^*\right)$ is given by

$$\phi_{j;F_{j-1}(y)} \mid z_j^* \sim D\left(\beta_{j;F_{j-1}(y)} + z_{j;F_{j-1}(y)}^* \right) \tag{7.54}$$

for all $F_{j-1}(y) \in F_{j-1}$.

Monotone data augmentation capitalizes on (7.53) to create an efficient simulation algorithm for situations where Y_{obs} is non-monotone. Suppose that Y_{obs} is no longer monotone, but we have identified a subset Y_{mis*} of Y_{mis}, such that (Y_{obs}, Y_{mis*}) is monotone. The monotone data augmentation algorithm alternates between the following two steps.

1. I-step: Simulate a value of Y_{mis*} from its predictive distribution given the current value of θ,

$$Y_{mis*}^{(t+1)} \sim P\left(Y_{mis*} \mid Y_{obs}, \theta^{(t)}\right).$$

2. P-step: Draw a new value of θ from its posterior distribution given Y_{obs} and the new value of Y_{mis*},

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis*}^{(t+1)}\right).$$

In practice, the I-step is identical to that of ordinary data augmentation (Section 7.3.3) except that we need only draw the elements of Y_{mis*} rather than the full Y_{mis}. The P-step is carried out by drawing $\phi_1, ...,\phi_p$ from the factored posterior (7.53), and then numerically transforming the resulting value of $\phi = (\phi_1, ...,\phi_p)$ back to the θ-scale using (7.47).

Interleaving the I- and P-steps

Notice that the simulation of ϕ_j within a P-step does not require knowledge of the most recent simulated value of the entire Y_{mis*} rather, it requires only the most recent value of the j-dimensional table z_j^*. This allows us to interleave portions of the I- and P-steps in the following manner. Suppose that the data are grouped by missingness pattern and sorted as shown in Figure 6.10. Let s_j denote the last pattern for which variable Y_j may need to be filled in to complete the overall monotone pattern, so that $s_p \leq s_{p-1} \leq \cdots \leq s_1$, and for convenience define $s_{p+1} = 0$. Let T_1, T_2 and T_3 be three workspace arrays, each of dimension $d_1 \times d_2 \times d_p$. Initialize T_1 and T_2 to be equal to the current parameter value $\theta^{(t)}$ and α, respectively, and initialize all the elements of T_3 to one. Then, for $j := p, p - 1, ..., 1$, perform the following steps:

1. If $s_j > s_{j+1}$, impute the missing data for variables $Y_1, ..., Y_j$ within patterns $s_{j+1} + 1$ up to s_j. These data should be drawn from their predictive distribution given the observed data and the parameters stored in T_1.

2. Cross-classify the units in patterns $s_{j+1} + 1$ up to s_j by their observed or imputed values for $Y_1, ..., Y_j$, and add the resulting counts into the corresponding cells of the workspace T_2. Upon completion of this step, T_2 will contain β_j plus the simulated value of z_j^*.

3. Draw a value of ϕ_j from its product-multinomial posterior distribution (7.54) given the value of $\beta_j + z_j^*$ in T_2. Multiply the elements of the array T_3 by the corresponding elements of this simulated ϕ_j.

4. If $j > 1$, collapse T_1 by summing along its jth dimension, thereby reducing its size to $d_1 \times \cdots \times d_{j-1}$. Now T_1 contains the current value of ξ_{j-1} (the parameters of the joint distribution of $Y_1, ..., Y_{j-1}$) which will be necessary for the next Step 1. Perform this same collapsing operation for T_2, preparing it for the next Step 2.

After all p-cycles of Steps 1-4 have been completed, the workspace T_3 will contain the updated parameter $\theta^{(t+1)}$.

Running this algorithm from a starting value $\theta^{(0)}$ generates a sequence of parameter values $\{\theta^{(t)} : t = 1, 2, ...\}$ which converges in distribution to the correct observed-data posterior,

$$P\left(\theta^{(t)} \mid Y_{obs}, \theta^{(0)}\right) \rightarrow P(\theta \mid Y_{obs}) \text{ as } t \rightarrow \infty$$

Convergence tends to be faster than for the ordinary data augmentation algorithm described in Section 7.3, because Y_{mis*} contains less information about the parameter than does Y_{mis}. The most dramatic improvements are seen when Y_{obs} is nearly monotone, because then Y_{mis*} is only a small subset of Y_{mis}. When the observed data happen to be monotone, Y_{mis*} is empty and the algorithm converges from any starting value in a single step.

This algorithm can be used to generate proper multiple imputations of the missing data Y_{mis} as follows. First, simulate a small number of independent draws of θ from $P(\theta|Y_{obs})$, either by running multiple chains or subsampling a single chain. Then, under each of these θ values, impute the full set of missing data Y_{mis} using the ordinary data augmentation I-step (Figure 7.5).

7.4.3 Example: driver injury and seatbelt use

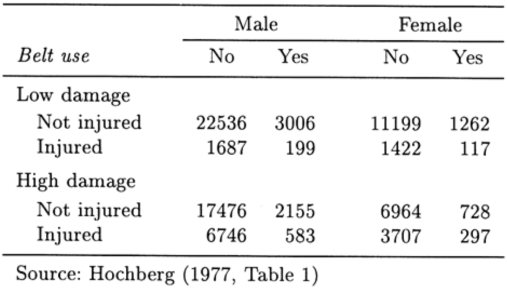
The data in Tables 7.4 and 7.5, previously analyzed by Hochberg (1977) and Chen (1989), concern the effectiveness of seatbelts in reducing the risk of driver injury in automobile accidents. Table 7.4 classifies 80 084 automobile accidents according to four variables obtained from police reports: driver's sex, car damage (low, high), belt use (no, yes) and injury (no, yes). At first glance, these data suggest that the use of seatbelts substantially reduces the risk of injury. The estimated odds of injury are

$$\frac{199+117+583+297}{3006+1262+2155+728}=0.167$$

for belted drivers and

$$\frac{1687+1422+6746+3707}{22536+11199+17476+6964}=0.233$$

Table 7.4. *Classification of accidents by police reports of driver's sex, car damage, injury and belt use*



	Male		Female	
Belt use	No	Yes	No	Yes
Low damage				
Not injured	22536	3006	11199	1262
Injured	1687	199	1422	117
High damage				
Not injured	17476	2155	6964	728
Injured	6746	583	3707	297

Source: Hochberg (1977, Table 1)

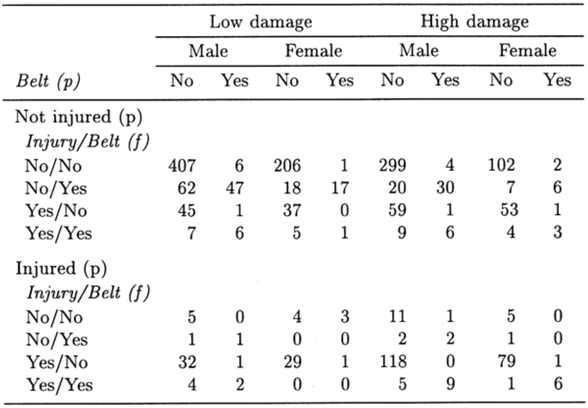
for unbelted drivers, giving an odds ratio of 0.717; an approximate 95% confidence interval for this ratio is (0.673,0.765). This simple analysis is unconvincing, however, for a number of reasons. First, the belted and unbelted groups tend to differ with respect to a variety of characteristics (e.g. sex), and to the extent that these characteristics may be related to the risk of injury, our estimate of the effectiveness of seatbelts may be biased upward or downward.

Another difficulty with this analysis is that the data provided by the police reports are not always accurate, especially with respect to belt use and injury. Experience has shown that the police were prone to overestimate the proportion of drivers who were not injured and unbelted, and that the biases toward not injured were especially severe for low-damage accidents. Even small rates of misclassification with respect to belt use and injury can have a large impact on the estimated effect of wearing a seatbelt.

To examine the effect of misclassification errors, followup data were collected for an additional sample of 1796 accidents. Subsequent to the police reports, investigators obtained more reliable data on belt use and injury from hospital records and personal interviews. We will assume that the information obtained in this followup effort is correct. Data from the followup study are shown in Table 7.5, with the police-reported and followup values of belt use and injury indicated by (p) and (f), respectively.

The followup data in Table 7.5 may be used in a variety of ways. For example, we may ignore the police reports entirely and estimate the seatbelt effect from the followup data alone. Presumably, such estimates would be less biased than those we obtained from Table 7.4, because they would be less prone to misclassification error. On the other hand, they would have greater variability because they would be based on a much smaller sample. A more effective approach would be to combine the data from Tables 7.4 and 7.5 and analyze them as a six-variable dataset with two of the variables partially missing. Combining the two sources would allow us to make use of the police-report data for all 81880 accidents, but would calibrate them to correct for occasional misclassification errors in keeping with the error rates seen in the followup study. In other words, a combined analysis would allow the police-report data to serve as a proxy for the followup data among the

Table 7.5. *Classification of accidents by driver's sex, car damage, injury and belt use obtained from police reports (p), and injury and belt use obtained from followup (f)*



	Low damage				High damage			
	Male		Female		Male		Female	
Belt (p)	No	Yes	No	Yes	No	Yes	No	Yes
Not injured (p)								
Injury/Belt (f)								
No/No	407	6	206	1	299	4	102	2
No/Yes	62	47	18	17	20	30	7	6
Yes/No	45	1	37	0	59	1	53	1
Yes/Yes	7	6	5	1	9	6	4	3
Injured (p)								
Injury/Belt (f)								
No/No	5	0	4	3	11	1	5	0
No/Yes	1	1	0	0	2	2	1	0
Yes/No	32	1	29	1	118	0	79	1
Yes/Yes	4	2	0	0	5	9	1	6

Source: Hochberg (1977, Table 2)

initial 80084 cases, taking into account the fact that the correlation between the two data sources is less than perfect.

The six-variable combined dataset has a monotone pattern, with followup belt use and injury missing for 97.8% of the cases. Because of the high rate of missingness for these two variables, the EM and ordinary data augmentation algorithms described in Section 7.3 converge very slowly. To illustrate, we ran a single chain of ordinary data augmentation for 5000 steps beginning from the ML estimate using the Jeffreys prior (all hyperparameters equal to 1/2), and monitored the worst linear function of θ as estimated from the trajectory of EM (Section 4.4.3). The sample autocorrelation function for this parameter, plotted in Figure 7.8, reveals extreme long-range dependence. Monotone data augmentation, however, converges in a single step because the observed data are precisely monotone. A sequence of 0 values generated by monotone data augmentation will be

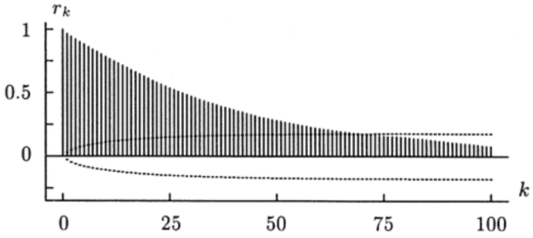


Figure 7.8. *Sample ACFs for the worst linear function of θ, estimated from 5000 iterations of ordinary data augmentation, with dashed lines indicating approximate critical values for testing* $\rho_k=\rho_{k+1}=\cdots=0$.

an actual independent sample from the observed-data posterior $P(\theta|Y_{obs})$.

Using monotone data augmentation, we simulated 1000 independent draws of θ from the observed-data posterior under the Jeffreys prior, and calculated the odds ratios relating seatbelt use to driver injury (both from the police reports and from the followup reports) within each of the four sex-by-damage cells. Boxplots of these simulated odds ratios are shown in Figure 7.9. The odds ratios based on the police reports are highly concentrated to the left of one; the beneficial effects of seatbelts thus appear to be 'statistically significant' if we ignore the problem of misclassification. The odds ratios based on followup data, however, are much more dispersed, with all of the distributions straddling one; when misclassification errors are taken into account, the evidence that seatbelts reduce the risk of injury is no longer overwhelming. Simulated posterior means, 95% interval estimates and p-values for these odds ratios are shown in Table 7.6. The p-values are simply the proportions of simulated odds ratios exceeding one; they are appropriate for testing whether a given odds ratio is one, versus the one-sided alternative that it is less than one.

Because the police-report versions of belt use and injury are

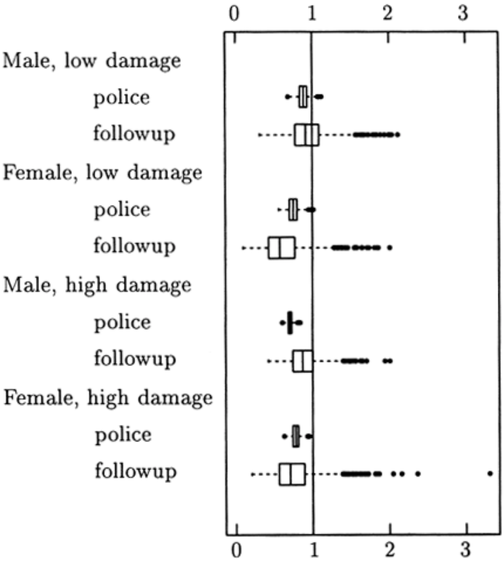


Figure 7.9. *Boxplots of 1000 simulated odds ratios showing the relationship of seatbelt use and injury within classes of damage and sex, both from police reports and from followup data.*

highly correlated with the followup versions, one might think that the rates of missing information for the followup variables should be much smaller than their actual missingness rates (98%). The fact that the followup-based intervals are so much wider than the police-based intervals, however, indicates that rates of missing information for these variables are still quite high. The main reason for this is the complexity of the saturated multinomial model. The saturated model allows for a full six-way association among the variables. The misclassification mechanism is described by the four-way table that relates the followup versions of belt-use and injury to the police versions. The saturated model estimates a full four-way association in this table; moreover, it allows the four-way association to vary freely across the four sex-by-damage cells. It is apparent that some of these high-order associations are poorly estimated, because the data in some parts of Table 7.5 are sparse. We will address this issue in Chapter 8 by applying models that are more parsimonious.

