Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress.

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

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

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 imputa-

tion model is effectively applied not to the entire dataset but only to its missing part. For example, it may be quite reasonable to use a normal model to impute a variable that is ordinal (consisting of a 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, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. The ith row of Y, expressed as a column vector (all vectors will be regarded as column vectors), is

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})^T.$$

We assume that $y_1, y_2, \ldots, y_n$ are independent realizations of a random vector, denoted symbolically as $(Y_1, Y_2, \ldots, Y_p)^T$, which has a multivariate normal distribution with mean vector μ and covariance matrix Σ; that is,

$$y_1, y_2, \ldots, y_n \mid \theta \sim \text{iid}\, N(\mu, \Sigma),$$

where $\theta = (\mu, \Sigma)$ is the unknown parameter. Throughout the chapter, we assume no prior restrictions on θ other than the positive definiteness of Σ ($\Sigma > 0$); that is, we allow θ to lie anywhere within its natural parameter space. Because the density of a single row is

$$P(y_i \mid \theta) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right\},$$

the complete-data likelihood is, discarding a proportionality constant,

$$L(\theta \mid Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\right\}. \quad (5.1)$$

Maximum-likelihood estimates

By expanding the exponent in (5.1) and using the fact that

$$\begin{aligned} y_i^T \Sigma^{-1} y_i &= \text{tr}\, y_i^T \Sigma^{-1} y_i \\ &= \text{tr}\, \Sigma^{-1} y_i y_i^T, \end{aligned}$$

it follows that the complete-data loglikelihood can be written as

$$\begin{aligned} l(\theta \mid Y) &= -\tfrac{n}{2}\log|\Sigma| - \tfrac{n}{2}\mu^T \Sigma^{-1}\mu \\ &\quad + \mu^T \Sigma^{-1} T_1 - \tfrac{1}{2}\text{tr}\, \Sigma^{-1} T_2, \end{aligned} \quad (5.2)$$

where

$$T_1 = \sum_{i=1}^{n} y_i = Y^T \mathbf{1}, \quad (5.3)$$

$$T_2 = \sum_{i=1}^{n} y_i y_i^T = Y^T Y \quad (5.4)$$

are the complete-data sufficient statistics, and $\mathbf{1} = (1, 1, \ldots, 1)^T$. Note that T_1 is the vector of column sums,

$$T_1 = \left(\textstyle\sum_{i=1}^n y_{i1},\ \sum_{i=1}^n y_{i2},\ \ldots,\ \sum_{i=1}^n y_{ip}\right)^T,$$

and T_2 is the matrix of columnwise sums of squares and cross-products,

$$T_2 = \begin{bmatrix} \sum_{i=1}^n y_{i1}^2 & \sum_{i=1}^n y_{i1}y_{i2} & \cdots & \sum_{i=1}^n y_{i1}y_{ip} \\ \sum_{i=1}^n y_{i2}y_{i1} & \sum_{i=1}^n y_{i2}^2 & \cdots & \sum_{i=1}^n y_{i2}y_{ip} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n y_{ip}y_{i1} & \sum_{i=1}^n y_{ip}y_{i2} & \cdots & \sum_{i=1}^n y_{ip}^2 \end{bmatrix}.$$

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 frequentist 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 $P(\theta \mid Y) \propto \pi(\theta) L(\theta \mid Y)$ 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 $m \times p$ 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

$$A \sim W(m, \Lambda). \tag{5.7}$$

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 frequentist 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

$$P(B \mid m, \Lambda) \propto |B|^{-\left(\frac{m+p+1}{2}\right)} \exp\left\{ -\tfrac{1}{2} \mathrm{tr} \Lambda^{-1} B^{-1} \right\} \tag{5.8}$$

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

$$E(B \mid m, \Lambda) = \frac{1}{m-p-1} \Lambda^{-1} \tag{5.9}$$

provided that $m \geq p+2$. In the special case of $p = 1$, the inverted-Wishart reduces to a scaled inverted-chisquare, $c\chi_m^{-2}$, 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

$$\mathrm{mode}\,(B \mid m, \Lambda) = \frac{1}{m+p+1} \Lambda^{-1}. \tag{5.10}$$

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 $\theta = (\mu, \Sigma)$ under a multivariate normal model, let us apply the following prior distribution. Suppose that, given Σ, μ is assumed to be conditionally multivariate normal,

$$\mu \mid \Sigma \sim N(\mu_0, \tau^{-1}\Sigma), \tag{5.11}$$

where the hyperparameters $\mu_0 \in \mathcal{R}^p$ and $\tau > 0$ 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

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{m+p+2}{2}\right)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\Lambda^{-1}\Sigma^{-1}\right\} \times \exp\left\{-\tfrac{\tau}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\}. \quad (5.13)$$

Following some matrix algebra, the complete-data likelihood function (5.1) can be rewritten as

$$L(\theta|Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{n}{2}\operatorname{tr}\Sigma^{-1}S\right\} \times \exp\left\{-\tfrac{n}{2}(\bar{y}-\mu)^T\Sigma^{-1}(\bar{y}-\mu)\right\}. \quad (5.14)$$

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 $(\tau, m, \mu_0, \Lambda)$; that is, the complete-data posterior is normal inverted-Wishart,

$$\mu \mid \Sigma, Y \sim N(\mu_0', (\tau')^{-1}\Sigma), \quad (5.15)$$

$$\Sigma \mid Y \sim W^{-1}(m', \Lambda'), \quad (5.16)$$

where the updated hyperparameters are

$$\tau' = \tau + n,$$

$$m' = m + n,$$

$$\mu_0' = \left(\frac{n}{\tau+n}\right)\bar{y} + \left(\frac{\tau}{\tau+n}\right)\mu_0,$$

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

$$\mu \mid \Sigma, Y \sim N(\mu_0', (\tau')^{-1}\Sigma),$$

$$\Sigma \mid Y \sim c'\chi^{-2}_{m'},$$

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 $c = \Lambda^{-1}$ 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 μ 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 $\nu = m - p + 1$ degrees of freedom. The mean of this distribution is μ_0 provided that $\nu > 1$, and the covariance matrix is $(\nu-2)^{-1}\tau^{-1}\Lambda^{-1}$ provided that $\nu > 2$. 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 $\xi = a^T\mu$, where a is a constant vector of length p. The marginal prior distribution of ξ implied by (5.11)–(5.12) is then $(\xi - \xi_0)/\sigma \sim t_\nu$ where $\nu = m - p + 1$, $\xi_0 = a^T\mu_0$, and

$$\sigma = \sqrt{\frac{a^T\Lambda^{-1}a}{\tau\nu}}.$$

The marginal prior density is

$$P(\xi) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left[1 + \frac{(\xi-\xi_0)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2}, \quad (5.17)$$

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 $(\tau, m, \mu_0, \Lambda)$ 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

$$\begin{aligned} \mu \mid \Sigma &\sim N(\mu_0, \tau^{-1}\Sigma), \\ \Sigma &\sim W^{-1}(m, \Lambda). \end{aligned}$$

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

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}, \tag{5.18}$$

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

$$\mu \mid \Sigma, Y \sim N(\bar{y}, n^{-1}\Sigma), \tag{5.19}$$

$$\Sigma \mid Y \sim W^{-1}(n-1, (nS)^{-1}). \tag{5.20}$$

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 $(1-\alpha)100\%$ confidence region for μ from sampling theory, and for inferences about μ the Bayesian and frequentist 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 frequentist 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 $(n-1)^{-1}nS$. 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

$$\mu \mid \Sigma, Y \sim N(\bar{y}, n^{-1}\Sigma), \tag{5.21}$$

$$\Sigma \mid Y \sim W^{-1}(m+n, [\Lambda^{-1}+nS]^{-1}), \tag{5.22}$$

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 $m = \epsilon$ for some $\epsilon > 0$ and $\Lambda^{-1} = \epsilon S^*$ for some covariance matrix S^*, then our rule-of-thumb estimate of Σ is

$$\frac{1}{m+n}(\Lambda^{-1}+nS) = \left(\frac{\epsilon}{n+\epsilon}\right) S^* + \left(\frac{n}{n+\epsilon}\right) S,$$

a weighted average of S and S^* with weights determined by the relative sizes of n and ϵ.

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,

$$S^* = \operatorname{Diag} S. \tag{5.23}$$

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 ϵ and n, and ϵ can be regarded as an imaginary number of prior degrees of freedom added to the inference. Note that ϵ need not be an integer, and in some cases even a small fractional value of ϵ 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 $z^T = (z_1^T, z_2^T)$ where z_1 and z_2 are subvectors of lengths p_1 and $p_2 = p - p_1$, 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 $\mu^T = (\mu_1^T, \mu_2^T)$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

are the partitions of μ and Σ corresponding to $z^T = (z_1^T, z_2^T)$. Moreover, the conditional distributions are also normal; in particular, the distribution of z_2 given z_1 is normal with mean

$$\begin{aligned} E(z_2 \mid z_1) &= \mu_2 + B_{2\cdot 1}(z_1 - \mu_1) \\ &= \alpha_{2\cdot 1} + B_{2\cdot 1} z_1 \end{aligned}$$

and covariance matrix $\Sigma_{22\cdot 1}$, where

$$\begin{aligned} \alpha_{2\cdot 1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \\ B_{2\cdot 1} &= \Sigma_{21}\Sigma_{11}^{-1}, \\ \Sigma_{22\cdot 1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{5.24}$$

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 $\phi_2 = (\alpha_{2\cdot 1}, B_{2\cdot 1}, \Sigma_{22\cdot 1})$. 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

$$\begin{aligned} \mu_2 &= \alpha_{2\cdot 1} + B_{2\cdot 1}\mu_1, \\ \Sigma_{12} &= \Sigma_{11} B_{2\cdot 1}^T, \\ \Sigma_{22} &= \Sigma_{22\cdot 1} + B_{2\cdot 1}\Sigma_{11} B_{2\cdot 1}^T. \end{aligned} \tag{5.25}$$

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 $\alpha_{2\cdot1}$, $B_{2\cdot1}$ and $\Sigma_{22\cdot1} > 0$ will produce a valid $\theta = (\mu, \Sigma)$ with $\Sigma > 0$.