Table 7.6. *Simulated posterior means, 95% intervals and p-values for odds ratios from 1000 iterations of monotone data augmentation*

	mean	interval	p-value
Male, low damage			
police	0.89	(0.77, 1.03)	0.06
followup	0.95	(0.57, 1.59)	0.36
Female, low damage			
police	0.75	(0.62, 0.90)	0.00
followup	0.63	(0.24, 1.28)	0.09
Male, high damage			
police	0.70	(0.64, 0.77)	0.00
followup	0.89	(0.56, 1.35)	0.26
Female, high damage			
police	0.78	(0.68, 0.88)	0.00
followup	0.76	(0.37, 1.51)	0.17

CHAPTER 8

Loglinear Models

8.1 Introduction

In Chapter 7 we examined methods based on the saturated multinomial model. That model was quite general, allowing the associations among the categorical variables to be arbitrarily complex. In many realistic examples, however, unless the number of variables is very small, the observed data cannot support such complexity. This chapter presents methods for a flexible class of models which allows the associations among variables to be simplified.

Loglinear models have been used extensively, particularly in the social sciences, for almost two decades. In loglinear models, the cell probabilities for the cross-classified contingency table are decomposed into multiplicative effects for each variable and for the associations among them. Eliminating certain terms from this decomposition imposes equality constraints on odds ratios in the cross-classified table. A large part of this chapter is devoted to loglinear modeling with complete data, in particular, to the classical estimation technique of iterative proportional fitting (IPF) and a new simulation algorithm known as Bayesian IPF. These methods, which will be unfamiliar to many readers, are easily extended to calculate ML estimates and simulate posterior draws of parameters and missing data. Sections 8.3 and 8.4 concentrate on IPF and Bayesian IPF, respectively, and extensions to incomplete-data problems are presented in Section 8.5.

8.2 Overview of loglinear models

8.2.1 Definition

Suppose $x = (x_1, x_2, ..., x_D)$ is a contingency table having a multinomial distribution,

$$\chi \,|\, \theta \sim M(n, \theta), \tag{8.1}$$

where the cell probabilities $\theta = \left(\theta_1, \theta_2, ..., \theta_D\right)$ lie within the simplex

$$\Theta = \left\{\theta : \theta_d \; \geq \; \text{for all } d \text{ and } \Sigma_{d=1}^{D} \theta_d = 1\right\}.$$

A loglinear model does not alter the distributional assumption (8.1), but imposes further constraints on the elements of θ. Let

$$\eta_d = \log\theta_d, \;\; d{=}1, 2, ..., D$$

and

$$\eta = (\eta_1, \eta_2, \eta_D)^T .$$

In the most general sense, a loglinear model is any constraint of the form

$$\eta = M\lambda \tag{8.2}$$

where λ is an $r \times 1$ parameter vector and M is a fixed and known $D \times r$ design matrix. Thus, in addition to requiring that the elements of θ sum to one, we also require η=log θ to lie in the linear subspace spanned by the columns of M. The meaning of the elements of λ will depend on the coding method used in M. In typical applications of loglinear modeling, x represents a cross-classification of sample units by categorical variables $Y_1, Y_2, ..., Y_p$, and M is a design matrix of the type used in the analysis of variance (ANOVA) for factorial experiments; each variable Y_j represents a 'factor,' and the elements of λ represent the 'main effects' and 'interactions' associated with the factors.

Models for three categorical variables

For expository purposes, let us temporarily assume that there are only three categorical variables (p = 3). This assumption is purely a matter of convenience, and all results will immediately generalize to any number of variables. Also, we will temporarily switch to a notation more consistent with that of standard texts on categorical data (e.g. Agresti, 1990); in later sections, we will return to the notation developed in Chapter 7. Suppose we have three categorical variables:

A with levels i = 1, 2,..., I;
B with levels j = 1, 2,..., J;
C with levels k = 1, 2,..., K.

Let x_{ijk} denote the number of sample units for which we observe $A = i$, $B = j$, $C = k$. Let $\theta_{ijk} = P(A = i, B = j, C = k)$ and $\eta_{ijk} = \log \theta_{ijk}$. Finally, let '+' in place of a subscript denote summation over that subscript, as in

$$x_{+jk} = \sum_{i=1}^{I} x_{ijk} \text{ and } \theta_{i++} = \sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{ijk} .$$

The total sample size is n=x_{+++}.

As in a factorial ANOVA model, we can decompose n_{ijk} into additive terms corresponding to the 'main effects' and 'interactions' of$A$, B, and C,

$$\eta_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}, \qquad (8.3)$$

where for identifiably the λ terms are constrained to sum to zero over any subscript,

$$\sum_{i=1}^{I} \lambda_i^A = 0, \ \sum_{i=1}^{I} \lambda_{ij}^{AB} = \sum_{j=1}^{J} \lambda_{ij}^{AB} = 0, \tag{8.4}$$

and so on. To see how this relates to the general specification (8.2), consider the special case where $I = J = K = 2$; taking

$$\eta = \begin{bmatrix} \eta_{111} \\ \eta_{211} \\ \eta_{121} \\ \eta_{221} \\ \eta_{112} \\ \eta_{212} \\ \eta_{122} \\ \eta_{222} \end{bmatrix}, M = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix}$$

yields

$$\lambda = \left[\lambda_0, \lambda_1^A, \lambda_1^B, \lambda_1^C, \lambda_{11}^{AB}, \lambda_{11}^{AC}, \lambda_{11}^{BC}, \lambda_{111}^{ABC}\right]^T,$$

and the other λ terms follow from the identifiably constraints,

$$\lambda_1^A = -\lambda_2^A, \ \lambda_{11}^{AB} = -\lambda_{12}^{AB} = -\lambda_{21}^{AB} = \lambda_{22}^{AB},$$

and so on.

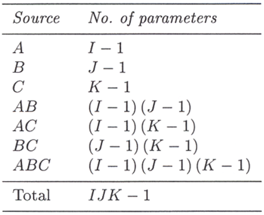
In many respects the loglinear model (8.3) is like the classical linear model for a factorial experiment. There are two important differences, however, that distinguish the loglinear model from its linear counterpart. First, the term λ_0, which appears to be like the ’grand mean,’ is not a free parameter but a normalizing constant chosen to make the cell probabilities sum to one,

$$\lambda_0 = -\log\left\{\Sigma_{ijk} \exp\left(\lambda_i^A + \lambda_j^B + \cdots + \lambda_{ijk}^{ABC}\right)\right\}.$$

Second, the linear equation (8.3) does riot represent the mean of a response variable given $A = i$, $B = j$, $C = k$; rather, it represents the log-probability of the event $A = i$, $B = j$, $C = k$ itself. The loglinear model is not a regression model describing the effects of A, B and C on an additional response variable, but a true multivariate model describing the relationships among the variables A, B and C. Thus the meaning of the λ terms is quite different from the usual interpretation of main effects and interactions in a linear model. For example, the set of terms $\lambda^{AB} = \left\{ \lambda_{ij}^{AB} \right\}$ describes the association between A and B, not their *interaction* with respect to a third variable. The terms in λ^{AB} are essentially the log-odds ratios describing the association between A and B, and the terms $\lambda^{ABC} = \left\{ \lambda_{ijk}^{ABC} \right\}$ are the differences in log-odds ratios describing how the association between any two variables varies across levels of the third. For details on the exact correspondence between the λ terms and log-odds ratios, see Bishop, Fienberg and Holland (1975) or Agresti (1990).

8.2.2 Eliminating associations

The number of free parameters in the loglinear model (8.3) can be counted in the same manner as the number of degrees of freedom in a factorial ANOVA.



Source	*No. of parameters*
A	$I-1$
B	$J-1$
C	$K-1$
AB	$(I-1)(J-1)$
AC	$(I-1)(K-1)$
BC	$(J-1)(K-1)$
ABC	$(I-1)(J-1)(K-1)$
Total	$IJK-1$

Notice that the total number of free parameters in (8.3), IJK-1, is the same as in the saturated multinomial; hence (8.3) is nothing more than a re-parameterization of the saturated model, with the cell probabilities θ re-expressed in terms of the loglinear coefficients λ. The loglinear representation has a great advantage, however, in that it allows us to selectively eliminate associations among variables by setting groups of λ terms to zero. Suppose we set all the terms in $\lambda^{AB} = \left\{ \lambda_{ij}^{AB} \right\}, \lambda^{AC} = \left\{ \lambda_{ik}^{AC} \right\}, \lambda^{BC} = \left\{ \lambda_{jk}^{BC} \right\},$ and

$\lambda^{ABC} = \left\{\lambda_{ijk}^{ABC}\right\}$ to zero. The loglinear model can then be written as

$$\theta_{ijk} \propto \exp\left(\lambda_i^A + \lambda_j^B + \lambda_k^C\right),$$

which implies that A, B and C are mutually independent,

$$\theta_{ijk} = P(A = i)P(B = j)P(C = k)$$

Setting $\lambda^{BC} = \lambda^{AC} = \lambda^{ABC} = 0$ but allowing λ^{AB} to vary leads to

$$\theta_{ijk} = P(A = i, B = j)P(C = k),$$

which means that A and B may be related but requires them to be jointly independent of C. Setting $\lambda^{AB} = \lambda^{ABC} = 0$ gives

$$\theta_{ijk} = P(A = i \mid C = k)P(B = j \mid C = k)P(C = k),$$

which means that A and B are conditionally independent given C. Finally, setting λ^{ABC}=0 results in a model of *homogeneous association*. This model does not imply any form of independence or conditional independence, but has the property that the association between any two variables (in terms of odds ratios) is constant across levels of the third.

Hierarchical models

In most applications of loglinear modeling, it would not make sense to specify a model that contains an association but omits a main effect. A model that includes λ^{AB} but omits λ^{A} allows A to be related to B, but requires the average log-probability across levels of B to be the same within every level of A. Under ordinary circumstances one would not expect this to happen except by chance. Similarly, it would rarely make sense to fit a model that contains the three-way association λ^{ABC} but omits one or more of the two-way associations λ^{AB}, λ^{AC} or λ^{BC}.

A loglinear model is said to be hierarchical if omitting a λ term implies that all higher-order associations containing that term are omitted as well; for example, if setting λ^{AB}=0 requires that we also set $\lambda^{ABC} = 0$. Putting it another way, a model is hierarchical if no association is present unless all lower-order terms within that association are also present. Thus a hierarchical model containing λ^{ABC} must also contain λ^{A}, λ^{B}, λ^{C}, λ^{AB}, λ^{AC} and λ^{BC}. The class of hierarchical models includes models of independence and conditional independence, as well as some other models that may be of interest, e.g. the model of homogeneous association in a three-way table (λ^{ABC}=0). Non-hierarchical models, however, rarely correspond to sensible hypotheses about the underlying categorical variables. For the remainder of this book, we will restrict our attention to hierarchical models only.

8.2.3 Sufficient statistics

The loglikelihood for the saturated model in terms of the cell probabilities θ is

$$l(\theta \mid x) = \Sigma_{ijk} x_{ijk} \log \theta_{ijk}.$$

When expressed in terms of the loglinear coefficients, the loglikelihood becomes

$$\begin{aligned} l(\lambda \mid x) &= \Sigma_{ijk} \quad x_{ijk}\left(\lambda_0 + \lambda_i^A + \cdots + \lambda_{ijk}^{ABC}\right) \\ &= n\lambda_0 \quad + \Sigma_i x_{i++} \lambda_i^A + \Sigma_j x + j + \lambda_j^B \\ &\qquad + \Sigma_k x_{++k} \lambda_k^C + \Sigma_{ij} x_{ij} + \lambda_{ij}^{AB} + \Sigma_{ik} x_{i+k} \lambda_{ik}^{AC} \\ &\qquad + \Sigma_{jk} x_{+jk} \lambda_{jk}^{BC} + \Sigma_{ijk} x_{ijk} \lambda_{ijk}^{ABC} \end{aligned}$$

We will use $x^A = \{x_{i++}\}, x^{AB} = \{x_{ij+}\}, x^{ABC} = \{x_{ijk}\}$ and so on to denote the marginal frequencies that result when units are cross-classified by subsets of variables. Following Bishop, Fienberg and Holland (1975), we will call these marginal tables *configurations*. Because the configurations x^A, x^B, x^C,

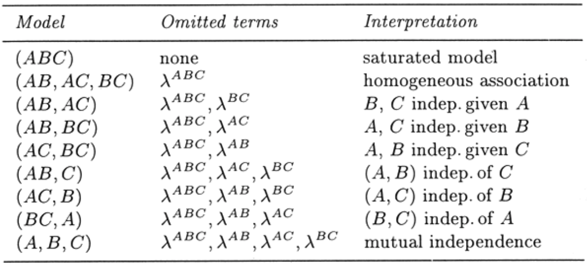
x^{AB}, x^{AC}, and x^{BC} can be obtained by summing the elements of x^{ABC} the loglikelihood for the saturated model is a linear function of the configuration x^{ABC}.