When a probability distribution is applied to $\theta = (\mu, \Sigma)$, 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 $\phi = \phi(\theta)$ induced by f. The relationship between g and f is

$$g(\phi) = f(\phi^{-1}(\phi)) \, ||J||^{-1},$$

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 $\alpha_{2\cdot1}$, $B_{2\cdot1}$ and $\Sigma_{22\cdot1}$ are of the same dimension as μ_2, Σ_{21} and Σ_{22}, respectively, so J can be partitioned as

$$J = \begin{bmatrix} \frac{\partial\mu_1}{\partial\mu_1} & \frac{\partial\mu_1}{\partial\Sigma_{11}} & \frac{\partial\mu_1}{\partial\mu_2} & \frac{\partial\mu_1}{\partial\Sigma_{21}} & \frac{\partial\mu_1}{\partial\Sigma_{22}} \\ \frac{\partial\Sigma_{11}}{\partial\mu_1} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{11}} & \frac{\partial\Sigma_{11}}{\partial\mu_2} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{21}} & \frac{\partial\Sigma_{11}}{\partial\Sigma_{22}} \\ \frac{\partial\alpha_{2\cdot1}}{\partial\mu_1} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{11}} & \frac{\partial\alpha_{2\cdot1}}{\partial\mu_2} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{21}} & \frac{\partial\alpha_{2\cdot1}}{\partial\Sigma_{22}} \\ \frac{\partial B_{2\cdot1}}{\partial\mu_1} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{11}} & \frac{\partial B_{2\cdot1}}{\partial\mu_2} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{21}} & \frac{\partial B_{2\cdot1}}{\partial\Sigma_{22}} \\ \frac{\partial\Sigma_{22\cdot1}}{\partial\mu_1} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{11}} & \frac{\partial\Sigma_{22\cdot1}}{\partial\mu_2} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{21}} & \frac{\partial\Sigma_{22\cdot1}}{\partial\Sigma_{22}} \end{bmatrix},$$

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 $\times$ denotes a matrix that is neither I nor 0. It is a well-known property of determinants that

$$\begin{vmatrix} A & B \\ 0 & C \end{vmatrix} = |A| \, |C| \tag{5.26}$$

for square A and C. Applying (5.26) repeatedly, the determinant of J reduces to

$$|J| = \left| \frac{\partial B_{2\cdot1}}{\partial\Sigma_{21}} \right|. \tag{5.27}$$

With Σ_{11} held fixed, $B_{2\cdot1} = \Sigma_{21}\Sigma_{11}^{-1}$ 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

$$||J|| = |\Sigma_{11}|^{-p_2}. \tag{5.28}$$

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 $p \times p$ symmetric matrix with elements g_{ij}. The sweep operator SWP[k] operates on G by replacing it with another $p \times p$ symmetric matrix H,

$$H = \text{SWP}[k]\, G,$$

where the elements of H are given by

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk} \text{ for } j \neq k, \\ h_{jl} &= h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \text{ for } j \neq k \text{ and } l \neq k. \end{aligned} \tag{5.29}$$

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 $g_{jk} = g_{kj}$ in row and column k with $h_{jk} = -g_{jk}h_{kk}$; 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 $1, 2, \ldots, p_1$, the matrix becomes

$$\mathrm{SWP}[1, 2, \ldots, p_1]\, G = \begin{bmatrix} -G_{11}^{-1} & G_{11}^{-1} G_{12} \\ G_{21} G_{11}^{-1} & G_{22} - G_{21} G_{11}^{-1} G_{12} \end{bmatrix},$$

which is recognizable as a matrix version of (5.29). The notation $\mathrm{SWP}[1, 2, \ldots, p_1]$ indicates successive application of (5.29),

$$\mathrm{SWP}[1, 2, \ldots, p_1]\, G = \mathrm{SWP}[p_1] \cdots \mathrm{SWP}[2]\, \mathrm{SWP}[1]\, G.$$

Sweeps on multiple positions need not be carried out in any particular order, because the sweep operator is commutative,

$$\mathrm{SWP}[k_2]\, \mathrm{SWP}[k_1]\, G = \mathrm{SWP}[k_1]\, \mathrm{SWP}[k_2]\, G.$$

Sweeping a $p \times p$ matrix G on positions $1, 2, \ldots, p$ 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 γ_k denote the kth diagonal element of the matrix after it is swept on positions $1, 2, \ldots, k-1$,

$$\gamma_k = \left(\mathrm{SWP}[1, 2, \ldots, k-1]\, G\right)_{kk}.$$

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

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} &= h_{kj} = -g_{jk}/g_{kk} \text{ for } j \neq k, \\ h_{jl} &= h_{lj} = g_{jl} - g_{jk} g_{kl}/g_{kk} \text{ for } j \neq k \text{ and } l \neq k. \end{aligned} \tag{5.31}$$

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 $z^T = (z_1^T, z_2^T)$ 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 $1, 2, \ldots, p_1$; the result will be, by the matrix analogue of (5.29),

$$\begin{bmatrix} -1 - \mu_1^T \Sigma_{11}^{-1} \mu_1 & \mu_1^T \Sigma_{11}^{-1} & \mu_2^T - \mu_1^T \Sigma_{11}^{-1} \Sigma_{12} \\ \Sigma_{11}^{-1} \mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1} \Sigma_{12} \\ \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1 & \Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \end{bmatrix}.$$

Comparing this to (5.24), we see that the last $p - p_1$ rows and columns contain $\alpha_{2\cdot1}$, $B_{2\cdot1}$, and $\Sigma_{22\cdot1}$, 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 $(p_1+1)\times(p_1+1)$ submatrix contains in swept form the parameters of the marginal distribution of z_1,

$$\begin{bmatrix} -1 & \mu_1^T \\ \mu_1 & \Sigma_{11} \end{bmatrix} = \mathrm{RSW}[1,\ldots,p_1] \begin{bmatrix} -1-\mu_1^T\Sigma_{11}^{-1}\mu_1 & \mu_1^T\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} \end{bmatrix}.$$

We have thus shown that $\phi = (\mu_1, \Sigma_{11}, \alpha_{2\cdot1}, B_{2\cdot1}, \Sigma_{22\cdot1})$, expressed in matrix form as

$$\phi = \begin{bmatrix} -1 & \mu_1^T & \alpha_{2\cdot1}^T \\ \mu_1 & \Sigma_{11} & B_{2\cdot1}^T \\ \alpha_{2\cdot1} & B_{2\cdot1} & \Sigma_{22\cdot1} \end{bmatrix}, \tag{5.33}$$

can be computed from the θ-matrix by first sweeping the full matrix on positions $1,2,\ldots,p_1$, and then reverse-sweeping the upper-left $(p_1+1)\times(p_1+1)$ 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

$$\mathrm{RSW}[0] \begin{bmatrix} -1 & \mu^T \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} 1 & \mu^T \\ \mu & \Sigma+\mu\mu^T \end{bmatrix}, \tag{5.34}$$

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 $T_1 = Y^T\mathbf{1}$ and $T_2 = Y^TY$ into a $(p+1)\times(p+1)$ matrix

$$T = [\mathbf{1},Y]^T[\mathbf{1},Y] = \begin{bmatrix} n & T_1^T \\ T_1 & T_2 \end{bmatrix}, \tag{5.35}$$

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

$$\hat\theta = \mathrm{SWP}[0]\,n^{-1}T.$$

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.

	variables				
	Y_1	Y_2	Y_3	$\cdots$	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 $s = 1,2,\ldots,S$, 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 r_{sj}, where

$$r_{sj} = \begin{cases} 1 & \text{if } Y_j \text{ is observed in pattern } s, \\ 0 & \text{if } Y_j \text{ is missing in pattern } s. \end{cases}$$

The matrix R is shown in Figure 5.1. For each missingness pattern s, let $\mathcal{O}(s)$ and $\mathcal{M}(s)$ denote the subsets of the column labels $\{1, 2, \ldots, p\}$ corresponding to variables that are observed and missing, respectively,

$$\begin{aligned} \mathcal{O}(s) &= \{j : r_{sj} = 1\}, \\ \mathcal{M}(s) &= \{j : r_{sj} = 0\}. \end{aligned}$$

Finally, let $\mathcal{I}(s)$ denote the subset of $\{1, 2, \ldots, n\}$ 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 $1, \ldots, 10$; the first row of R is then $(1, 1, \ldots, 1)$, and

$$\begin{aligned} \mathcal{O}(1) &= \{1, 2, \ldots, p\}, \\ \mathcal{M}(1) &= \emptyset, \\ \mathcal{I}(1) &= \{1, 2, \ldots, 10\}. \end{aligned}$$

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 $P(Y_{mis} \mid Y_{obs}, \theta)$ for an assumed value of θ. These statistics are of the form $\sum_i y_{ij}$ and $\sum_i y_{ij} y_{ik}$, so to perform the E-step we need to find the expectations of y_{ij} and $y_{ij} y_{ik}$ over $P(Y_{mis} \mid Y_{obs}, \theta)$.

Because the rows $y_1, y_2, \ldots, y_n$ of Y are independent given θ, we can write

$$P(Y_{mis} \mid Y_{obs}, \theta) = \prod_{i=1}^{n} P(y_{i(mis)} \mid y_{i(obs)}, \theta),$$

where $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing subvectors of y_i, respectively. The distribution $P(y_{i(mis)} \mid y_{i(obs)}, \theta)$ 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

$P(y_{i(mis)} \mid y_{i(obs)}, \theta)$ are contained in $\mathrm{SWP}[\mathcal{O}(s)]\,\theta$ in the rows and columns labeled $\mathcal{M}(s)$. Let A denote the swept parameter matrix

$$A = \mathrm{SWP}[\mathcal{O}(s)]\,\theta,$$

and let a_{jk} denote the (j,k)th element of A, $j, k = 0, 1, \ldots, p$. Using the results of Section 5.2.4, the reader may verify that the first two moments of $y_{i(mis)}$ with respect to $P(Y_{mis} \mid Y_{obs}, \theta)$ are given by

$$\begin{aligned} E(y_{ij} \mid Y_{obs}, \theta) &= a_{0j} + \sum_{k \in \mathcal{O}(s)} a_{kj} y_{ik}, \\ \mathrm{Cov}(y_{ij}, y_{ik} \mid Y_{obs}, \theta) &= a_{jk} \end{aligned}$$

for each $i \in \mathcal{I}(s)$ and $j, k \in \mathcal{M}(s)$. For any $j \in \mathcal{O}(s)$, of course, the moments are

$$\begin{aligned} E(y_{ij} \mid Y_{obs}, \theta) &= y_{ij}, \\ \mathrm{Cov}(y_{ij}, y_{ik} \mid Y_{obs}, \theta) &= 0, \end{aligned}$$

because y_{ij} is regarded as fixed. Applying the relation

$$\begin{aligned} E(y_{ij} y_{ik} \mid Y_{obs}, \theta) &= \mathrm{Cov}(y_{ij}, y_{ik} \mid Y_{obs}, \theta) \\ &\quad + E(y_{ij} \mid Y_{obs}, \theta)\, E(y_{ik} \mid Y_{obs}, \theta), \end{aligned}$$

it follows that

$$E(y_{ij} \mid Y_{obs}, \theta) = \begin{cases} y_{ij} & \text{for } j \in \mathcal{O}(s), \\ y^*_{ij} & \text{for } j \in \mathcal{M}(s), \end{cases}$$

and

$$E(y_{ij} y_{ik} \mid Y_{obs}, \theta) = \begin{cases} y_{ij} y_{ik} & \text{for } j, k \in \mathcal{O}(s), \\ y^*_{ij} y_{ik} & \text{for } j \in \mathcal{M}(s), k \in \mathcal{O}(s), \\ a_{jk} + y^*_{ij} y^*_{ik} & \text{for } j, k \in \mathcal{M}(s), \end{cases}$$

where

$$y^*_{ij} = a_{0j} + \sum_{k \in \mathcal{O}(s)} a_{kj} y_{ik}. \tag{5.36}$$

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 $E(T \mid Y_{obs}, \theta)$, 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 $A = \text{SWP}[\mathcal{O}(s)]\,\theta$ needed for the E-step depends on the missingness pattern s, and thus in practice the elements of $E(T \mid Y_{obs}, \theta)$ must be calculated by first summing expected values of y_{ij} and $y_{ij}y_{ik}$ for $i \in \mathcal{I}(s)$, and then summing across patterns $s = 1, 2, \ldots, S$, with a new A-matrix being calculated for each missingness pattern.

5.3.3 *Implementation of the algorithm*

Once $E(T \mid Y_{obs}, \theta)$ 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 $E(T \mid Y_{obs}, \theta)$ rather than T. A single iteration of EM can thus be written succinctly as

$$\theta^{(t+1)} = \text{SWP}[0]\,n^{-1}E(T \mid Y_{obs}, \theta^{(t)}). \tag{5.37}$$

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 tradeoffs 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 $E(T \mid Y_{obs}, \theta)$ 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

$$T(s) = \begin{bmatrix} n_s & \sum y_{i1} & \sum y_{i2} & \cdots & \sum y_{ip} \\ & \sum y_{i1}^2 & \sum y_{i1}y_{i2} & \cdots & \sum y_{i1}y_{ip} \\ & & \sum y_{i2}^2 & \cdots & \sum y_{i2}y_{ip} \\ & & & \ddots & \vdots \\ & & & & \sum y_{ip}^2 \end{bmatrix},$$

where all sums are taken over $i \in \mathcal{I}(s)$, and $n_s = \sum_{i \in \mathcal{I}(s)} 1$ 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 $\mathcal{M}(s)$ 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 $\mathcal{M}(s)$, 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

$$T_{obs}(s) = \begin{bmatrix} n_s & \sum y_{i1} & 0 & \sum y_{i3} \\ & \sum y_{i1}^2 & 0 & \sum y_{i1}y_{i3} \\ & & 0 & 0 \\ & & & \sum y_{i3}^2 \end{bmatrix},$$

$$T_{mis}(s) = \begin{bmatrix} 0 & 0 & \sum y_{i2} & 0 \\ & 0 & \sum y_{i1}y_{i2} & 0 \\ & & \sum y_{i2}^2 & \sum y_{i2}y_{i3} \\ & & & 0 \end{bmatrix},$$

where all sums are taken over $i \in \mathcal{I}(s)$. Finally, define

$$T_{obs} = \sum_{s=1}^{S} T_{obs}(s) \quad \text{and} \quad T_{mis} = \sum_{s=1}^{S} T_{mis}(s),$$

```
T := T_obs
for s := 1 to S do
    for j := 1 to p do
        if r_sj = 1 and θ_jj > 0 then θ := SWP[j] θ
        if r_sj = 0 and θ_jj < 0 then θ := RSW[j] θ
        end do
    for i ∈ I(s) do
        for j ∈ M(s) do
            c_j := θ_0j
            for k ∈ O(s) do c_j := c_j + θ_kj y_ik
            end do
        for j ∈ M(s) do
            T_0j := T_0j + c_j
            for k ∈ O(s) do T_kj := T_kj + c_j y_ik
            for k ∈ M(s) and k ≥ j do T_kj := T_kj + θ_kj + c_k c_j
            end do
        end do
    end do
θ := 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 $c = (c_1, \ldots, c_p)$, is needed as a temporary workspace to hold the values of y_{ij}^* 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[$\mathcal{O}(s)$] 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 $j \leq k$ or the (k,j)th element if $j > k$. Memory requirements can be reduced by retaining only the upper-triangular parts of T and θ 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 $\mu^{(0)} = (0, 0, \ldots, 0)^T$ 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 $P(\theta|Y_{obs})$ 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 frequentist inferences about μ 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 frequentist 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 $Y_1, Y_2, \ldots, Y_p$ 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 Σ 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 $\tau = 0$, $m = \epsilon$ for some $\epsilon > 0$, and $\Lambda^{-1} = \epsilon \,\mathrm{Diag}\, S$, 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 Σ 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 $\Lambda^{-1} = \epsilon I$.

Modifications to the M-step

The joint mode of the normal inverted-Wishart distribution,

$$\begin{aligned} \mu \mid \Sigma &\sim N(\mu_0, \tau^{-1}\Sigma), \\ \Sigma &\sim W^{-1}(m, \Lambda), \end{aligned}$$

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 $\tilde{\theta} = (\tilde{\mu}, \tilde{\Sigma})$, is

$$\tilde{\mu} = \mu_0' \quad \text{and} \quad \tilde{\Sigma} = \frac{1}{m' + p + 2}\,(\Lambda')^{-1},$$

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,

$$\mathrm{RSW}[0] \begin{bmatrix} -1 & \tilde{\mu}^T \\ \tilde{\mu} & \tilde{\Sigma} \end{bmatrix} = \begin{bmatrix} 1 & \tilde{\mu}^T \\ \tilde{\mu} & \tilde{\Sigma} + \tilde{\mu}\tilde{\mu}^T \end{bmatrix} = n^{-1} \begin{bmatrix} n & \tilde{T}_1^T \\ \tilde{T}_1 & \tilde{T}_2 \end{bmatrix},$$

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

$$\tilde{T}_2 = \frac{n}{n+m+p+2}\left(T_2 - \frac{1}{n}T_1T_1^T + \Lambda^{-1} + A\right) + \frac{1}{n}\tilde{T}_1\tilde{T}_1^T$$

as the modified sufficient statistics, where

$$A = \frac{\tau}{n(\tau+n)}\,(T_1 - n\mu_0)(T_1 - n\mu_0)^T.$$

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 $\theta := \mathrm{SWP}[0]\, n^{-1}T$.

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 $l(\theta \mid Y_{obs})$, would be very tedious to differentiate or maximize by gradient-based methods. Evaluation of $l(\theta \mid Y_{obs})$ 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

$$\sum_{s=1}^{S} \sum_{i \in \mathcal{I}(s)} \left\{ -\tfrac{1}{2} \log|\Sigma_s^*| - \tfrac{1}{2}\,(y_{i(obs)} - \mu_s^*)^T \Sigma_s^{*-1} (y_{i(obs)} - \mu_s^*) \right\},$$

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

$$l(\theta|Y_{obs}) = \sum_{s=1}^{S} \left\{ -\tfrac{n_s}{2} \log|\Sigma_s^*| - \tfrac{1}{2} \operatorname{tr} \Sigma_s^{*-1} M_s \right\}, \qquad (5.38)$$

where n_s is the number of observations in missingness pattern s and

$$M_s = \sum_{i \in \mathcal{I}(s)} (y_{i(obs)} - \mu_s^*)(y_{i(obs)} - \mu_s^*)^T.$$