If we simplify the model by eliminating some of the λ terms, the corresponding configurations drop out of the loglikelihood. For example, if we set $\lambda^{BC} = \lambda^{ABC} = 0$ the loglikelihood becomes

$$\begin{aligned} \ell(\lambda \mid x) = \quad & n\lambda_0 + \Sigma_i x_{i++}\lambda_i^A + \Sigma_j x_{+j+}\lambda_j^B \\ & + \Sigma_k x_{++k}\lambda_k^C + \Sigma_{ij} x_{ij+}\lambda_{ij}^{AB} + \Sigma_{ik} x_{i+k}\lambda_{ik}^{AC} \end{aligned}$$

Because x^A, x^B and x^C follow from x^{AB} and x^{AC}, the latter two configurations constitute a minimal set of sufficient statistics for this model. If a model is hierarchical, then the configuration for any set of variables present in the model can be derived from the highest-order configuration containing that set. Consequently, the configurations for these highest-order terms form a minimal set of sufficient statistics. We will call these the sufficient configurations. It has become standard practice to identify loglinear models by their sufficient configurations. For example, the model

Table 8.1. *Hierarchical loglinear models for three categorical variables*



Model	*Omitted terms*	*Interpretation*
(ABC)	none	saturated model
(AB, AC, BC)	λ^{ABC}	homogeneous association
(AB, AC)	$\lambda^{ABC}, \lambda^{BC}$	B, C indep. given A
(AB, BC)	$\lambda^{ABC}, \lambda^{AC}$	A, C indep. given B
(AC, BC)	$\lambda^{ABC}, \lambda^{AB}$	A, B indep. given C
(AB, C)	$\lambda^{ABC}, \lambda^{AC}, \lambda^{BC}$	(A, B) indep. of C
(AC, B)	$\lambda^{ABC}, \lambda^{AB}, \lambda^{BC}$	(A, C) indep. of B
(BC, A)	$\lambda^{ABC}, \lambda^{AB}, \lambda^{AC}$	(B, C) indep. of A
(A, B, C)	$\lambda^{ABC}, \lambda^{AB}, \lambda^{AC}, \lambda^{BC}$	mutual independence

λ^{ABC}=0 can be denoted by (x^{AB}, x^{AC}, x^{BC}) or, more simply, (*AB*, *AC*, *BC*).

8.2.4 Model interpretation

The hierarchical models that can be fitted to a three-variable dataset are listed in Table 8.1, along with their sufficient configurations. This list does not include any model that omits one or more of the main effects λ^A, λ^B, λ^C. Setting a main effect to zero is equivalent to saying that the marginal distribution of the corresponding variable is uniform across its levels, a hypothesis which is rarely of interest.

Models in four or more dimensions are interpreted in a similar fashion. For example: (AB, CD) means that *A* and *B* are jointly independent of *C* and *D*; (ABC, BCD) means that *A* and *D* are conditionally independent given (*B*, *C*); (AB, AC, AD, BC, BD, CD) means that the odds ratios for any two variables are constant across levels of the other two; and (ABC, ABD, ACD, BCD) means that the associations among any three variables are constant across levels of the fourth.

Correspondence to logit models

If one of the variables is regarded as a response and the others are regarded as potential predictors, then certain loglinear models are equivalent to standard logistic regression or logit models (Goodman, 1970). Consider the saturated model for *A*, *B* and *C*, where *C* is a binary variable considered to be a response, and let

$$\pi_{ij} = P(C = 1 \mid A = i, B = j).$$

The logit model for predicting the probability of $C = 1$ from *A* and *B* can be written as

$$\begin{aligned}\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \log\left(\frac{\theta_{ij1} \mid \theta_{ij+}}{\theta_{ij2} \mid \theta_{ij+}}\right), \\ &= \eta_{ij1} - \eta_{ij2},\end{aligned}$$

where $n_{ijk} = \log\theta_{ijk}$. But notice that

$$\begin{aligned}\eta_{ij1}-\eta_{ij2} &= \left(\lambda_1^C-\lambda_2^C\right)+\left(\lambda_{i1}^{AC}-\lambda_{i2}^{AC}\right)\\ &\quad+\left(\lambda_{j1}^{BC}-\lambda_{j2}^{BC}\right)+\left(\lambda_{ij1}^{ABC}-\lambda_{ij2}^{ABC}\right)\\ &= 2\lambda_1^C+2\lambda_{i1}^{AC}+2\lambda_{j1}^{BC}+2\lambda_{ij1}^{ABC}\end{aligned}$$

so this logit model is of the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\beta_0+\beta_i^A+\beta_j^B+\beta_{ij}^{AB}, \qquad (8.5)$$

where the coefficients satisfy

$$\Sigma_i\beta_i^A=\Sigma_j\beta_j^B=\Sigma_i\beta_{ij}^{AB}=\Sigma_j\beta_{ij}^{AB}=0.$$

Thus, the saturated model (ABC) implies a standard logit model for *C* that includes main effects for *A* and *B* as well as the AB interaction.

Notice that if we set $\lambda^{ABC}=0$ in the loglinear model, (8.5) becomes

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\beta_0+\beta_i^A+\beta_j^B$$

a logit model with main effects only. Setting $\lambda^{ABC}=\lambda^{AC}=0$ and $\lambda^{ABC}=\lambda^{BC}=0$ removes the effects of *A* and *B*, respectively, and $\lambda^{ABC}=\lambda^{AC}=\lambda^{BC}=0$ produces the null model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\beta_0$$

Omitting λ^{AB}, however, would require *A* and *B* to be conditionally independent, an assumption not found in the standard logit model. The standard logit model, like other regression models, makes no assumptions about the predictors; it allows their joint distribution to be arbitrary.

The relationships between loglinear models and logit models for three categorical variables are summarized in Table 8.2. In general,

Table 8.2. *Loglinear and corresponding logit models for three categorical variables*

Model	*Implied logit model for C*
(ABC)	A, B main effects and AB interaction
(AB, AC, BC)	A, B main effects only
(AB, AC)	A main effect
(AB, BC)	B main effect
(AB, C)	null model

a loglinear model is equivalent to a standard logit model provided that it includes all potential associations among the variables considered to be predictors. A two-way association between a response and a predictor in the loglinear model introduces a main effect for that predictor in the logit model; a three-way association between a response and two predictors introduces an interaction between the two predictors; and so on. The response variable need not be binary; if it has more than two categories, the loglinear model implies a generalized logit model for an unordered multinomial response (e.g. Agresti, 1990).

8.3 Likelihood-based inference with complete data

8.3.1 Maximum-likelihood estimation

To derive the ML estimates for a loglinear model, we could try to differentiate the loglikelihood with respect to some set of free parameters and set the resulting expressions to zero. However, as demonstrated by Birch (1963), we may also apply the method that leads to ML estimates for any regular exponential family model: solve the system of equations that results when the minimal sufficient statistics are set equal to their expectations (Section 3.2.1). In many cases, this system can be solved immediately to yield the ML estimates for the cell probabilities θ.

For example, consider the saturated model (*ABC*). Setting the elements of the sufficient configuration x^{ABC} equal to their expectations $E\left(x_{ijk} \mid \theta\right) = n\theta_{ijk}$ produces

$$\hat{\theta}_{ijk} = x_{ijk} / n$$

for all *i*, *j* and *k*. For the model (*AB*, *C*), the moment equations are

$$x_{ij+} = n\hat{\theta}_{ij+},$$

$$x_{++k} = n\hat{\theta}_{++k}.$$

But because this model implies $\theta_{ijk} = \theta_{ij+}\theta_{++k}$, we obtain

$$\hat{\theta}_{ijk} = \left(x_{ij+}x_{++k}\right) / n^2 .$$

The model (*AB*, *BC*) gives

$$x_{ij+} = n\hat{\theta}_{ij+},$$

$$x_{+jk} = n\hat{\theta}_{+jk},$$

and because this model implies

$$\theta_{ijk} = \theta_{+j+}\left(\frac{\theta_{ij+}}{\theta_{+j+}}\right)\left(\frac{\theta_{+jk}}{\theta_{+j+}}\right),$$

The model (*AB*, *AC*, *AC*), however, produces

$$
\begin{aligned}
x_{ij+} &= n\hat{\theta}_{ij+},\\
x_{i+k} &= n\hat{\theta}_{i+k},\\
x_{+jk} &= n\hat{\theta}_{+jk},
\end{aligned}
$$

a system for which there is no closed-form solution.

It turns out that (AB, AC, BC) is the only hierarchical model in three dimensions for which the ML estimates cannot be written in closed form. In four or more dimensions, however, there are many more models for which this is so. When the moment equations do not yield explicit ML estimates for θ, they may be solved numerically by the method of iterative proportional fitting.

8.3.2 Iterative proportional fitting

Iterative proportional fitting (IPF) is a simple and intuitive method for solving the moment equations. First, start with an arbitrary value of θ that satisfies the loglinear constraints, typically a uniform table (all cell probabilities equal). Then proportionately adjust the elements of θ to satisfy the moment equations for a single configuration. Do this for each sufficient configuration in turn, and repeat the entire process until the elements of θ stabilize.

For example, consider the model (AB, AC, BC). Given the current estimate $\theta^{(t)}$, IPF updates it as follows:

$$\theta_{ijk}^{(t+1/3)} = \theta_{ijk}^{(t+0/3)} \left(\frac{x_{ij+}/n}{\theta_{ij+}^{(t+0/3)}} \right) \text{ for all } i, j, k; \tag{8.6}$$

$$\theta_{ijk}^{(t+2/3)} = \theta_{ijk}^{(t+1/3)} \left(\frac{x_{i+k}/n}{\theta_{i+k}^{(t+1/3)}} \right) \text{ for all } i, j, k; \tag{8.7}$$

$$\theta_{ijk}^{(t+3/3)} = \theta_{ijk}^{(t+2/3)} \left(\frac{x_{+jk} / n}{\theta_{+jk}^{(t+2/3)}} \right) \text{ for all } i, j, k. \qquad (8.8)$$

Notice that $\theta^{(\tau+1/3)}$ satisfies the required conditions for x^{AB} but not necessarily those for x^{AC} or x^{BC}. Similarly, $\theta^{(\tau+2/3)}$ satisfies those for x^{AC} but not necessarily for x^{AB} or x^{BC}, and $\theta^{(\tau+3/3)}$ satisfies those for x^{BC} but not necessarily x^{AB} or x^{BC}. Repeating the cycle (8.6)-(8.8) produces a sequence $\theta^{(1)}\ \theta^{(2)}$ which converges to a value $\theta^{(\infty)} = \hat{\theta}$ satisfying all three sets of moment equations simultaneously; this is the unique ML estimate of θ.

This IPF algorithm immediately generalizes to loglinear models of any dimension. When the moment equations can be solved in closed form, IPF may still be used, and it typically converges to the correct ML estimates in a single cycle. In other cases, IPF exhibits linear convergence near the mode. Proofs of the convergence of IPF are given by Bishop, Fienberg and Holland (1975) and their references. Because IPF operates on the cell probabilities θ, it does not automatically yield explicit estimates of the loglinear coefficients λ. Estimates of λ may be obtained in a variety of ways. One simple way is to define a full-rank design matrix M such that log logθ=$M\lambda$, and calculate the ordinary least-squares estimates

$$\hat{\lambda} = \left(M^T M \right)^{-1} M^T \log \hat{\theta}$$

using standard regression software. Except for rounding errors, the regression model will have perfect fit because $\log \hat{\theta}$ is required to lie in the space spanned by the columns of M. Another way to obtain the elements of λ is to express them as linear contrasts of the elements of logθ; see Bishop, Fienberg and Holland (1975) for details.

Comparison with other methods

IPF has been in existence for more than half a century. Deming and Stephan (1940) discussed *raking*, a method of proportionately adjusting survey data to make the observed distributions of certain variables agree with census totals; their algorithm is essentially equivalent to IPF. Early work on loglinear modeling (e.g. Bishop, Fienberg and Holland, 1975) relied almost exclusively on IPF, but more recent books emphasize other methods. Software packages capable of fitting loglinear models typically use Newton-Raphson or Fisher scoring (e.g. Agresti, 1990). These methods tend to be quicker than IPF because their convergence behavior is quadratic; moreover, they provide asymptotic standard errors as an automatic by-product, because they make use of the loglikelihood function's second derivatives. Standard errors can be obtained with IPF, but computing them requires additional formulas that are not an integral part of the fitting algorithm. Yet IPF maintains some advantages because of its simplicity and computational stability. In this text, we focus on IPF because of its intimate relationships to ECM and the simulation algorithms described later in this chapter.

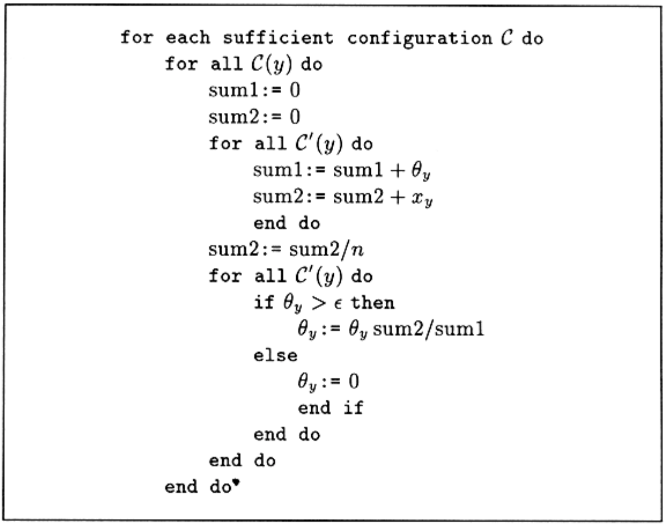
Random and structural zeroes

If the contingency table contains random zeroes, $\hat{\theta}$ may lie on the boundary of its parameter space with estimated probabilities of zero in one or more cells. When this occurs, some of the IPF equations may be undefined at $\hat{\theta}$ because they may involve division by zero. This difficulty is easily overcome by the following modification: if any probability falls below a small positive constant c, set it to zero and omit that cell from further iterations.

Structural zeroes, whose cell probabilities are taken to be zero a priori, can also be handled quite easily. The usual way of handling them is to omit them from the model and assume that the loglinear specification (8.2) holds for the remaining cells. With IPF, we simply choose a starting value of $\hat{\theta}$ that has zeroes in the structural-zero cells and uniform values elsewhere. Because $0 < \in$, the estimated probabilities for these cells will remain at zero for all iterations.

An implementation in pseudocode

Returning now to the general notation of Chapter 7, suppose that Y_1, $Y_2, \ldots, Y_p$ are categorical variables recorded for a sample of n units, where Y_j takes values 1, 2,..., d_j. Let $y = (y_1, \ldots, y_p)$ denote a generic realization of $(Y_1, \ldots, Y_p)$; let $\mathcal{Y}$ be the set of all possible values of y; and let x_y and θ_y be the frequency and cell probability, respectively, associated with cell y of the p-dimensional cross-classified table. Let C be a subset of $\{1, 2, \ldots, p\}$ identifying a generic configuration; for example, $C = \{1, 3\}$ indicates the configuration Y_1Y_3. For brevity, we will also use C to denote the function that extracts from $y = (y_1, \ldots, y_p)$ the elements corresponding to the configuration C, for example, if $C = \{2, 4\}$ then $C = (y_2, y_4)$. Finally, a generic marginal count



```
for each sufficient configuration $C$ do
    for all $C(y)$ do
        sum1:= 0
        sum2:= 0
        for all $C'(y)$ do
            sum1:= sum1 $+ \theta_y$
            sum2:= sum2 $+ x_y$
            end do
        sum2:= sum2$/n$
        for all $C'(y)$ do
            if $\theta_y > \epsilon$ then
                $\theta_y$:= $\theta_y$ sum2/sum1
            else
                $\theta_y$:= 0
                end if
            end do
        end do
    end do
```

Figure 8.1. *Single cycle of iterative proportional fitting.*

within C and its corresponding marginal probability will be denoted by

$$x_{C(y)} = \sum_{C'(y)} x_y \text{ and } \theta_{C(y)} = \sum_{C'(y)} \theta_y,$$

respectively, where C is the complement of C.

Pseudocode for a general implementation of IPF is shown in Figure 8.1. Given the observed counts in the workspace x and the current parameter $\theta^{(t)}$ in θ, this code performs one cycle of IPF, overwriting $\theta^{(t)}$ with the updated cell means $n\theta^{(t+1)}$. Prior to execution, the values stored in θ need not sum to one; we will obtain the same result if this workspace contains $c\ \theta^{(t)}$ for any constant c. Therefore, the updated cell means $n\theta^{(t+1)}$ from one cycle may be used directly as input to the next cycle; there is no need to rescale them to sum to one at each cycle.

8.3.3 Hypothesis testing and goodness of fit

A loglinear model's quality of fit may be assessed by its deviance. The deviance, typically denoted by G_2, is the likelihood-ratio statistic for testing the current model against the alternative of a saturated model,

$$G^2 = 2 \sum_{y \in \mathcal{Y}} x_y \log \frac{x_y}{n\hat{\theta}_y} \tag{8.9}$$

where $\hat{\theta} = \{\theta_y : y \in \mathcal{Y}^*\}$ is the maximizer of the likelihood under the current model, and $\mathcal{Y}^*$ is the set of all cells excluding structural zeroes. The deviance is asymptotically equivalent to the well-known goodness-of-fit statistic due to Pearson,

$$X^2 = \sum_{y \in \mathcal{Y}} \frac{\left(x_y - n\hat{\theta}_y\right)^2}{n\hat{\theta}_y} \tag{8.10}$$

If the sample size is sufficiently large, G_2 and X_2 are distributed approximately as x^2_{df} under the null hypothesis that the current model is true, where df is equal to the difference in the number of free parameters in the saturated and current models. An approximate p-value for testing the current model against a general alternative is thus $P\left(x^2_{df} \geq G^2\right)$ or $P\left(x^2_{df} \geq X^2\right)$. The chisquare approximation for these goodness-of-fit tests is traditionally regarded as accurate if $n\hat{\theta}_y \geq 5$ for all $y \in \mathcal{Y}^*$; in addition, some empirical studies have shown that it may be reasonably accurate if a small proportion of the cells have $n\hat{\theta}_y$ as small as 2 or even 1; see Agresti (1990, pp. 246-247) for further details.

The G^2 and X^2 statistics also provide a basis for general comparisons between models that are nested. Suppose we want to test the null hypothesis that Model *A* is true against the alternative that Model *B* is true, where Model *A* is a special case of Model *B*. Then

$$\Delta G^2 = G^2 \text{ for Model A — } G^2 \text{ for Model B}$$

and

$$\Delta X^2 = X^2 \text{ for Model A — } X^2 \text{ for Model B}$$

are distributed approximately as chisquare with degrees of freedom equal to

$$\Delta df = df \text{ for Model A — } df \text{ for Model B}$$

Large values of ΔG^2 or ΔX^2 indicate that Model *B* fits the data substantially better than Model *A*. Asymptotic arguments suggest that the chisquare approximations for ΔG^2 and ΔX^2

may be quite accurate even when the approximations for the individual goodness-of-fit statistics for Model A and Model B are poor. Consequently, for ΔG^2 and ΔX^2 can be useful even with sparse tables, provided that (a) the number of observations is large relative to Δdf, and (b) the observed frequencies are of approximately the same order of magnitude (Haberman, 1977).

Effects of random zeroes and boundary estimates

Notice that the presence of a random zero ($x_y = 0$) will cause G_2 to be undefined. This problem can be overcome by taking 0 log 0 to be 0. When a pattern of random zeroes causes the ML estimate to lie on the boundary ($\theta_y = 0$ for some y), however, neither G_2 nor X_2 can be calculated directly. In these situations, it is customary to omit the cells with zero estimates from consideration and adjust the degrees of freedom to reflect the fact that some parameters may not be estimable. Rules for adjusting the degrees of freedom (e.g. Bishop, Fienberg and Holland, 1975) are quite complicated and are difficult to implement in general-purpose computer code. Users of general-purpose software for loglinear modeling should be wary of these adjustments, because situations exist for which nearly every popular software package gives misleading results (Clogg *et al.*, 1991). When sparseness in the data table x leads to boundary estimates, a simpler and more reliable procedure is to introduce a small amount of prior information to smooth the data and move $\hat{\theta}$ away from the boundary; see Clogg *et al.* (1991) and Section 8.4.2 below.

8.3.4 Example: misclassification of seatbelt use and injury

Recall the data of Table 7.5 from a followup study on misclassification error in seatbelt use and injury in automobile accidents. Hochberg (1977) and Chen (1989) investigated loglinear models for the six dichotomous variables:

Code	Variable
D	car damage (1=low, 2=high)
S	driver's sex (1=male, 2=female)
B_1	belt use, police report (1=no, 2=yes)

I_1	injury, police report (1=no, 2=yes)
B_2	belt use, followup study (1=no, 2=yes)
I_2	injury, followup study (1=no, 2=yes)

Table 8.3. *Goodness-of-fit statistics for nine loglinear models fitted to the automobile accident followup data*

Model	G^2	X^2	*df*
1. (DSB_2I_2, E_BE_I)	1056.46	1726.43	45
2. $(DSB_2I_2, B_2I_2E_BE_I)$	64.59	62.14	36
3. $(DSB_2I_2, B_2I_2E_BE_I, DE_B)$	60.40	57.09	35
4. $(DSB_2I_2, B_2I_2E_BE_I, DE_I)$	57.51	57.67	35
5. $(DSB_2I_2, B_2I_2E_BE_I, DE_B, DE_I)$	53.99	53.47	34
6. $(DSB_2I_2, B_2I_2E_BE_I, DE_BE_I)$	53.05	51.99	33
7. $(DSB_2I_2, B_2I_2E_BE_I, DE_B, DE_I, SE_B)$	53.90	53.03	33
8. $(DSB_2I_2, B_2I_2E_BE_I, DE_B, DE_I, SE_I)$	52.48	51.29	33
9. $(DSB_2I_2, B_2I_2E_BE_I, DSE_I)$	52.37	51.07	32

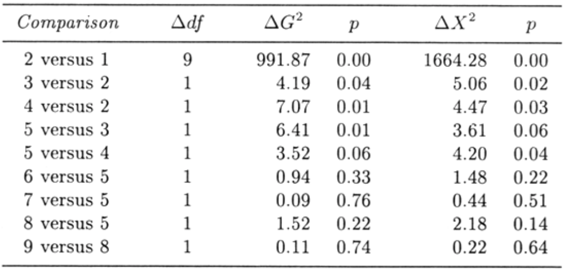
Instead of working with these six variables directly, let us consider models for D, S, B_2, I_2 and the two error indicators

E_B = 1 if $B_1 = B_2$, *0* otherwise;
E_1 = 1 if $I_1 = I_2$, *0* otherwise.

Working with D, S, B_2, I_2, E_B and E_I rather than the six original variables may result in a model for the misclassification mechanism that is easier to interpret, because the associations between the underlying 'true' state of an accident DSB_2I_2 and the error indicators E_BE_{I} may be somewhat simpler than the associations between DSB_2I_2 and the police report B_1I_1. Regarding E_B and E_I as response variables and D, S, B_2, and I_2 as potential predictors, perhaps the simplest loglinear model worth considering is (DSB_2I_2, E_BE_I), which states that the bivariate response is unrelated to the predictors. Goodness-of-fit statistics for this model and eight other loglinear models are shown in Table 8.3.

None of the nine models shown in Table 8.3 produced ML estimates on the boundary of the parameter space, but all of them had estimated expected counts falling below 1.0 for some cells; p-values based on the chisquare approximations for G_2 and X_2 are not shown because they are not trustworthy. Chisquare approximations for comparisons among these models are probably more accurate, however, and results from hypothesis tests for various pairs of nested models are shown in Table 8.4. Model 1, the null model of no relationships between the response and predictors, appears to fit the data very poorly. The fit improves dramatically when the actual belt use/injury status B_2I_2 is allowed to influence the response (Model 2). In addition, the data provide fairly strong evidence for the associations D_{EB}, DE_I, and perhaps SE_I. Among these nine, the simplest models that capture the essential relationships between the response and the predictors appear to be Models 5 and 8.

Table 8.4. *Hypothesis tests for various pairs of nested models*



Comparison	Δdf	ΔG^2	p	ΔX^2	p
2 versus 1	9	991.87	0.00	1664.28	0.00
3 versus 2	1	4.19	0.04	5.06	0.02
4 versus 2	1	7.07	0.01	4.47	0.03
5 versus 3	1	6.41	0.01	3.61	0.06
5 versus 4	1	3.52	0.06	4.20	0.04
6 versus 5	1	0.94	0.33	1.48	0.22
7 versus 5	1	0.09	0.76	0.44	0.51
8 versus 5	1	1.52	0.22	2.18	0.14
9 versus 8	1	0.11	0.74	0.22	0.64

8.4 Bayesian inference with complete data

8.4.1 Prior distributions for loglinear models

In our work with the saturated multinomial model in the last chapter, we adopted the simple Dirichlet prior distribution $\theta \sim D(\alpha)$, with the elements of a typically chosen to be equal. This prior is 'naive' in the sense that it treats the cell probabilities in an unordered fashion, i.e. it does not describe the special structure that exists in a cross-classified contingency table. This was not regarded as a serious drawback, because the saturated multinomial model does not make use of this cross-classified structure either. The fundamental quality of loglinear models, however, is that they take this structure into account.

Many alternative types of prior distributions have been proposed to sensibly incorporate prior information about the structure of a loglinear model. Bishop, Fienberg and Holland

(1975) decomposed the Dirichlet hyperparameters a into ’main effects’ and ’associations’ in a loglinear fashion. Good (1967) proposed a second-stage prior distribution on a, resulting in a mixture of Dirichlet priors that can potentially reflect a cross-classified structure. Several authors have applied normal prior distributions to the loglinear coefficients λ; variations of this approach are discussed by Good (1956), Leonard (1975), Laird (1978), and Knuiman and Speed (1988). The normal priors, although conceptually attractive, lead to nonnormal posteriors that are computationally more difficult to handle than the Dirichlet.

The constrained Dirichlet prior

For our purposes, it will be convenient to adopt a prior distribution that has the same functional form as the Dirichlet, but which requires the parameters to satisfy the constraints imposed by a loglinear model. Let M denote the design matrix for a loglinear model

$$\log \theta = M\lambda \tag{8.11}$$

and let Θ_M denote the set of all parameters $\theta = \{\theta_y : y \in \mathcal{Y}\}$ that lie in the simplex and satisfy (8.11) for some λ. Let us take the prior density for θ to be

$$\pi(\theta) \propto \prod_{y \in \mathcal{Y}} \theta_y^{\alpha_y - 1}$$

for $\theta \in \Theta_M$ and zero elsewhere. We will call this the constrained Dirichlet prior with hyperparameter $\alpha = \{\alpha_y : y \in \mathcal{Y}\}$. This is not a Dirichlet distribution per se, but the conditional distribution of $\theta \sim D(\alpha)$ given that the event $\theta \in \Theta_M$ has occurred. The normalizing constant

$$\int_{\Theta_M} \prod_{y \in \mathcal{Y}} \theta_y^{\alpha_y - 1} d\theta$$

is not generally tractable, but this will not be a problem because in the algorithms to follow this integral will not be explicitly evaluated.

The advantage of the constrained Dirichlet prior is that it retains the same functional form as the multinomial likelihood and thus forms a conjugate class; the posterior distribution of θ given the contingency table x is another constrained Dirichlet with updated hyperparameters $\alpha' = \alpha + x$. A potential disadvantage is that this prior makes the strong assumption that the given loglinear model is true; it assigns zero probability to values of θ not satisfying (8.11). Just as in likelihood-based inference, however, it will be possible to examine the adequacy of a model by performing goodness-of-fit tests against alternative models that are more general.

8.4.2 Inference using posterior modes

Under the constrained Dirichlet prior, the complete-data posterior density for θ is

$$P(\theta \mid x) \propto \prod_{y \in \mathcal{Y}} \theta_y^{x_y + \alpha_y - 1} \tag{8.12}$$

for $\theta \in \Theta_M$ and zero elsewhere. Notice that this is equivalent to the likelihood function for θ given a modified contingency table with cell counts $x'_y = x_y + \alpha_y - 1$. Any algorithm that computes ML estimates for loglinear models can thus be trivially modified to find posterior modes for θ as well; all we need to do is to augment each cell count x_y by the amount $\alpha_y - 1$. In particular, the IPF algorithm of Section 8.3.2 will find posterior modes if we simply replace each x_y by $x'_y = x_y + \alpha_y - 1$.