```
d:=0
l:=0
for j:= 1 to p do c_j:=θ_0j
for s:=1 to S do
    for j:= 1 to p do
        if r_sj=1 and θ_jj>0 then
            d:=d+log θ_jj
            θ:=SWP[j]θ
        else if r_sj=0 and θ_jj<0 then
            θ:=RSW[j]θ
            d:=d-log θ_jj
            end if
        end do
    M:=0
    for i∈I(s), j,k∈O(s) and j≤k do
        M_jk:=M_jk+(y_ij-c_j)(y_ik-c_k)
        end do
    t:=0
    for j,k∈O(s) do t:=t-θ_jk M_jk
    l:=l-(n_s d+t)/2
    end do
```

Figure 5.3. *Calculation of observed-data loglikelihood function.*

Pseudocode for calculating $l(\theta \mid Y_{obs})$ is shown in Figure 5.3. This algorithm requires a $p \times p$ matrix workspace M to hold values of M_s, and a $p \times 1$ vector c for temporary storage of μ. The constants d and t hold $\log|\Sigma_s^*|$ and $\operatorname{tr} \Sigma_s^{*-1} M_s$, respectively, and after execution the loglikelihood value is contained in l. 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 $l(\theta \mid Y_{obs})$ 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 $l(\theta^{(t)} \mid Y_{obs})$ 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

$$\log \pi(\theta) \;=\; -\tfrac{m+p+2}{2} \log|\Sigma| - \tfrac{1}{2} \operatorname{tr} \Sigma^{-1} M_0,$$

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

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

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 σ_{jk}, respectively, for $j, k = 1, 2, 3$, and let $\rho_{jk} = \sigma_{jk}(\sigma_{jj}\sigma_{kk})^{-1/2}$ denote the correlations. From starting values chosen based on a crude guess, $\mu^{(0)} = (200, 200, 200)^T$ and $\Sigma^{(0)} = (50)^2 I$, convergence within four significant digits to

$$\hat{\mu} = \begin{bmatrix} 253.9 \\ 230.6 \\ 222.2 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 2195 & 1455 & 835.4 \\ & 2127 & 1515 \\ & & 1952 \end{bmatrix}$$

was achieved in just three iterations. Because no data are 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).

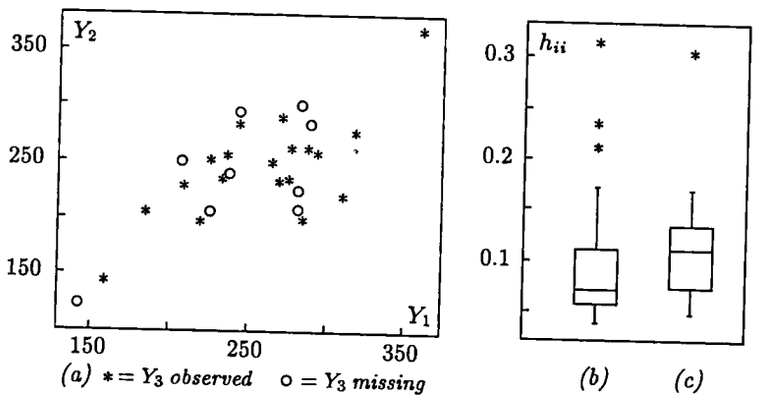

Figure 5.4. *(a) Scatterplot of Y_1 versus Y_2 for all cases, and boxplots of leverage values h_{ii} for cases having (b) Y_3 observed and (c) Y_3 missing.*

The hat matrix for linear regression is defined to be

$$H = X(X^TX)^{-1}X^T,$$

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 $\mathbf{1} = (1, 1, \ldots,)^T$. Boxplots of the diagonal elements h_{ii} 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

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)

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.

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 ex-

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

ample? 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

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

and the P-step simulates

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

Because the rows $y_1, y_2, \ldots, y_n$ of Y are conditionally independent given θ, the I-step is carried out by drawing

$$y_{i(mis)}^{(t+1)} \sim P(y_{i(mis)} | y_{i(obs)}, \theta^{(t)})$$

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

$$E(y_{ij} | Y_{obs}, \theta) = a_{0j} + \sum_{k \in \mathcal{O}(s)} a_{kj} y_{ik} \tag{5.39}$$

and covariances

$$\mathrm{Cov}(y_{ij}, y_{ik} | Y_{obs}, \theta) = a_{jk} \tag{5.40}$$

for $j, k \in \mathcal{M}(s)$, where a_{jk} denotes an element of the matrix

$$A = \mathrm{SWP}[\mathcal{O}(s)]\,\theta. \tag{5.41}$$

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 ∈ S do
    a_ii := ( a_ii − Σ_{k∈S,k<i} a_ki^2 )^{1/2}
    for j ∈ S, j > i do
        a_ij := a_ii^{-1} ( a_ij − Σ_{k∈S,k<i} a_ki a_kj )
    end do
end do
```

Figure 5.5. *Calculation of* $A := \mathrm{Chol}_{\mathcal{S}} A$.

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 $C^T C = A$. 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 $\mathcal{M}(s)$. For a set $\mathcal{S}$ of row labels of a matrix A, let us use

$$A := \mathrm{Chol}_{\mathcal{S}} A \tag{5.42}$$

to indicate the operation that overwrites (the upper triangular portion of) the square submatrix $\{a_{jk} : j, k \in \mathcal{S}\}$ 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 $s = 1, \ldots, S$, calculating

$$\mathrm{Chol}_{\mathcal{M}(s)}\ \mathrm{SWP}[\mathcal{O}(s)]\,\theta$$

for each s, and simulating $y_{i(mis)}$ for each $i \in \mathcal{I}(s)$. 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 θ_jj > 0 then θ := SWP[j] θ
           if r_sj = 0 and θ_jj < 0 then θ := RSW[j] θ
           end do
       C := Chol_{M(s)} θ
       for i ∈ I(s) do
           for j ∈ M(s) do
               y_ij := θ_0j
               for k ∈ O(s) do y_ij := y_ij + θ_kj y_ik
               draw z_j ~ N(0,1)
               for k ∈ M(s) and k ≤ j do y_ij := y_ij + C_kj z_k
C              T_0j := T_0j + y_ij
C              for k ∈ O(s) do T_kj := T_kj + y_ij y_ik
C              for k ∈ M(s) and k ≤ 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 $P(\theta \mid Y_{obs}, Y_{mis})$ is a normal inverted-Wishart distribution. The P-step of data augmentation, therefore,

is merely a simulation of the normal inverted-Wishart distribution,

$$\begin{aligned} \mu \mid \Sigma &\sim N(\mu_0, \tau^{-1}\Sigma), \\ \Sigma &\sim W^{-1}(m, \Lambda), \end{aligned}$$

for some $(\tau, m, \mu_0, \Lambda)$ 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 $(\tau, m, \mu_0, \Lambda)$ 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 $\Sigma = (X^TX)^{-1}$, 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 $A = B^TB$ where B is an upper-triangular matrix whose elements are independently distributed as

$$b_{jj} \sim \sqrt{\chi^2_{m-j+1}}, \quad j = 1, \ldots, p, \tag{5.43}$$

$$b_{jk} \sim N(0,1), \quad j < k. \tag{5.44}$$

Suppose that we generate an upper-triangular matrix B according to (5.43)–(5.44), so that $B^TB \sim W(m, I)$, and take

$$M = (B^T)^{-1}C,$$

where C is the Cholesky factor of Λ^{-1} (i.e. $C^TC = \Lambda^{-1}$). Then $\Sigma = M^TM$ will be distributed as $W^{-1}(m, \Lambda)$, because

$$\begin{aligned} (M^TM)^{-1} &= C^{-1}B^TB(C^T)^{-1} \\ &\sim W(m, (C^TC)^{-1}). \end{aligned}$$

(Here we have made use of the property that $D \sim W(n, \Gamma)$ implies $C^TDC \sim W(n, C^T\Gamma C)$, which follows immediately from the definition of the Wishart distribution.) Moreover, taking

$$\mu = \mu_0 + \tau^{-1/2}M^Tz,$$

where $z \sim N(0, I)$ is a $p \times 1$ vector of independent standard normal variates, results in $\mu \mid \Sigma \sim N(\mu_0, \tau^{-1}\Sigma)$. 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 $\sigma_3 = \sqrt{\sigma_{33}}$, 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

$$\begin{bmatrix} \beta_{31\cdot 12} \\ \beta_{32\cdot 12} \end{bmatrix}^T = \begin{bmatrix} \sigma_{31} & \sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1},$$

the intercept

$$\beta_{30\cdot 12} = \mu_3 - \begin{bmatrix} \sigma_{31} & \sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

and the residual standard deviation $\sigma_{3\cdot 12} = \sqrt{\sigma_{33\cdot 12}}$, where

$$\sigma_{33\cdot 12} = \sigma_{33} - \begin{bmatrix} \sigma_{31} & \sigma_{32} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix};$$

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	$\beta_{30\cdot12}$	$\beta_{31\cdot12}$	$\beta_{32\cdot12}$	$\sigma_{3\cdot12}$	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 $\mu = (\mu_1, \mu_2, \mu_3)^T$. For illustration, we will focus attention on three quantities: μ_3, the average cholesterol level at 14 days;