It might seem natural to call this modified IPF algorithm ’Bayesian IPF.’ However, we will reserve that name for another algorithm, to be discussed shortly, for simulating random draws from a constrained Dirichlet distribution.

Notice that the posterior mode for θ is identical to the ML estimate under the uniform prior $\alpha = (1, 1,..., 1)$. A prior $\alpha = (c, c,..., c)$ for some $c > 1$ has a 'flattening' effect; the posterior mode under this prior will represent a compromise between the ML estimate and a uniform table in which all cell probabilities are equal (Section 7.2.5). Flattening priors can be especially useful for ensuring that the mode lies within the interior of the parameter space Θ_M, avoiding complications that arise with sparse data when ML estimates lie on the boundary.

Posterior modes and goodness of fit

The goodness-of-fit statistics G_2 and X_2 defined in Section 8.3.3 can be used with posterior modes. One possibility is to use the same expressions (8.9)-(8. 10) and simply replace the ML estimate $\hat{\theta}$ with a posterior mode. In many situations, however, it will be more natural to work with statistics that are based on the augmented cell counts $x'_y = x_y + \alpha_y - 1$. For a given loglinear model, let $\tilde{\theta} = \left\{\tilde{\theta}_y : y \in \mathcal{Y}^*\right\}$ represent the posterior mode under the constrained Dirichlet prior with hyperparameter α where $\mathcal{Y}^*$ is the set of all cells excluding structural zeroes. The posterior mode under the saturated model and the unrestricted $D(\alpha)$ prior occurs at

$$\theta_y = \frac{x'_y}{n'}, \quad n' = \sum_{y \in \mathcal{Y}^*} x'_y . \qquad (8.13)$$

The statistic

$$G^2 = 2 \sum_{y \in \mathcal{Y}^*} x'_y \log \frac{x'_y}{n' \hat{\theta}_y} \qquad (8.14)$$

represents twice the increase in the log-posterior density as we move from the mode under the current model to the mode under the saturated model. Like its likelihood-based counterpart, the modified statistic (8.14) is approximately distributed as χ^2_{df} over repeated samples, where df is the difference in the number of free parameters under the two models. This result holds because as the sample size grows, the cell counts x_y become appreciably larger than the hyperparameters α_y, and the influence of the prior becomes negligible. The analogue of (8.14) corresponding to Pearson's goodness-of-fit statistic is

$$X^2 = \sum_{y \in \mathcal{Y}^*} \frac{\left(x'_y - n'\tilde{\theta}_y\right)^2}{n'\tilde{\theta}_y} \tag{8.15}$$

One attractive feature of (8.14) and (8.15) is that these two statistics are easy to calculate; they will be generated automatically by standard software for loglinear modeling if the cell counts x_y are replaced by x'_y. Moreover, if all the hyperparameters α_y are greater than one, both $\hat{\theta}$ and the unrestricted mode (8.13) are guaranteed to lie in the interior of the parameter space, and there is no need to worry about adjusting df for estimates on the boundary (Clogg *et al.*, 1991).

8.4.3 Inference by Bayesian IPF

Here we present a clever but still relatively unknown technique for simulating random draws from a constrained Dirichlet posterior (8.12). This iterative method, first presented by Gelman *et al.* (1995), bears a striking resemblance to iterative proportional fitting and has thus been named *Bayesian IPF*. For simplicity, we describe Bayesian IPF for three categorical variables A, B and C under the model of homogeneous association (AB, AC, BC). Following the notation of Section 8.2, let x_{ijk}, θ_{ijk} and α_{ijk} denote the observed frequency, probability and prior hyperparameter, respectively, corresponding to the event ($A = i$, $B = j$, $C = k$). Let $\theta^{(t)}$ denote the simulated value of the parameter $\theta = \left\{\theta_{ijk}\right\}$ at cycle t. Bayesian IPF updates the parameter in three steps: first,

$$\theta_{ijk}^{(t+1/3)} = \theta_{ijk}^{(t+0/3)} \left(\frac{g_{ij+} / g_{+++}}{\theta_{ij+}^{(t+0/3)}} \right) \text{ for all } i, j, k, \qquad (8.16)$$

where the g_{ij+} are independent random variates drawn from standard gamma distributions with shape parameters

$$\alpha'_{ij+} = \sum_k \left(\alpha_{ijk} + x_{ijk} \right)$$

(Section 7.2.3), and $g_{+++} = \Sigma_{ij} g_{ij+}$ is their sum; second,

$$\theta_{ijk}^{(t+2/3)} = \theta_{ijk}^{(t+1/3)} \left(\frac{g_{i+k} / g_{+++}}{\theta_{i+k}^{(t+1/3)}} \right) \text{ for all } i, j, k, \qquad (8.17)$$

where the g_{i+k} are standard gamma variates with shape parameters

$$\alpha'_{i+k} = \sum_j \left(\alpha_{ijk} + x_{ijk} \right)$$

drawn independently of those in (8.16), and $g_{+++} = \Sigma_{ik} g_{i+k}$ is the new sum; and third,

$$\theta_{ijk}^{(t+3/3)} = \theta_{ijk}^{(t+2/3)} \left(\frac{g_{+jk} / g_{+++}}{\theta_{+jk}^{(t+2/3)}} \right) \text{ for all } i, j, k, \qquad (8.18)$$

where the g_{+jk} are standard gamma variates with shape parameters

$$\alpha'_{+jk} = \sum_i \left(\alpha_{ijk} + x_{ijk} \right)$$

drawn independently of those in the first two steps, and $g_{+++} = \Sigma_{jk} g_{+jk}$ is the new sum. Given any starting value that satisfies the constraints of the loglinear model (AB, AC, BC), these three steps (8.16)-(8.18) define a Markov chain

$\left\{\theta^{(t)}: t=1,2,\ldots\right\}$ which converges in distribution to the constrained Dirichlet posterior with hyperparameters $\alpha'_{ijk}=\alpha_{ijk}+x_{ijk}$; a heuristic argument for this result will be given below. Thus, for a suitably large value of t, we can regard $\theta^{(t)}$ as a random draw from the correct posterior $P(\theta|x)$. The subsequent output stream $\theta^{(t+1)}, \theta^{(t+2)}, \ldots$ represents a dependent sample from $P(\theta|x)$ which can be summarized by any of the methods described in Chapter 4.

Relationship to conventional IPF

It is easy to see the relationship between this algorithm and conventional IPF. Consider the first step of conventional IPF under the constrained Dirichlet prior,

$$\theta_{ijk}^{(t+1/3)}=\left(\frac{\theta_{ijk}^{(t+0/3)}}{\theta_{ij+}^{(t+0/3)}}\right)\left(\frac{x'_{ij+}}{n'}\right) \text{ for all } i, j, k, \tag{8.19}$$

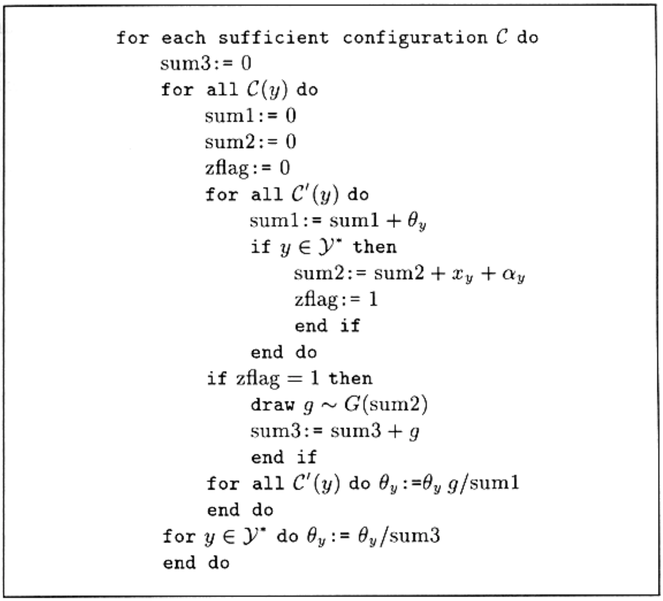
where $x'_{ijk}=x_{ijk}+\alpha_{ijk}-1$ and $n'=\Sigma_{ijk}x'_{ijk}$. The first term in parentheses on the right-hand side of (8.19) is the estimate of the conditional probability of $C=k$ given ($A=i$, $B=j$) from the previous step. The second term in parentheses is the posterior mode of the marginal probability of ($A=i$, $B=j$). Thus (8.19) represents the marriage between an old estimate of $P(C=k \mid A=i, B=j)$ and a new estimate of $P(A=i$, $B=j)$. Now consider the first step of Bayesian IPF,

$$\theta_{ijk}^{(t+1/3)}=\left(\frac{\theta_{ijk}^{(t+0/3)}}{\theta_{ij+}^{(t+0/3)}}\right)\left(\frac{g_{ij+}}{g_{+++}}\right) \text{ for all } i, j, k. \tag{8.20}$$

The first term in (8.20) is the old simulated value of $P(C=k \mid A=i, B=j)$. The second term, g_{ij+}/g_{+++}, simulates new values of the marginal probabilities $\theta_{ij+}\Sigma_k\theta_{ijk}$ from the Dirichlet distribution with parameters $\left\{\alpha'_{ij+}\right\}$, the marginal posterior distribution of θ_{ij+} given the data. Thus (8.19) represents the marriage between an old draw of $P(C=k \mid A=i, B=j)$ and a new draw of $P(A=i, B=J)$. The second and third steps of Bayesian IPF continue in a similar vein, updating the parameters by taking new random draws of the marginal probabilities $P(A=i, C=k)$ and $P(B=j, C=k)$, respectively.

An implementation in pseudocode

Like its conventional counterpart, Bayesian IPF generalizes immediately to hierarchical loglinear models for any number of variables. Pseudocode for a general implementation of Bayesian IPF is shown in Figure 8.2. This code, which uses the same notation as that of Figure 8.1, performs one cycle of Bayesian IPF and overwrites θ with its updated value. The starting value of θ must lie in the interior of the parameter space; in particular, it must satisfy the following requirements. (a) If any structural zeroes are present, the starting value must have zeroes in those positions. (b) All other elements of the starting value must be nonzero. (c) The starting



```
for each sufficient configuration $C$ do
    sum3:= 0
    for all $C(y)$ do
        sum1:= 0
        sum2:= 0
        zflag:= 0
        for all $C'(y)$ do
            sum1:= sum1 + $\theta_y$
            if $y \in \mathcal{Y}^*$ then
                sum2:= sum2 + $x_y + \alpha_y$
                zflag:= 1
                end if
            end do
        if zflag $= 1$ then
            draw $g \sim G$(sum2)
            sum3:= sum3 + $g$
            end if
        for all $C'(y)$ do $\theta_y$ :=$\theta_y$ $g$/sum1
        end do
    for $y \in \mathcal{Y}^*$ do $\theta_y$ := $\theta_y$/sum3
    end do
```

Figure 8.2. *Single cycle of Bayesian iterative proportional fitting*

.

value must satisfy the constraints of the loglinear model. One way to create a starting value with these properties is to fill θ with zeroes corresponding to the structural zeroes and uniform values elsewhere. Another good choice is a posterior mode under a constrained Dirichlet prior with all hyperparameters $\alpha_{(y)} > 1$, which can be obtained from conventional IPF.

One unusual feature of Bayesian IPF is that if a cell probability $\theta_{(\mathrm{y})}$ ever becomes zero, it remains at zero for all subsequent iterations. In theory, this should never happen for a non-structural-zero cell, because true gamma random variates are always positive. In practice, however, if the prior hyperparameters α_y are close to zero and the observed contingency table contains random zeroes, the pseudorandom number generator used in

$$\text{draw } g \sim G(\text{sum2}) \tag{8.21}$$

may occasionally produce a value with a floating-point representation of zero. The resulting value of θ will then fall on the boundary and remain trapped there for all future iterations, and the Markov chain will fail to converge to the correct posterior distribution. The problem is that boundary values are absorbing states, whose presence violates the regularity conditions necessary for an iterative simulation algorithm to converge (Section 3.5.2). In principle the probability of ever reaching the boundary should be zero, but due to the limitations of computer arithmetic, the chance of falling within machine precision of the boundary might be non-negligible. This difficulty is easily overcome by adding a very small positive constant (say 10^{-20}) to the value of g in (8.21), the effect of which will be imperceptible in the statistical inference.

8.4.4 Why Bayesian IPF works

Here we present heuristic arguments to establish that Bayesian IPF does indeed converge to the constrained Dirichlet posterior for the given loglinear model. For simplicity, let us consider the homogeneous-association model for three variables, (AB, AC, BC), where each of the variables A, B and C is binary; extensions to other loglinear models will be immediate. The relationships among the log-cell probabilities $\eta = \log\theta$ can be expressed as $\eta = M\lambda$, where

$$\eta = \begin{bmatrix} \eta_{111} \\ \eta_{211} \\ \eta_{121} \\ \eta_{221} \\ \eta_{112} \\ \eta_{212} \\ \eta_{122} \\ \eta_{222} \end{bmatrix}, M = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix},$$

and

$$\lambda = \left[\lambda_0, \lambda_1^A, \lambda_1^B, \lambda_1^C, \lambda_{11}^{AB}, \lambda_{11}^{AC}, \lambda_{11}^{BC}\right]^T.$$

The remaining loglinear coefficients follow from the identifiability constraints

$$\lambda_1^A = -\lambda_2^A,\ \lambda_{11}^{AB} = -\lambda_{12}^{AB} = -\lambda_{21}^{AB} = \lambda_{22}^{AB},$$

and so on.

Until now we have been assuming that the cell frequencies x follow a multinomial distribution,

$$x \mid \theta \sim M(n,\theta),$$

where the sample size $n = x_{+++}$ is considered fixed, and the cell probabilities $\theta = \{\theta_{ijk}\}$ follow a Dirichlet distribution with hyperparameters $\alpha = \{\alpha_{ijk}\}$. With the restrictions imposed by the loglinear model, η $\log\theta$ must lie in $R(M)$, the seven-dimensional linear space spanned by the columns of M; in addition, θ must satisfy $\theta_{+++} = 1$. This combination of linear and loglinear constraints on the elements of θ makes the parameter space somewhat difficult to visualize and understand. The geometric features can be simplified, however, if we expand the model by allowing the total sample size n to vary.

The Poisson/gamma representation

Consider an expanded model in which the cell counts are Poisson,

$$x_{ijk} \mid \mu \sim Poisson(\mu_{ijk}) \tag{8.22}$$

independently for all i, j, k, and the cell means $\mu = \{\mu_{ijk}\}$ are a priori distributed as independent gamma variates

$$\mu_{ijk} \sim cG(\alpha_{ijk}) \tag{8.23}$$

with a common scaling factor c. By well-known properties of the Poisson model, (8.22) implies that

$$\eta \mid \mu \sim \text{Poisson}(\mu_{+++})$$

and

$$x \mid n, \mu \sim M(\eta, \theta),$$

where $\mu_{+++} = \Sigma_{ijk}\mu_{ijk}$ and $\theta_{ijk} = \mu_{ijk} / \mu_{+++}$ (e.g. Agresti, 1990). Moreover, it can be shown that the product-gamma prior (8.23) implies that μ_{+++} and θ are independently distributed as

$$\mu_{+++} \sim cG(\alpha_{+++})$$

and

$$\theta \sim D(\alpha),$$

respectively; the proof is a standard exercise in transformation and will be left to the reader.

Thus the expanded model (8.22)-(8.23) for x and μ implies our usual multinomial-Dirichlet model for x and μ; the only difference is that the expanded model allows estimation of an overall intensity parameter μ_{+++} which is independent of θ in both the prior and the posterior distributions. By standard Bayesian arguments, the posterior distributions of the cell means μ are

$$\mu_{ijk} \mid x \sim c' G(\alpha'_{ijk}) \tag{8.24}$$

where $\alpha'_{ijk} = \alpha_{ijk} + x_{ijk}$ and $c' = c/(c+1)$, and the posterior distribution of the intensity parameter is

$$\mu_{+++} \mid x \sim c' G(\alpha'_{+++})$$

where $\alpha'_{+++} = \Sigma_{ijk}\alpha_{ijk} + n$. As before, the posterior distribution of θ is $D(\alpha'), \alpha' = \alpha + x$.

The Poisson/gamma representation is geometrically convenient because, unlike θ, the cell means μ are not required to sum to any particular value. The loglinear model for the cell probabilities, $\log\theta = M\lambda$ implies a similar model for the cell means,

$$\log\mu = M\lambda*, \tag{8.25}$$

where

$$\lambda* = \left[\lambda_0^*, \lambda_1^A, \lambda_1^B, \lambda_1^C, \lambda_{11}^{AB}, \lambda_{11}^{AC}, \lambda_{11}^{BC}\right]^T$$

and $\lambda_0^* = \lambda_0 + \log \mu_{+++}$. Unlike λ_0 the new intercept λ_0^* is a free parameter that can take any value on the real line. Thus the parameter space for p is simpler than the space for θ, because $\log\mu$ is allowed to lie anywhere in $R(M)$.

The cell-means version of Bayesian IPF

Under the expanded model, we can define a version of Bayesian IPF that operates on the cell means μ. It is similar to the version for θ except that we do not rescale the parameters to sum to one at every step. The cell-means version is

$$\mu_{ijk}^{(t+1/3)} = \mu_{ijk}^{(t+0/3)} \left(\frac{c' g_{ij+}}{\mu_{ij+}^{(t+0/3)}} \right) \text{ for all } i, j, k, \tag{8.26}$$

$$\mu_{ijk}^{(t+2/3)} = \mu_{ijk}^{(t+1/3)} \left(\frac{c' g_{i+k}}{\mu_{i+k}^{(t+1/3)}} \right) \text{ for all } i, j, k, \tag{8.27}$$

$$\mu_{ijk}^{(t+3/3)} = \mu_{ijk}^{(t+2/3)} \left(\frac{c' g_{jk}}{\mu_{ijk}^{(t+2/3)}} \right) \text{ for all } i, j, k, \tag{8.28}$$

where the g_{ij+}, g_{i+k} and g_{+jk} are independent gamma variates as before. To see the relationship to the previous version, notice that we can rewrite the first step as

$$\begin{aligned} \mu_{ijk}^{(t+1/3)} &= \left(\frac{\theta_{ijk}^{(t+0/3)}}{\theta_{ij+}^{(t+0/3)}} \right) \left(\frac{g_{ij+}}{g_{+++}} \right) c' g_{+++} \\ &= \theta_{ijk}^{(t+1/3)} c' g_{+++}, \end{aligned}$$

and similarly for the second and third steps. The value of θ at every step is the same under the new version; the only difference is that the intensity parameter

$$\mu_{+++} = \sum_{i,j,k} \theta_{ijk} c' g_{+++} = c' g_{+++}$$

is updated at every step to be a random draw from $c' G(\alpha'_{+++})$.

Without constraints, the product-gamma posterior (8.24) for μ implies that the posterior distribution of θ is $D(\alpha')$. Constraining μ and θ to lie in $R(M)$ does not change the functional form of the densities for μ or θ, but only their normalizing constants; so if we are able to show that the cell-means version of Bayesian IPF converges to the constrained product-gamma posterior over $R(M)$, then we have successfully shown that the original version converges to the constrained Dirichlet posterior over $R(M)$.

Heuristic argument for convergence

The three steps (8.26)-(8.28) define the transition rule of a Markov chain for μ. To show that a particular distribution F is the stationary distribution for this chain, one must establish two facts. First, one must show that the chain ’maps F onto itself,’ in other words, that $\mu^{(t)} \sim F$ implies $\mu^{(t+1)} \sim F$. Second, one must show that the chain is ergodic, containing no periodicities or absorbing states: it must be possible for $\mu^{(t)}$ to reach any value in the support of F for a sufficiently large t, and for every step thereafter, from any starting value $\mu^{(0)}$ within the support.

To establish the first condition, notice that the first step (8.26) combines conditional probabilities $P(C = k \mid A = i, B = j)$ from the previous cycle with updated values for the $(A = i, B = j)$ marginal rates,

$$\mu_{ijk}^{(t+1/3)} = \theta_{(ij)k}^{(t+0/3)} \mu_{ij+}^{(t+1/3)},$$

where

$$\theta_{(ij)k}^{(t+0/3)} = \frac{\theta_{ijk}^{(t+0/3)}}{\theta_{ij+}^{(t+0/3)}} = \frac{\mu_{ijk}^{(t+0/3)}}{\mu_{ij+}^{(t+0/3)}}$$

and

$$\mu_{ij+}^{(t+1/3)} \sim c'G\left(\alpha'_{ij+}\right) \tag{8.29}$$

independently for all i, j. Because (8.29) is the marginal distribution of $\left\{\mu_{ij+}\right\}$ implied by (8.24), this first step represents a draw from the conditional posterior of μ with $\left\{\theta_{(ij)k}\right\}$ fixed at its previous value,

$$\mu^{(t+1/3)} \sim P\left(\mu \mid x, \left\{\theta_{(ij)k}\right\} = \left\{\theta_{(ij)k}^{(t+0/3)}\right\}\right).$$

Now if the old value of μ is drawn from the actual posterior distribution,

$$\mu^{(t+0/3)} \sim P(\mu \mid x),$$

then the old values of $\theta_{(ij)k}$ are drawn from their actual posterior as well,

$$\left\{\theta_{(ij)k}^{(t+0/3)}\right\} \sim P\left(\left\{\theta_{(ij)k}\right\} \mid x\right)$$

which implies

$$\mu^{(t+1/3)} \sim P(\mu \mid x)$$

Thus we have established that the first step (8.26) maps $P(\mu|x)$ onto itself. By similar arguments this result holds for the second and third steps (8.27)-(8.28) as well.

To establish ergodicity, we must examine the structure of the design matrix M in the loglinear model (8.25) for μ. Let

$$M_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, M_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

and

$$M_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

denote the portions of M corresponding to the AB, AC and BC effects, respectively. The space spanned by the columns of M_1 is the same as that of

$$M_1^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly, the range spaces of M_2 and M_3 are the same as those of

$$M_2^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, M_3^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

respectively. Notice that the first step of Bayesian IPF represents a proportionate adjustment for each group of cells that contributes to a mean μ_{ij+} for the AB marginal table.

This first step can thus be written in terms of the log-cell means as

$$\log \mu^{(t+1/3)} = \log \mu^{(t+0/3)} + M_1^* \log \gamma_1$$

where γ_1 is a vector of four gamma variates. Similarly, the second and third steps can be written

$$\log \mu^{(t+2/3)} = \log \mu^{(t+1/3)} + M_2^* \log \gamma_2$$

$$\log \mu^{(t+2/3)} = \log \mu^{(t+2/3)} + M_3^* \log \gamma_3$$

The complete cycle consisting of all three steps is thus

$$\log \mu^{(t+3/3)} = \log \mu^{(t+2/3)} + M_3^* \log \gamma_3 \qquad (8.30)$$

where $M^* = \left(M_1^*, M_2^*, M_3^*\right)$ and $\gamma = \left(\gamma_1^T, \gamma_2^T, \gamma_3^T\right)^T$. Because the columns of M^* span the same space as those of M, and the elements of γ are random gamma variates whose logarithms can lie anywhere on the real line, (8.30) ensures that $\mu^{(t+1)}$ has a nonzero chance of falling anywhere in $R(M)$ provided that $\mu^{(t)} \in R(M)$. Thus it follows that this Markov chain can reach any state in a single cycle from any state in the parameter space of μ, and ergodicity is established.

Further notes

Bayesian IPF was first presented by Gelman et al. (1995) who gave brief arguments for convergence under a Poisson/gamma loglinear model. Our version (8.26)-(8.28) differs from theirs in that we have included a factor c' arising from the scaling parameter in the prior distribution (8.23). Without this factor, the resulting posterior could give misleading inferences about the overall intensity μ+++. Inferences about the cell probabilities $\theta_{ijk} = \mu_{ijk} / \mu_{+++}$, however, will be the same under both versions.

Bayesian IPF bears an interesting relationship to Gibbs sampling (Section 3.4.1). In Gibbs sampling, we partition a random vector Z into non-overlapping subvectors (Z_1, Z_2,..., Z_J) and draw from the full conditionals $P\left(Z_j \mid \left\{Z_k : k \neq j\right\}\right)$ for j = 1,..., J in turn. In Bayesian IPF, however, the vector is partitioned differently at each step of the cycle; in our example we partition μ as

$$\left(\left\{\mu_{ij}\right\}, \left\{\mu_{ijk} / \mu_{ij+}\right\}\right) \text{ at Step 1,}$$

$$\left(\left\{\mu_{i+k}\right\}, \left\{\mu_{ijk} / \mu_{i+k}\right\}\right) \text{ at Step 2,}$$

$$\left(\left\{\mu_{+jk}\right\}, \left\{\mu_{ijk} / \mu_{+jk}\right\}\right) \text{ at Step 3.}$$

As noted by several authors (e.g. Gelfand and Smith, 1990), any partitioning scheme will work provided that the complete cycle is ergodic, allowing the random vector to eventually reach any state from any other state.

8.4.5 Example: misclassification of seatbelt use and injury

In Section 8.3.4 we examined loglinear models pertaining to errors in police reporting of seatbelt use and injury in automobile accidents. Regarding the error indicators E_B and E_I, as response variables and D, S, B_2 and I_2 as predictors, we found convincing evidence for the associations $B_2I_2E_BE_I$, D_{EB}, DE_I and mild evidence for SE_I. This evidence was based on p-values from chisquare approximations for the test statistics ΔG^2 and ΔX^2. Using Bayesian IPF, we can make Bayesian inferences about these associations directly without large-sample approximations. To illustrate, we ran Bayesian IPF under Model 8,

$$\left(DSB_2I_2, B_2I_2E_BE_I, DE_B, DE_I, SE_I\right) \tag{8.31}$$

Recall that the ML estimate for this model lies in the interior of the parameter space. Taking the ML estimate from IPF as a starting value, we ran 5100 cycles of Bayesian IPF under the Jeffreys

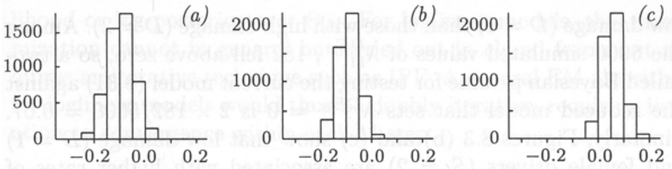


Figure 8.3. *Histograms of 5000 simulated values of (a)* $\lambda_{11}^{DE_B}$ *(b)* $\lambda_{11}^{DE_I}$ *and (c)* $\lambda_{11}^{SE_I}$ *, respectively, from Bayesian IPF.*

prior with hyperparameters 0.5; ignoring the first 100 cycles, we stored the results of the remaining 5000. Bayesian IPF appears to converge very quickly in this example; autocorrelation plots of a variety of parameters revealed no significant correlations beyond lag 5. This behavior is consistent with that of ordinary IPF, which converged in only 11 cycles.