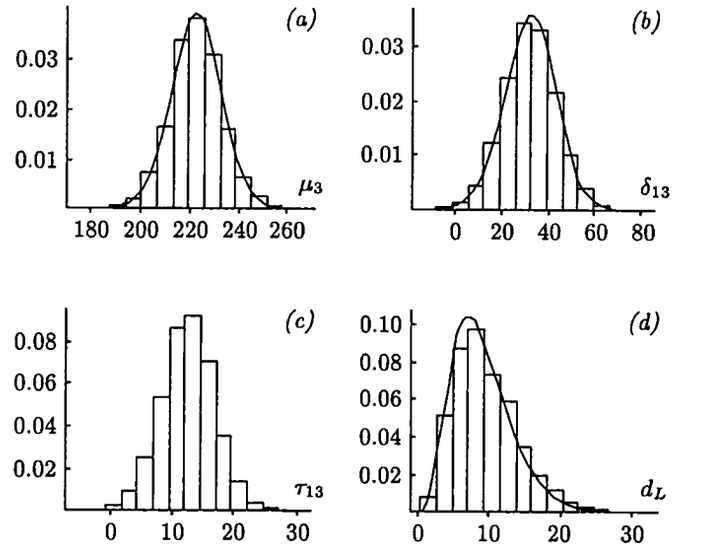

Figure 5.7. *Histograms of sample values of (a) μ_3, (b) δ_{13}, (c) τ_{13} and (d) d_L from 5000 consecutive iterations of data augmentation.*

$\delta_{13} = \mu_1 - \mu_3$, the average decrease in cholesterol level from day 2 to day 14; and $\tau_{13} = 100(\mu_1 - \mu_3)/\mu_1$, 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

$$P(\eta|Y_{obs}, Y_{mis}) = k\left[1 + \frac{(\eta - a^T\bar{y})^2}{(n-p)\sigma^2}\right]^{-(n-p+1)/2}, \qquad (5.45)$$

where $n = 28$ and $p = 3$ are the number of observations and variables, respectively; $\sigma^2 = (n-p)^{-1}a^TSa$; $\bar{y}$ and S are the sample mean vector (5.5) and covariance matrix (5.6) computed from

$Y = (Y_{obs}, Y_{mis})$; 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 $\mu_3 = (0,0,1)\mu$ and $\delta_{13} = (1,0,-1)\mu$ estimated from the first 1000 iterations after the 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

$$d_L = d_L(\theta) = 2[l(\hat{\theta}|Y_{obs}) - l(\theta|Y_{obs})]$$

over the 5000 iterations, where $\hat{\theta}$ is the ML estimate. For large samples, the posterior distribution of d_L is approximately χ^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 $\delta_{jk} = \mu_j - \mu_k$, $j,k = 1,\ldots,6$. 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,

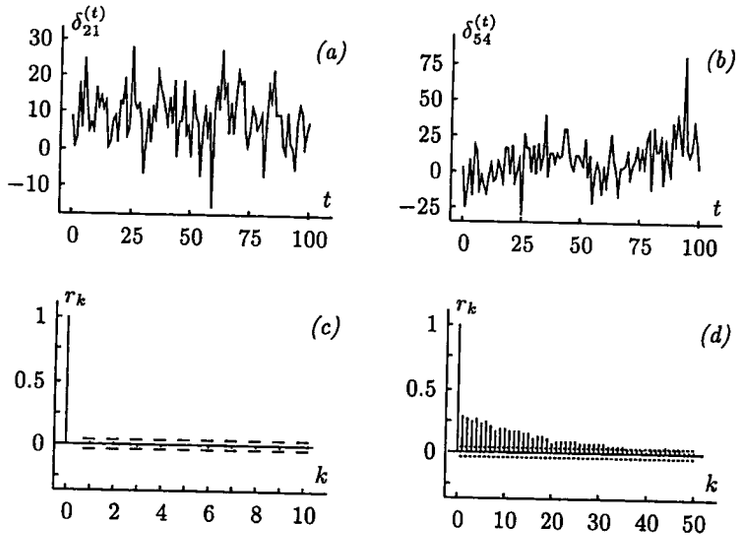

Figure 5.8. *Time-series plots of (a) δ_{21} and (b) δ_{54} over the first 100 iterations of data augmentation, and sample ACFs for (c) δ_{21} and (d) δ_{54} estimated from 10 000 iterations, with dashes indicating approximate 0.05-level critical values for testing $\rho_k = \rho_{k+1} = \cdots = 0$.*

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 $\hat{\theta}$ 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 δ_{jk} 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

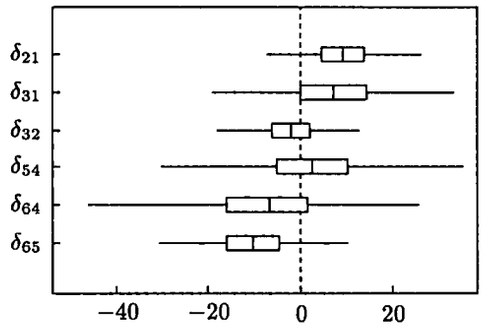

Figure 5.9. *Simulated posterior medians, quartiles and 95% equal-tailed intervals for six contrasts.*

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 δ_{jk} 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;

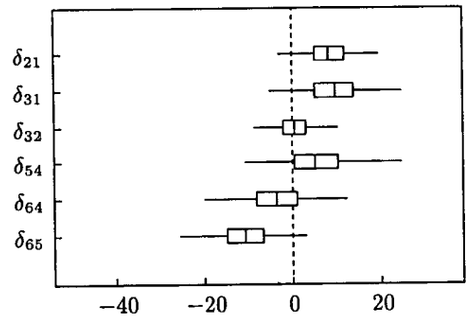

Figure 5.10. *Simulated posterior medians, quartiles and 95% equal-tailed intervals for six contrasts using a bivariate approach.*

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, $P(Y_{mis} \mid Y_{obs})$. 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 requires 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

$$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \quad \text{and} \quad S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k)$$

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

$$\begin{aligned} \hat{Q} &= \bar{y}_1 - \bar{y}_3, \\ U &= (S_{11} - 2S_{13} + S_{33})/n. \end{aligned}$$

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 $\tau_{13} = 100(\mu_1 - \mu_3)/\mu_1$, a first-order Taylor expansion of the function $(\bar{y}_1 - \bar{y}_3)/\bar{y}_1$ about (μ_1, μ_3),

$$\frac{\bar{y}_1 - \bar{y}_3}{\bar{y}_1} - \frac{\mu_1 - \mu_3}{\mu_1} \approx \frac{\mu_3}{\mu_1^2}(\bar{y}_1 - \mu_1) - \frac{1}{\mu_1}(\bar{y}_3 - \mu_3),$$

suggests that the complete-data point estimate

$$\hat{Q} = 100(\bar{y}_1 - \bar{y}_3)/\bar{y}_1$$

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 $(47.7/\sqrt{28})/253.9 = 0.036$, 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 $\bar{Q}$, standard errors $\sqrt{T}$, degrees of freedom ν for the t-approximation and 95% interval estimates are displayed in Table 6.3. The values of ν 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

Table 6.3. *Results of multiple-imputation inference for* μ_1, 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

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 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

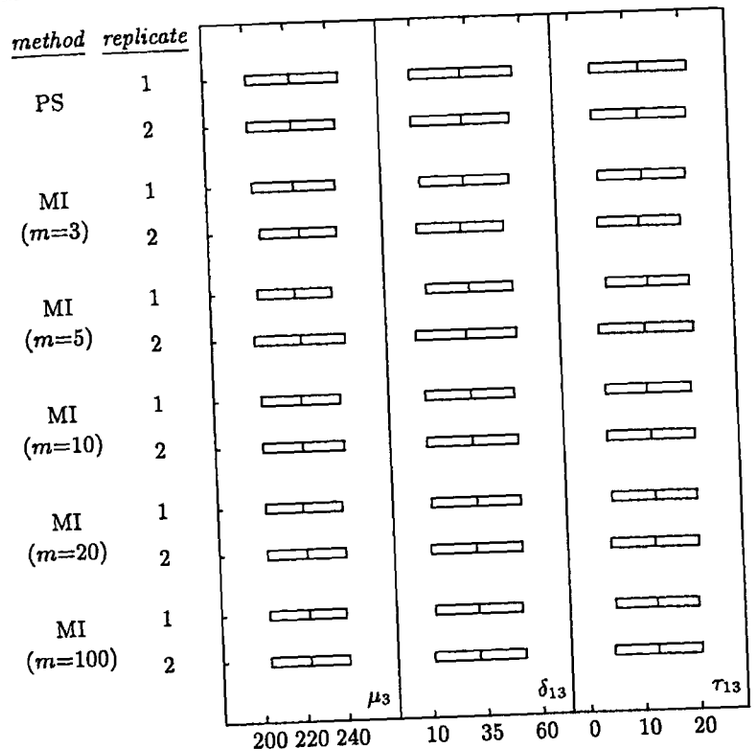

Figure 6.1. *Point and 95% interval estimates for* μ_3, δ_{13} *and* τ_{13} *from parameter simulation (PS) and multiple imputation (MI).*

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*

m	replicate	Parameter 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 ($10 \times 9 = 90$ locations to hold imputations of Y_{mis}, versus $5000 \times 3 = 15\,000$ 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 λ 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.