Let us consider how to draw inferences about the coefficients λ of the loglinear model. The output of each cycle of Bayesian IPF is a table of simulated cell probabilities θ, and each coefficient in λ is a linear contrast among the elements of log θ. Consider the terms in λ^{DE_B}, which pertain to the effect of damage on errors in the police report of belt use. If we average the elements of the 2^6 array log θ over the dimensions corresponding to S, B_2, I_2 and E_I, all the coefficients pertaining to these variables drop out due to the linear constraints imposed on them. The result of this averaging is a 2 × 2 table with elements

$$\gamma_{ij}^{DE_B} = \lambda_0 + \lambda_i^D + \lambda_j^{E_B} + \lambda_{ij}^{DE_B}$$

for $i, j = 1, 2$. The loglinear coefficients can then be obtained as

$$\begin{aligned}\lambda_0 &= \tfrac{1}{4}\Sigma_{ij}\gamma_{ij}^{DE_B},\\ \lambda_i^D &= \tfrac{1}{2}\Sigma_j\gamma_{ij}^{DE_B} - \lambda_0,\\ \lambda_j^{E_B} &= \tfrac{1}{2}\Sigma_i\gamma_{ij}^{DE_B} - \lambda_0,\\ \lambda_{ij}^{DE_B} &= \gamma_{ij}^{DE_B} - \lambda_i^D - \lambda_j^{E_B} - \lambda_0.\end{aligned}$$

By similar manipulations of the elements of log θ we can derive any coefficient in the loglinear model. Histograms of the 5000 simulated values of $\lambda_{11}^{DE_B}$, $\lambda_{11}^{DE_I}$ and $\lambda_{11}^{SE_I}$ are shown in Figure 8.3 (a), (b) and (c), respectively.

Notice that the distribution in Figure 8.3 (a) is located primarily to the left of zero, providing evidence that errors in reporting of belt use (E_B = 2) tend to occur more frequently for accidents with low damage (D = 1) than those with high damage (D = 2). Among the 5000 simulated values of $\lambda_{11}^{DE_B}$ fell above zero, so a two-tailed Bayesian p-value for testing the current model (8-31) against the reduced model that sets $\lambda^{DE_B} = 0$ is 2 × 182/5000. Similarly, Figures 8.3 (b) and (c) show that low damage (D = 1) and female drivers (S = 2) are associated with higher rates of reporting errors for injury (E_I = 2); two-tailed Bayesian p-values for testing λ^{DE_I} and λ^{SE_I} are 0.03 and 0.13, respectively.

8.5 Loglinear modeling with incomplete data

8.5. 1 ML estimates and posterior modes

The two algorithms we have discussed thus far, IPF and Bayesian IPF, can be extended quite easily to handle missing values in the original data matrix. With conventional IPF, the extension is an example of the ECM algorithm, a generalization of EM discussed briefly in Chapter 3.

EM for loglinear models

Let us now return to our general notation, with Y_1, Y_2,..., Y_p representing categorical variables, and $x = \left\{ x_y : y \in \mathcal{Y} \right\}$ and $\theta = \left\{ \theta_y : y \in \mathcal{Y} \right\}$ the cell counts and probabilities, respectively, in the p-dimensional cross-classified contingency table. As described in Section 7.3, the actual cell counts x are not observed when the data are incomplete; rather, we observe a table $z^{(s)}$ of potentially smaller dimension for each missingness pattern $s = 1, 2, \ldots, S$ where $z^{(s)}$ contains marginal frequencies for the variables observed in pattern s. The basic EM algorithm (Section 7.3.2) updates the estimate of θ in two steps: the E-step, in which we calculate the predicted mean of x given $z^{(1)}, \ldots, z^{(S)}$ under the current estimate of θ; and the M-step, in which we re-estimate θ from the predicted mean of x. Under the saturated multinomial model, the M-step has a particularly simple form because the complete-data ML estimates are $\hat{\theta} = x_y / n$ for all $y \in \mathcal{Y}$.

Now consider what happens to EM when we move from the saturated model to a loglinear model, which requires the parameter θ to lie in a restricted space Θ_M. The E-step, which is performed under an assumed value of θ, does not change at all, because the conditional expectation of x given $z^{(1)}, \ldots, z^{(S)}$ and θ has the same form whether $\theta \in \Theta_M$ or $\theta \notin \Theta_M$. The M-step, however, becomes a constrained maximization of the (expected) complete-data loglikelihood or log-posterior over Θ_M. For loglinear models, this maximization cannot in general be carried out in closed form, but requires an iterative technique such as IPF. A general EM algorithm for loglinear models would thus be doubly iterative, requiring iteration to convergence within each M-step.

The ECM algorithm

In many situations, EM for loglinear models is not unduly cumbersome, especially in modern computing environments. Several authors speculated, however, that it might not be necessary to iterate until full convergence at each M-step; rather, running only a single cycle of IPF might be enough (e.g. Fuchs, 1982). As shown by Meng and Rubin (1993), this modification produces an example of an algorithm called Expectation-Conditional Maximization or ECM. ECM possesses the same reliable convergence properties as EM, increasing the observed-data loglikelihood at each step. The key idea of ECM is that the full M-step is replaced by a quicker CM-step, a single cycle of constrained maximizations which, if repeated over and over, would eventually result in a maximization over the full parameter space Θ_M.

Each step in a cycle of IPF is a constrained maximization. Consider the three steps of IPF for the model (AB, AC, BC). The first step (8.6) is based on a factorization of the complete-data likelihood for the cell probabilities $\theta_{ijk} = P(A = i, B = j, C = k)$ into independent factors corresponding to the conditional probabilities for C given A and B,

$$P(C = k \mid A = i, B = j) = \theta_{ijk} / \theta_{ij+},$$

and the marginal probabilities for A and B,

$$P(A = i, B = j) = \theta_{ij+}.$$

The first step fixes $\{\theta_{ijk} / \theta_{i+k}\}$ at its previous value but replaces $\{\theta_{ij+}\}$ by an ML estimate or posterior mode; thus it represents a constrained maximization of the likelihood or posterior density for θ. Similarly, the second and third steps (8.7)-(8.8) represent maximizations subject to fixed values of $\{\theta_{ijk} / \theta_{i+k}\}$ and $\{\theta_{ijk} / \theta_{+jk}\}$, respectively.

For a cycle of constrained maximizations to form a valid CM step, it must satisfy a set of technical requirements known as the space-filling conditions (Meng and Rubin, 1993). These conditions, which have been demonstrated to hold for a single cycle of IPF (Meng and Rubin, 1991b), are similar to those needed to establish ergodicity in a Markov chain (Section 8.4.4). The fact that IPF can reach any point in Θ_M from any other point in a single cycle ensures that the ECM algorithm will eventually converge to an unconstrained maximum rather than a constrained one.

To obtain a general ECM algorithm for loglinear models, we only need to replace the M-step of the basic EM (the last line of pseudocode in Figure 7.2) by a single call to the IPF algorithm presented in Figure 8.1. As in ordinary IPF, the starting value for θ should lie in the interior of the parameter space; one choice that always works is to assign zeroes to the structural-zero cells and uniform values elsewhere.

8.5.2 Goodness-of-fit statistics

The goodness-of-fit statistics G_2 and X_2 can be readily extended to handle incomplete data. Using the notation of Section 7.3, let $y = (y_1, y_2,..., y_p)$ denote a generic realization of the variables $(Y_1, Y_2,..., Y_p)$, and let $O_s(y)$ and $M_s(y)$, respectively, denote the subvectors of y corresponding to the variables that are observed and missing in pattern s. Let O_s and M_s denote the sets over which $O_s(y)$ and $M_s(y)$ can vary, excluding structural zeroes. The marginal table $z^{(s)}$ that we observe for pattern s has counts

$$z^{(s)}_{O_s(y)} = \sum_{M_s(y)\in M_s} x^{(s)}_{y} \text{ for all } O_s(y) \in O_s \tag{8.32}$$

and the marginal probabilities corresponding to these counts are

$$\beta_{O_s(y)} = \sum_{M_s(y)\in M_s} \theta_y . \tag{8.33}$$

The most obvious way to extend G_2 is to take

$$G^2_{raw} = 2\sum_{s=1}^{S} \sum_{O_s(y)\in O_s} z^{(s)}_{O_s(y)} \log \frac{z^{(s)}_{O_s(y)}}{n_s \beta_{O_s(y)}}, \tag{8.34}$$

where n_s represents the total sample size in pattern s. This statistic, when evaluated at an ML estimate $\hat{\theta}$, increases as the observed frequencies deviate from their estimated expected values. The corresponding extension of Pearson's X^2 is

$$X^2_{raw} = \sum_{s=1}^{S} \sum_{O_s(y)\in O_s} \frac{\left(z^{(s)}_{O_s(y)} - n_s\beta_{O_s(y)}\right)^2}{n_s\beta_{O_s(y)}}. \tag{8.35}$$

As noted by Fuchs (1982) and Little and Rubin (1987), these statistics differ from their complete-data counterparts in that they are typically nonzero even when evaluated at the ML estimate for the saturated model. The reason, which is somewhat technical, is buried in the definition of the observed-data likelihood in Chapter 2. The marginal probabilities (8.33) are not really the expected proportions for the observed counts (8.32) within each missingness pattern, unless the missingness happens to be MCAR. Under the less restrictive assumption of MAR, the true expected proportions may differ from (8.33), because MAR does not require the distribution of observed data to be identical across patterns. Using the same cell probabilities θ in the calculation of (8.33) for all patterns is merely a matter of convenience, because, as argued in Section 2.3, likelihood-based inferences for parameters of the complete-data model are identical under any ignorable mechanism.

The practical effect of using a common θ for all patterns is that G^2_{raw} and X^2_{raw} as defined above may be drastically different from zero even when evaluated at the ML estimate for the saturated model. In fact, if the sample is large enough, these statistics can be used to test the null hypothesis that the missingness data are MCAR against the alternative of MAR. In most situations, such a test will not be of great interest, because we are concerned primarily with the parameters of the complete-data model; the parameters of the missingness mechanism are a nuisance. Moreover, in all but the most trivial real data examples, the expected cell counts within missingness patterns are rarely large enough for the chisquare approximation to work well. For these reasons, we will not attempt to interpret a value of G^2_{raw} or X^2_{raw} from the saturated model, except as a baseline for assessing the fit of a smaller model.

Adjusted goodness-of-fit statistics

Let G_0^2 denote the value of (8.34) evaluated at the ML estimate for the saturated model. Consider the adjusted goodness-of-fit measure

$$G^2 = 2\sum_{s=1}^{S} \sum_{O_s(y)\in O_s} z^{(s)}_{O_s(y)} \log \frac{z^{(s)}_{O_s(y)}}{n_s\beta_{O_s(y)}} - G_0^2 \qquad (8.36)$$

regarded as a function of θ. This represents twice the difference in the observed-data loglikelihood evaluated at the current value of θ and at the global maximum over the entire simplex Θ. When evaluated at the ML estimate for the saturated model, (8.36) is zero. When evaluated at the ML estimate for a non-saturated loglinear model, it becomes the likelihood-ratio statistic for testing the fit of the model against the saturated alternative. Because this statistic behaves in much the same manner as the deviance (8.9) for complete data, we will adopt (8.36) as our definition for the deviance with incomplete data. The Pearson counterpart to (8.36) is

$$X^2 = \sum_{s=1}^{S} \sum_{O_s(y)\in O_s} \frac{\left(z^{(s)}_{O_s(y)} - n_s\beta_{O_s(y)}\right)^2}{n_s\beta_{O_s(y)}} - X_0^2, \qquad (8.37)$$

where X_0^2 represents the raw version (8.35) evaluated at the MLE for the saturated model.

Just as in the complete-data case, the chisquare approximation for the distributions of G^2 and X^2 may be poor when the data are sparse. Even with sparse data, however, chisquare approximations for the differences ΔG^2 and ΔX^2 can be quite reliable for nested model comparisons, particularly when Δdf is small. Finally, just as with complete data, G^2 and X^2 become problematic when ML estimates lie on the boundary. The easiest way to handle such situations is to add a small amount of prior information, e.g. in the form of a

Dirichlet prior with all hyperparameters greater than one, to smooth the posterior mode away from the boundary.

8.5.3 Data augmentation and Bayesian IPF

We have seen that IPF can be extended in a straightforward way to handle missing data, resulting in an ECM algorithm. In a similar fashion, we can extend Bayesian IPF to create an algorithm for parameter simulation and multiple imputation.

Recall the basic data augmentation procedure for the saturated multinomial model (Section 7.3.3). In this algorithm, the observed marginal counts $z^{(s)}$ for missingness patterns $s = 1, 2, \ldots, S$ are randomly allocated to the cells of the full p-dimensional table x under an assumed value for θ (the I-step). Then a new value of θ is drawn from its complete-data Dirichlet posterior given the simulated version of x (the P-step). Repeating the I- and P-steps a large number of times eventually produces a draw of θ from its observed-data posterior $P(\theta \mid Y_{obs})$.

As we move from the saturated model to a loglinear model, the I-step remains unchanged, because the random allocation procedure has the same form regardless of the value of θ. The P-step, however, must in general be carried out iteratively, because creating posterior draws of θ under a loglinear model requires multiple cycles of Bayesian IPF. True data augmentation, like a true EM algorithm, would thus require undesirable nested iterations.

Suppose, however, that instead of iterating to full convergence within each P-step, we perform only a single cycle of Bayesian IPF. The resulting algorithm, which we call *data augmentation-Bayesian IPF* (DABIPF), still converges to the observed-data posterior under the constrained Dirichlet prior. Although we are not drawing from the correct conditional distribution $\mathrm{P}(\theta \mid x)$ in this modified P-step, we are simulating one step from the transition rule of a Markov chain whose stationary distribution is $\mathrm{P}(\theta \mid x)$. DABIPF is neither a true data augmentation algorithm nor a Gibbs sampler, but a hybrid algorithm of the type described in Section 3.4.5 with the same basic convergence properties. Combining the data augmentation I-step (Section 7.3.3) with the implementation of Bayesian IPF in Figure 8.2 produces a single iteration of DABIPF. The algorithm may be used for parameter simulation or (when combined with the imputation code in Figure 7.5) for multiple imputation of the unit-level missing data.