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-

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

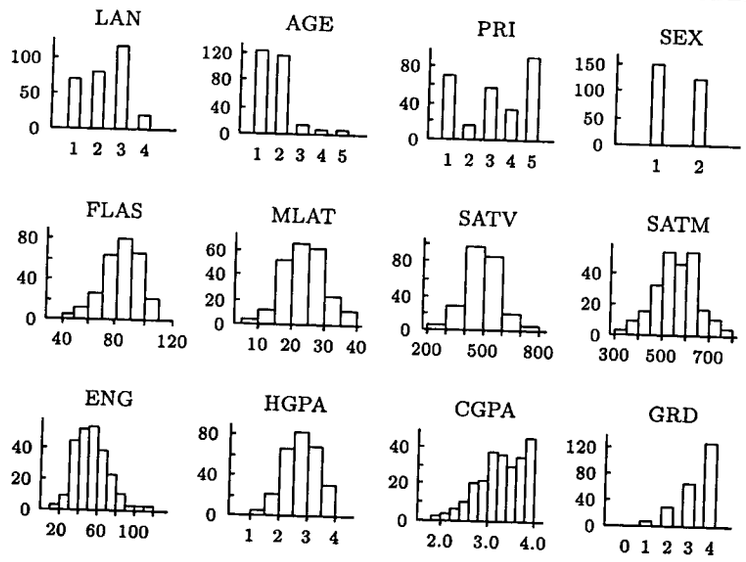

Figure 6.2. *Histograms of observed data for variables in foreign language achievement study.*

third of the entire dataset. Imputing for the missing values makes more efficient use of the available data.

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

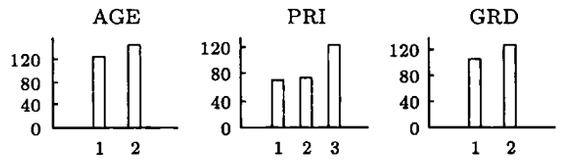

Figure 6.3. *Histograms of observed data for AGE, PRI and GRD after recoding.*

Table 6.6. *Definitions for AGE, PRI and GRD after recoding*

Variable	*Description*	*Missing*
AGE	age group (1=less than 20, 2=20+)	11
PRI	Number of prior foreign language courses (1=none, 2=1–2, 3=3+)	11
GRD	final grade in foreign language course (2=A, 1=B or lower)	47

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: $LAN_2 = 1$ if Spanish and 0 otherwise, $LAN_3 = 1$ if German and 0 otherwise, and $LAN_4 = 1$ 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

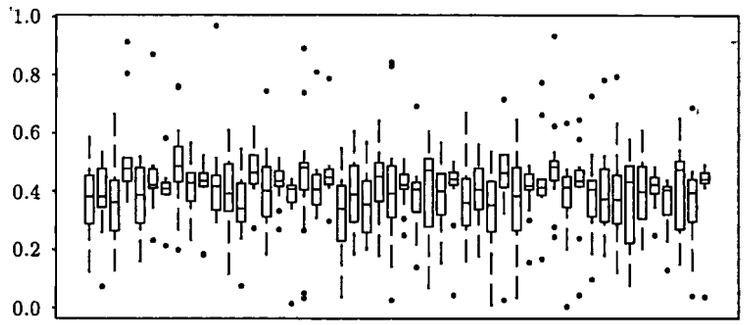

Figure 6.4. *Boxplots of estimated elementwise rates of convergence for 50 randomly selected parameters.*

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

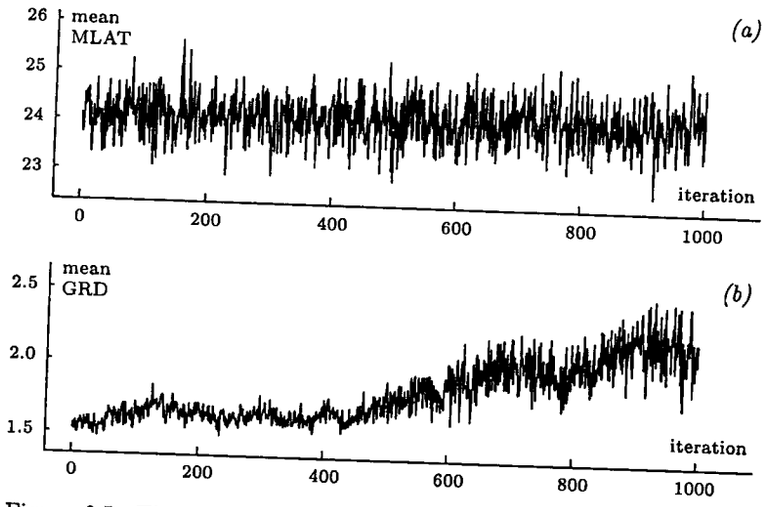

Figure 6.5. *Time series plots of (a) mean MLAT and (b) mean GRD over 1000 iterations of data augmentation.*

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

converging to any stationary distribution (Section 3.5.2).

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 $\tau = 0$, $m = \epsilon$, and $\Lambda^{-1} = \epsilon I$ for $\epsilon = 3$. This prior adds the equivalent of three degrees of freedom to the estimation of Σ and smooths the estimated correlation matrix toward I. 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

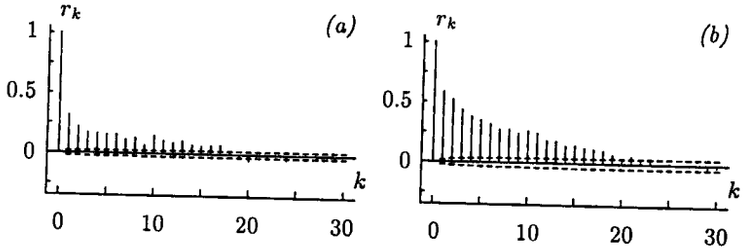

Figure 6.6. *Sample ACFs for (a) mean GRD and (b) the worst linear function of θ, estimated from 1000 iterations of data augmentation, with dashed lines indicating approximate critical values for testing $\rho_k = \rho_{k+1} = \cdots = 0$.*

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.

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 ($\text{AGE}_2 = 1$ if 20+ and 0 otherwise); an indicator for sex ($\text{SEX}_2 = 1$ if female and 0 otherwise); linear and quadratic contrasts for PRI ($\text{PRI}_\text{L} = -1, 0, 1$ and $\text{PRI}_\text{Q} =$

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

$1, -2, 1$ for PRI $= 1, 2, 3$, respectively); 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 $\bar{Q}$ and standard error $\sqrt{T}$, the t-statistic $\bar{Q}/\sqrt{T}$, the degrees of freedom ν 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

$$l(\beta|Y_{obs},Y_{mis}) = \sum_{i=1}^{n}\left[z_i \log\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} + (1-z_i)\log\frac{1}{1+e^{x_i^T\beta}}\right],$$

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 $\tilde{\beta}$, 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

$$d_L(\hat{\beta},\tilde{\beta}|Y_{obs},Y_{mis}) = 2[l(\hat{\beta}|Y_{obs},Y_{mis}) - l(\tilde{\beta}|Y_{obs},Y_{mis})],$$

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,

$$\bar{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L(\hat{\beta}^{(t)},\tilde{\beta}^{(t)}|Y_{obs},Y_{mis}^{(t)}).$$

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,

$$\tilde{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L(m^{-1}\textstyle\sum_{t=1}^{m}\hat{\beta}^{(t)}, m^{-1}\sum_{t=1}^{m}\tilde{\beta}^{(t)}|Y_{obs},Y_{mis}^{(t)}).$$

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 ν_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.

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

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 Third National Health and Nutrition Examination Survey (NHANES III) (NCHS,

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)

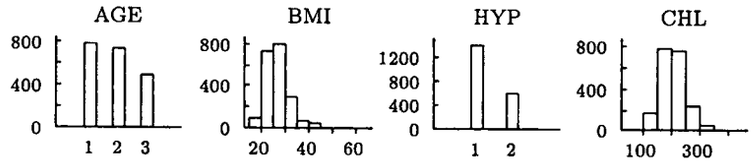

Figure 6.7. *Histograms of AGE, BMI, HYP and CHL in the population.*

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 NHANES 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

$$Q_1 \leq Q \leq Q_2$$

will be true if and only if

$$F(Q_1) \leq p \leq F(Q_2),$$

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

$$p_1 = p - 2\sqrt{\frac{p(1-p)}{n}} \quad \text{to} \quad p_2 = p + 2\sqrt{\frac{p(1-p)}{n}}.$$

If we set Q_1 and Q_2 equal to the p_1th and p_2th 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

$$\omega = \frac{P(Y_1 = 1|Y_2 = 1)/P(Y_1 = 2|Y_2 = 1)}{P(Y_1 = 1|Y_2 = 2)/P(Y_1 = 2|Y_2 = 2)},$$

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 $x_{11}^{-1} + x_{12}^{-1} + x_{21}^{-1} + x_{22}^{-1}$ (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 $\hat{\lambda}$ 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 $\bar{Q}$ 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

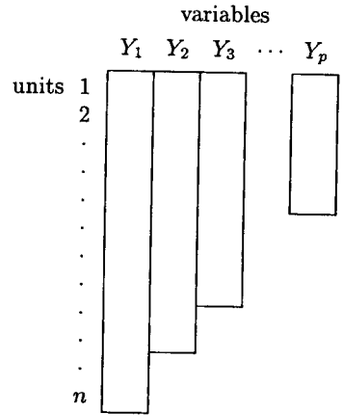

Figure 6.8. *Monotone missingness pattern.*

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 $Y_1, \ldots, Y_{j-1}$. 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

$$\begin{aligned} P(Y_1, \ldots, Y_p \,|\, \phi) &= P(Y_1 \,|\, \phi_1)\, P(Y_2 \,|\, Y_1, \phi_2) \\ &\quad \cdots P(Y_p \,|\, Y_1, \ldots, Y_{p-1}, \phi_p), \end{aligned}$$

which allows us to write the complete-data likelihood as

$$\begin{aligned} L(\phi|Y) &= \prod_{i=1}^{n} P(y_{i1}, \ldots, y_{ip} \,|\, \phi) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{p} P(y_{ij} \,|\, y_{i1}, \ldots, y_{i,j-1}, \phi_j) \\ &= \prod_{j=1}^{p} \prod_{i=1}^{n} P(y_{ij} \,|\, y_{i1}, \ldots, y_{i,j-1}, \phi_j). \end{aligned} \tag{6.3}$$

The inner product in (6.3),

$$\prod_{i=1}^{n} P(y_{ij}\,|\,y_{i1},\ldots,y_{i,j-1},\phi_j),$$

can also be written

$$\prod_{i=1}^{n_j} P(y_{ij}\,|\,y_{i1},\ldots,y_{i,j-1},\phi_j)\prod_{i=n_j+1}^{n} P(y_{ij}\,|\,y_{i1},\ldots,y_{i,j-1},\phi_j). \quad (6.4)$$

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 $1,\ldots,n_j$, whereas the second product integrates to unity because Y_j is missing in rows $n_j+1,\ldots,n$. It follows that

$$L(\phi\,|\,Y_{obs}) = \prod_{j=1}^{p} L(\phi_j\,|\,Y_{obs}), \quad (6.5)$$

where

$$L(\phi_j\,|\,Y_{obs}) = \prod_{i=1}^{n_j} P(y_{ij}\,|\,y_{i1},\ldots,y_{i,j-1},\phi_j). \quad (6.6)$$

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, $\mathrm{SWP}_A[k]$ for some $k \in A$ operates on G by replacing it with another $p \times p$ matrix H,

$$H = \mathrm{SWP}_A[k]\,G,$$

where the elements of H are given by

$$h_{kk} = -1/g_{kk},$$

$$h_{jk} = h_{kj} = \begin{cases} g_{jk}/g_{kk} & j \in A,\ j \neq k, \\ g_{jk} & j \notin A, \end{cases}$$

$$h_{jl} = h_{lj} = \begin{cases} g_{jl} - g_{jk}g_{kl}/g_{kk} & j \in A,\ l \in A,\ j \neq k,\ l \neq k, \\ g_{jl} & j \notin A \text{ or } l \notin A. \end{cases}$$

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, $\mathrm{RSW}_A[k]$ for some $k \in A$ operates on G by replacing it with another $p \times p$ matrix H,

$$H = \mathrm{RSW}_A[k]\,G,$$

where the elements of H are given by

$$h_{kk} = -1/g_{kk},$$

$$h_{jk} = h_{kj} = \begin{cases} -g_{jk}/g_{kk} & j \in A,\ j \neq k, \\ g_{jk} & j \notin A, \end{cases}$$

$$h_{jl} = h_{lj} = \begin{cases} g_{jl} - g_{jk}g_{kl}/g_{kk} & j \in A,\ l \in A,\ j \neq k,\ l \neq k, \\ g_{jl} & j \notin A \text{ or } l \notin A. \end{cases}$$

When the sweep or reverse sweep operators are written without a subscripting set, as in $\mathrm{SWP}[k]$ or $\mathrm{RSW}[k]$, 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

$$\phi_j = (\beta_j^T, \gamma_j)^T \quad (6.7)$$

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

$$Y_j \,|\, Y_1, \ldots, Y_{j-1}, \phi_j \sim N(\,(1, Y_1, \ldots, Y_{j-1})\,\beta_j,\, \gamma_j\,). \tag{6.8}$$

Let μ_j denote the jth element of μ and σ_{jk} the (j,k)th element of Σ, and define

$$\theta_j = (\mu_j, \sigma_{1j}, \sigma_{2j}, \ldots, \sigma_{jj})^T, \tag{6.9}$$

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 $1, 2, \ldots, j-1$ produces a new matrix whose jth column is ϕ_j. Therefore, if we sweep the full θ matrix on positions $1, 2, \ldots, p-1$, then ϕ_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 $A_j = \{0, 1, \ldots, j\}$ for $j = 1, 2, \ldots, p$. The following two lines will overwrite a θ matrix, replacing it with $\phi = \phi(\theta)$.

```
for j:=1 to p−1 do θ:=SWP[j] θ
for j:=p−1 down to 1 do θ:=RSW_{A_j}[j] θ
```