Factorizations for near-monotone patterns

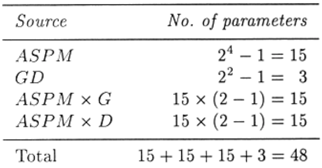
In the last chapter, we derived a monotone data augmentation procedure for the saturated model that tends to converge faster than ordinary data augmentation for near-monotone missingness patterns (Section 7.4). The algorithm was based on a factorization of the multinomial likelihood into a sequence of multinomial likelihoods pertaining to the distribution of each variable given the previous ones. For loglinear models, we can again exploit factorizations of the likelihood to improve the performance of DABIPF, but only in certain special cases. For many loglinear models, the parameters corresponding to the sequence of conditional distributions are not distinct, due to the loglinear restrictions-, we cannot always separate the complete-data inference into a sequence of independent inferences corresponding to the near-monotone pattern in the dataset. For this reason, we will not consider further the possible monotone versions of DABIPF.

8.6 Examples

8.6.1 Protective Services Project for Older Persons

Recall the data introduced in the last chapter (Section 7.3.5) regarding the impact of social services on elderly clients. In this six

Table 8.5. *Parameters in the model (ASPMG, ASPMD, GD)*



Source	*No. of parameters*
$ASPM$	$2^4 - 1 = 15$
GD	$2^2 - 1 = 3$
$ASPM \times G$	$15 \times (2-1) = 15$
$ASPM \times D$	$15 \times (2-1) = 15$
Total	$15 + 15 + 15 + 3 = 48$

variable dataset, the main question of interest pertained to the effect of the treatment-group indicator G on survival D, controlling for the possible confounding effects of the covariates age A, sex S, physical status P and mental status M. With only $n = 164$ clients in the study, some of whom had missing values for P and M, the data were too sparse to allow for estimation of individual GD associations within each of the sixteen covariate patterns. Now we will fit a loglinear model that constrains these associations to be the same.

Consider the model (ASPMG, ASPMD, GD). The presence of the association ASPMG allows the distribution of G to vary freely across the sixteen ASPM covariate patterns; similarly, ASPMD allows the distribution of D to vary freely across the covariate patterns. The GD association allows G to have a direct influence on D beyond that provided by their mutual associations with A, S, P and M. The absence of any association between GD and ASPM, however, requires the conditional GD odds ratios within the sixteen covariate patterns to be equal. The number of free parameters in this model, 48, can be counted as shown in Table 8.5. Notice that the saturated model has $2^6 - 1 = 63$ free parameters; the difference is $63 - 48 = 15$ because the saturated model fits 16 conditional GD odds ratios rather than just one. By pooling information across covariate patterns, the reduced model can provide a stable estimate of a common GD effect even though the information within any single pattern may be weak. With complete data, inferences about the conditional GD association under this model would be similar to those given by the well known Mantel-Haenszel test (Mantel and Haenszel, 1959; Agresti, 1990).

Using the tools of Section 8.5, we can draw inferences about the effect of interest in two ways. First, we can perform

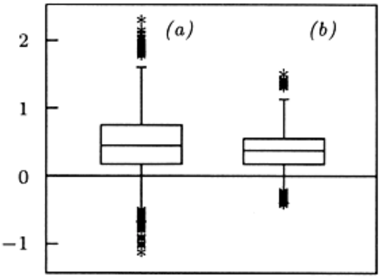


Figure 8.4. *Posterior draws of GD log-odds ratios under prior hyperparameters of (a) 0.1 and (b) 1.5, simulated from 5000 cycles of DABIPF.*

a single degree-of-freedom test of the null model (ASPMG, ASPMD) against the (ASPMG, ASPMD, GD) alternative using a large-sample approximation to ΔX^2 or ΔG^2. Under both of these models, the ECM algorithm converges to ML estimates on the boundary. To move the estimates away from the boundary, we re-ran ECM for each model using prior hyperparameters of 1.1. Comparing the observed-data loglikelihood at the two posterior modes, we find $\Delta G^2 = 0.826$. The corresponding p-value is 0.36, so there is essentially no evidence of any conditional *GD* association given *A*, *S*, *P* and *M*.

A second method, which does not rely on large-sample approximations, is to simulate posterior draws of the parameters under the larger model (ASPMG, ASPMD, GD) and examine the marginal distribution of the conditional GD association. Starting from the posterior mode that we obtained from ECM, we ran 5000 cycles of DABIPF after a burn-in period of 100 cycles. We did this under two alternative priors, setting the prior hyperparameters to 0. 1 and 1.5, respectively; the first results in very little smoothing, whereas the second pulls the estimates rather strongly toward a uniform table. Boxplots of the simulated log-odds ratios under these two priors are shown in Figure 8.4. Simulated posterior means and interval estimates for the odds ratio in (a) are 1.74 and (0.68, 3.75), respectively; the corresponding estimates for (b) are

1.50 and (0.82, 2.51). Thus there is no evidence to suggest that enriched social services are beneficial to clients. On the contrary, we find weak evidence that membership in the experimental group (G = 1) is associated with an increased rate of mortality (D = 1). The effects are not statistically significant,' however; simulated two-tailed Bayesian p-values are 0.26 for (a) and 0.20 for (b).

Recall that when we tried to draw inferences about the conditional GD associations under the saturated model (Section 7.3.5), we encountered difficulty because the parameters were so poorly estimated. Comparing the new boxplots in Figure 8.4 to the old ones in Figure 7.7, we see that the new inferences are much more plausible, and also much less sensitive to the choice of prior hyperparameters; pooling across covariate patterns to estimate a common odds ratio was indeed helpful.

8.6.2 Driver injury and seatbelt use

In the last chapter (Section 7.4.3) we examined data from a large sample of 80 084 automobile accidents and found apparently convincing evidence that seatbelt use reduces the risk of injury. Those data, however, were marred by errors in the police reports of injury and belt use. A followup study of an additional 1796 accidents provided information on error rates in the police reports. We attempted to use the followup data to calibrate the larger dataset, correcting the inference for potential biases due to misclassification. Those efforts were hindered by the complexity of the saturated model, the only model available to us at the time. Let us re-examine these data by applying a more parsimonious loglinear model to the combined sample of 81 880 accidents.

Our loglinear modeling of the 1796 followup cases (Section 8.3.4) provides some insight into the misclassification mechanism. Among various models relating damage D, driver sex S, true belt use B_2 and true injury I_2 to the error indicators for belt use E_B and injury E_I, we found that

$$\left(DSB_2I_2, B_2I_2E_BE_I, DE_B, DE_I, SE_I\right)$$

seemed to provide a good fit. For modeling the combined dataset of 81 880 cases, however, it is more convenient to work with D, S, B_2, I_2 and the original police-report variables B_1 and I_1 because E_B and E_I are not determined for the accidents not included in the followup study. Because E_B is a function of B_1 and B_2, and E_I is a function of I_1 and I_2, all the associations in the above model are present in

$$\left(DSB_2I_2B_1I_1B_2I_2, DB_1B_2, DI_1I_2, SI_1I_2\right).$$

Furthermore, we expect the full four-way association $\mathrm{DS}B_1I_1$ to be well estimated because these four variables are recorded for all cases in the combined dataset. Therefore, we will fit the model

$$\left(DSB_1I_1, DSB_2I_2, B_1I_1B_2I_2, DB_1B_2, DI_1I_2, SI_1I_2\right),$$

which has a total of 39 free parameters.

Under this model, both ECM and DABIPF converge rather slowly. This is not surprising, because B_2 and I_2 are missing for about 98% of the cases in the combined dataset. Because of the slow convergence, it would be difficult to draw inferences by parameter simulation; consecutive draws from DABIPF are so highly correlated that a very large number of cycles would be needed to obtain good posterior summaries. Instead of storing draws of parameters, we created ten multiple imputations of the followup belt use B_2 and injury status I_2. These imputations were created by running ten parallel chains of DABIPF for 2500 cycles each, using the ML estimate as a starting value and setting the prior hyperparameters equal to 0.5. The ten imputations are shown in Table 8.6. The imputed variables, denoted by B and I represent true belt use and injury, and the variation among the ten imputations reflects the uncertainty due to misclassification in the original data. After imputation the followup cases were removed, so that only the original 80 084 accidents are represented in the imputed data.

Using these ten imputations, we calculated point and interval estimates for the odds ratios relating seatbelt use to injury. The estimation was carried out on the logarithmic scale, as described in Section 6.4.2. Results for the overall odds ratio, and for the conditional odds ratios within each of the four $D \times S$ cells, are summarized in Table 8.7. Over the entire population, seatbelt use appears to reduce the odds of injury by about $1 - 0.73 = 27\%$, and the effect is statistically significant (p-value = 0.04). This marginal analysis, however, ignores the possible confounding effects due to the covariates D and S. Within the four $D \times S$ cells, the estimated odds ratios are all less than one, but the interval estimates are very wide; none of the effects is statistically significant. After controlling for damage and sex of driver, the evidence for any beneficial effect of seatbelt use is very weak.

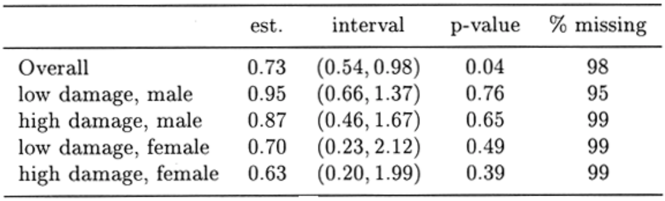
The fact that all four of the conditional odds ratios are less than one suggests that we may be able to strengthen our conclusions by assuming a common odds ratio across the four $D \times S$ cells; that is, by fitting the loglinear model (DSB, DSI, BI), we may be able to find a significant BI association. The loglinear model (DSB, DSI, BI) implies a logit model for I that includes main effects for D and S, a $D \times S$ interaction and a main effect

Table 8.6. *Multiple imputations of accident frequencies by damage D (1=low, 2=high), sex of driver S (1=male, 2=female), actual belt use B (1=no, 2=yes), and actual injury (1=not injured, 2=injured), reflecting errors of classification*

				Imputations 1–5				
D	*S*	*B*	*I*	1	2	3	4	5
1	1	1	1	18275	17698	18344	18570	18132
2	1	1	1	14667	15195	13739	14817	14218
1	2	1	1	9254	8735	8618	9341	8525
2	2	1	1	4311	4631	4565	4271	4516
1	1	2	1	5212	4954	4719	4865	4986
2	1	2	1	2558	2446	2860	2664	2670
1	2	2	1	2018	1471	1923	1607	1653
2	2	2	1	762	633	456	1117	703
1	1	1	2	3142	3847	3414	3173	3356
2	1	1	2	7912	7824	9113	8133	8499
1	2	1	2	2428	3544	3048	2740	3366
2	2	1	2	6165	5968	5897	5865	5741
1	1	2	2	799	929	951	820	954
2	1	2	2	1823	1495	1248	1346	1573
1	2	2	2	300	250	411	312	456
2	2	2	2	458	464	778	443	736

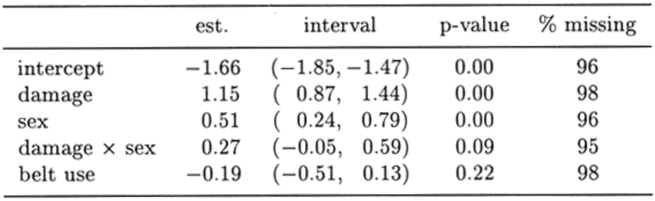
				Imputations 6–10				
D	*S*	*B*	*I*	1	2	3	4	5
1	1	1	1	18562	18433	18019	17766	18302
2	1	1	1	13612	14601	13405	13911	14021
1	2	1	1	8832	8671	9748	9157	8991
2	2	1	1	4149	4191	4395	4097	4535
1	1	2	1	4840	4742	5365	4966	5022
2	1	2	1	2979	2823	2907	3055	2862
1	2	2	1	2005	1724	1325	1585	1597
2	2	2	1	592	566	945	956	660
1	1	1	2	3346	3275	3142	3591	3422
2	1	1	2	9096	8413	9303	8454	8174
1	2	1	2	2966	3013	2506	2770	2821
2	2	1	2	6285	6088	5818	6042	5744
1	1	2	2	680	978	902	1105	682
2	1	2	2	1273	1123	1345	1540	1903
1	2	2	2	197	592	421	488	591
2	2	2	2	670	851	538	601	757

Table 8.7. *Multiple-imputation inferences for odds ratios relating to belt use to injury, overall and within cells of damage by sex of driver: estimates, intervals, p-values and percent missing information*



	est.	interval	p-value	% missing
Overall	0.73	(0.54, 0.98)	0.04	98
low damage, male	0.95	(0.66, 1.37)	0.76	95
high damage, male	0.87	(0.46, 1.67)	0.65	99
low damage, female	0.70	(0.23, 2.12)	0.49	99
high damage, female	0.63	(0.20, 1.99)	0.39	99

Table 8.8. *Multiple-imputation inferences for logistic-regression coefficients for predicting injury, assuming a common effect of belt use across classes of damage and sex*



	est.	interval	p-value	% missing
intercept	−1.66	(−1.85, −1.47)	0.00	96
damage	1.15	(0.87, 1.44)	0.00	98
sex	0.51	(0.24, 0.79)	0.00	96
damage × sex	0.27	(−0.05, 0.59)	0.09	95
belt use	−0.19	(−0.51, 0.13)	0.22	98

for B. We fit this logit model to each of the imputed datasets, coding dummy variables for the main effects of D (1 if high damage, 0 otherwise), S (1 if female, 0 otherwise), B (1 if belt used, 0 otherwise) and the $D \times S$ interaction (1 if high damage and female, 0 otherwise). The results of the multiply-imputed logit analysis are shown in Table 8.8. The estimate of the common odds ratio is exp(-0.19) = 0.83 and the 95% interval ranges from exp(-0.51) = 0.60 to exp(0.13) = 1.14, so the evidence is still weak. Accounting for occasional errors in the police reports greatly increases our uncertainty about the relationship between belt use and injury. These results are consistent with those of Chen (1989), who reached similar conclusions using likelihood-based methods.