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 φ:=SWP_{A_j}[j] φ
for j:=p−1 down to 1 do φ:=RSW[j] φ
```

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 $\phi = (\phi_1, \ldots, \phi_p)$,

$$L(\phi | Y_{obs}) = \prod_{j=1}^{p} L(\phi_j | Y_{obs}),$$

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 $j = 1, \ldots, p$. The ML estimate of ϕ is $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)$, 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 $1, \ldots, n_j$ of the data matrix. Let z_j denote the observed data in column j,

$$z_j = (y_{1j}, y_{2j}, \ldots, y_{n_j,j})^T, \tag{6.10}$$

and X_j the upper-left $n_j \times (j-1)$ 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

$$z_j \mid X_j, \phi_j \sim N(X_j \beta_j, \gamma_j I),$$

so the likelihood for ϕ_j is

$$L(\phi_j \,|\, Y_{obs}) \propto \gamma_j^{-n_j/2} \exp\left\{ -\frac{1}{2\gamma_j} (z_j - X_j\beta_j)^T (z_j - X_j\beta_j) \right\}.$$

Using well-known properties of the normal linear regression model, the ML estimate of $\phi_j = (\beta_j^T, \gamma_j)^T$ is given by

$$\hat{\beta}_j = (X_j^T X_j)^{-1} X_j^T z_j, \tag{6.12}$$

$$\hat{\gamma}_j = n_j^{-1}\hat{\epsilon}_j^T\hat{\epsilon}_j, \tag{6.13}$$

where $\hat{\epsilon}_j = z_j - X_j\hat{\beta}_j$ (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 $j = 1, 2, \ldots, p$ 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 $\hat{\theta} = \phi^{-1}(\hat{\phi})$.

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,

$$\pi(\phi) = \pi_1(\phi_1)\,\pi_2(\phi_2)\cdots\pi_p(\phi_p), \tag{6.14}$$

then it is obvious that the posterior distribution $P(\phi \mid Y_{obs})$ will also factor into independent posteriors for $\phi_1, \ldots, \phi_p$, 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

$$P(\phi_j \mid Y_{obs}) \propto L(\phi_j \mid Y_{obs})\,\pi_j(\phi_j)$$

for $j = 1, \ldots, p$. For example, we can simulate a value of ϕ from $P(\phi \mid Y_{obs})$ by drawing ϕ_j from $P(\phi_j \mid Y_{obs})$ independently for $j = 1, \ldots, p$. A simulated value of θ from $P(\theta \mid Y_{obs})$ 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 $\pi_\phi(\phi)$. The relationship between π_θ and π_ϕ is

$$\pi_\phi(\phi) = \pi_\theta(\phi^{-1}(\phi))\;\big|\big|J\big|\big|^{-1}, \tag{6.16}$$

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 $\big|\big|J\big|\big|$ is the absolute value of the determinant of J.

By a well known property of determinants, $|\Sigma|$ can be written as

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}|\,|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|$$

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 $\big|\big|J\big|\big|$. 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 $Y_1, \ldots, Y_{p-1}$, and ϕ_p pertains to the regression of Y_p on $Y_1, \ldots, Y_{p-1}$. From (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 ξ_{p-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 $(\gamma_1\gamma_2\cdots\gamma_{p-2})^{-1}$. 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

$$\big|\big|J\big|\big| = \gamma_1^{-(p-1)}\gamma_2^{-(p-2)}\cdots\gamma_{p-1}^{-1}. \tag{6.19}$$

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,

$$L(\phi_j \mid Y_{obs}) \propto \gamma_j^{-n_j/2} \exp\left\{-\frac{1}{2\gamma_j}(z_j - X_j\beta_j)^T(z_j - X_j\beta_j)\right\}.$$

With some algebraic manipulation, it can be shown that

$$(z_j - X_j\beta_j)^T(z_j - X_j\beta_j) = \hat{\epsilon}_j^T\hat{\epsilon}_j + (\beta_j - \hat{\beta}_j)^T X^T X(\beta_j - \hat{\beta}_j),$$

where $\hat{\beta}_j = (X^TX_j)^{-1}X_j^Tz_j$ and $\hat{\epsilon}_j = X_j\hat{\beta}_j$. When $L(\phi_j \mid Y_{obs})$ is combined with the factor in (6.20) involving ϕ_j,

$$\pi_j(\phi_j) \propto \gamma_j^{-\left(\frac{p+1}{2}-p+j\right)},$$

the resulting posterior can be written as

$$\begin{aligned} P(\phi_j \mid Y_{obs}) \;\propto\; & \gamma_j^{-j/2} \exp\left\{-\frac{1}{2\gamma_j}(\beta_j - \hat{\beta}_j)^T X^T X(\beta_j - \hat{\beta}_j)\right\} \\ & \times \gamma_j^{-((n_j-p+j-1)/2)-1} \exp\left\{-\frac{1}{2\gamma_j}\hat{\epsilon}_j^T\hat{\epsilon}_j\right\}, \end{aligned}$$

which is the product of a multivariate normal and a scaled inverted-chisquare density,

$$\beta_j \mid \gamma_j, Y_{obs} \;\sim\; N(\hat{\beta}_j, \gamma_j(X^TX)^{-1}), \tag{6.21}$$

$$\gamma_j \mid Y_{obs} \;\sim\; \hat{\epsilon}_j^T\hat{\epsilon}_j\, \chi^{-2}_{n_j-p+j-1}. \tag{6.22}$$

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^*},
$$Y_{mis^*}^{(t+1)} \sim P(Y_{mis^*} \mid Y_{obs}, \theta^{(t)}). \tag{6.23}$$
2. P-step: Conditioning on $Y_{mis^*}^{(t+1)}$, draw a new value of θ from its posterior given the now-completed monotone pattern,
$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis^*}^{(t+1)}). \tag{6.24}$$

In practice, the P-step will have to be carried out using the parameterization $\phi = (\phi_1, \ldots, \phi_p)$ that corresponds to the monotone pattern of (Y_{obs}, Y_{mis^*}). That is, we will have to draw

$$\phi^{(t+1)} = (\phi_1^{(t+1)}, \ldots, \phi_p^{(t+1)})$$

by drawing

$$\phi_j^{(t+1)} \sim P(\phi_j \mid Y_{obs}, Y_{mis^*}^{(t+1)})$$

independently for $j = 1, 2, \ldots, p$, and then calculate

$$\theta^{(t+1)} = \phi^{-1}(\phi^{(t+1)})$$

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

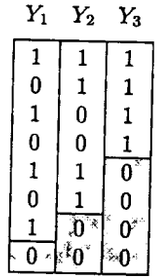

Figure 6.9. *Possible missingness patterns for a three-variable dataset, with observed and missing variables denoted by 1 and 0, respectively.*

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 $P(\theta \mid Y_{obs}, Y_{mis^*})$ 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 $Y_1, Y_2, \ldots, Y_p$ 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 $s = 1, 2, \ldots, S$. 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

$$r_{sj} = \begin{cases} 1 & \text{if } Y_j \text{ is observed in pattern } s, \\ 0 & \text{if } Y_j \text{ is missing in pattern } s. \end{cases}$$

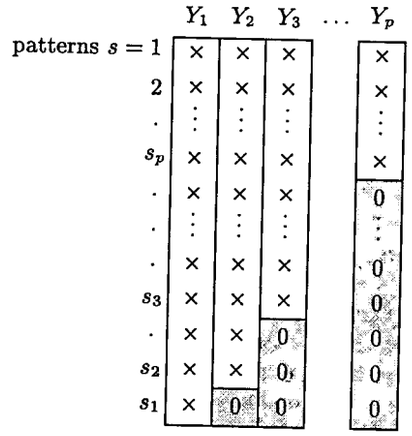

Figure 6.10. *Arrangement of missingness patterns for monotone data augmentation, with 0 denoting a variable that is missing and × denoting a variable that is either observed or missing.*

Let $\mathcal{O}(s)$ and $\mathcal{M}(s)$ denote the column labels corresponding to variables that are observed and missing, respectively, in pattern s,

$$\begin{aligned} \mathcal{O}(s) &= \{j : r_{sj}=1\}, \\ \mathcal{M}(s) &= \{j : r_{sj}=0\}. \end{aligned}$$

Also, let $\mathcal{M}^*(s)$ denote the subset of $\mathcal{M}(s)$ that must be filled in to complete the monotone pattern, and let $\mathcal{M}^{**}(s)$ be the remainder of $\mathcal{M}(s)$,

$$\begin{aligned} \mathcal{M}^*(s) &= \{j : r_{sj}=0 \text{ and } s_j \geq s\}, \\ \mathcal{M}^{**}(s) &= \{j : r_{sj}=0 \text{ and } s_j < s\}. \end{aligned}$$

For any s, $\mathcal{M}^*(s)$ lists the columns with missing values in the unshaded region of Figure 6.10, and $\mathcal{M}^{**}(s)$ lists the columns in the shaded region. Finally, let $\mathcal{I}(s)$ 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 $\mathcal{M}(s)$ with the potentially smaller set $\mathcal{M}^*(s)$. 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 $\hat{\beta}_j$, $(X_j^T X_j)^{-1}$, and $\hat{\epsilon}_j^T \hat{\epsilon}_j$, 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,

$$T(s) = \begin{bmatrix} n_s & \sum y_{i1} & \sum y_{i2} & \cdots & \sum y_{ip} \\ & \sum y_{i1}^2 & \sum y_{i1}y_{i2} & \cdots & \sum y_{i1}y_{ip} \\ & & \sum y_{i2}^2 & \cdots & \sum y_{i2}y_{ip} \\ & & & \ddots & \vdots \\ & & & & \sum y_{ip}^2 \end{bmatrix},$$

where all sums are taken over $i \in \mathcal{I}(s)$, and $n_s = \sum_{i \in \mathcal{I}(s)} 1$ 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 $j \in \mathcal{M}(s)$ or $k \in \mathcal{M}(s)$, 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 $j \in \mathcal{M}^{**}(s)$ or $k \in \mathcal{M}^{**}(s)$, 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 $0, 1, \ldots, j-1$, the result is

$$\begin{bmatrix} -(X_j^T X_j)^{-1} & (X_j^T X_j)^{-1} X_j^T z_j & 0 \\ z_j^T X_j (X_j^T X_j)^{-1} & z_j^T z_j - z_j^T X_j (X_j^T X_j)^{-1} X_j^T z_j & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

But notice that

$$(X_j^T X_j)^{-1} X_j^T z_j = \hat{\beta}_j$$

is the vector of estimated coefficients from ordinary least-squares regression of z_j on X_j. Moreover, it is straightforward to show that

$$z_j^T z_j - z_j^T X_j (X_j^T X_j)^{-1} X_j^T z_j = \hat{\epsilon}_j^T \hat{\epsilon}_j,$$

where $\hat{\epsilon}_j = z_j - X_j \hat{\beta}_j$ 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

$$\mathrm{SWP}_{A_j}[0, \ldots, j-1]\, T_j = \begin{bmatrix} -(X_j^T X_j)^{-1} & \hat{\beta}_j & 0 \\ \hat{\beta}_j^T & \hat{\epsilon}_j^T \hat{\epsilon}_j & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $A_j = \{0, 1, \ldots, j\}$.

An implementation of the P-step is shown in Figure 6.11. The components of ϕ are simulated in the reverse order $\phi_p, \phi_{p-1}, \ldots, \phi_1$ 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


```
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 := SWP_{A_j}[0, 1, ..., j - 1]T
    draw φ_jj ~ T_jj / χ²_{N_j - p + j - 1}
    C := Chol_{A_{j-1}}(-φ_jj T)
    for k := 0 to j - 1 do
        draw v_k ~ N(0, 1)
        φ_kj := T_kj
        for l := 0 to k do φ_kj := φ_kj + C_lk v_l
        end do
    if j > 1 and s_{j-1} > s_j then
        T := RSW_{A_{j-1}}[0, 1, ..., j - 1]T
    else if j > 1 and s_{j-1} = s_j then
        T := RSW_{A_{j-1}}[j - 1]T
        end if
    end do
```

Figure 6.11. *P-step for monotone data augmentation.*

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 $v = (v_0, v_1, \ldots, v_{p-1})$ 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

$$N_j = \sum_{s=1}^{s_j} n_s,$$

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 $Y_1, \ldots, Y_p$ are accumulated in T over missingness patterns $1, \ldots, s_p$. The matrix T is swept on positions $0, \ldots, p-1$, producing statistics from the regression of Y_p on $Y_1, \ldots, Y_{p-1}$. 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 $0, \ldots, p-1$ 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 $Y_1, \ldots, Y_{p-2}$. Continuing in this fashion, the algorithm draws $\phi_{p-1}, \phi_{p-2}, \ldots, \phi_1$. 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

$$T := T + \sum_{s=s_{j+1}+1}^{s_j} T_{obs}(s) + \sum_{s=s_{j+1}+1}^{s_j} T_{mis^*}(s). \qquad (6.25)$$

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 $j = 1, 2, \ldots, p$, 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 $P(\theta \mid Y_{obs})$, and (b) create proper multiple imputations of Y_{mis}. The output stream is a sequence

$$(Y_{mis^*}^{(1)}, \theta^{(1)}), (Y_{mis^*}^{(2)}, \theta^{(2)}), \ldots, (Y_{mis^*}^{(t)}, \theta^{(t)}), \ldots$$

with $P(Y_{mis^*}, \theta \mid Y_{obs})$ 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 $P(\theta \mid Y_{obs})$ 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,

$$\begin{aligned} Y_{mis}^{(1)} &\sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)}), \\ Y_{mis}^{(2)} &\sim P(Y_{mis} \mid Y_{obs}, \theta^{(t+k)}), \\ &\vdots \\ Y_{mis}^{(m)} &\sim P(Y_{mis} \mid Y_{obs}, \theta^{(t+mk)}), \end{aligned}$$

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. 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

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

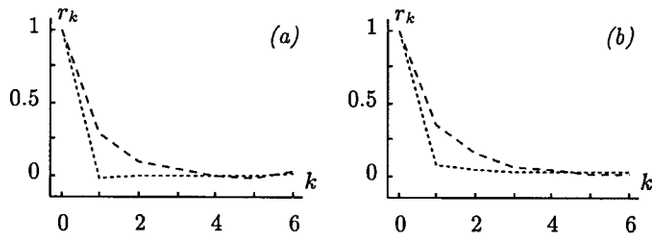

Figure 6.12. *Sample ACFs of series from ordinary data augmentation (dashed line) and monotone data augmentation (dotted line) for (a) the correlation between BMI and CHL, and (b) the worst-linear function of the parameter.*

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 $Y_1, Y_2, \ldots, Y_p$ 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 $j = 1, 2, \ldots, p$. Throughout this chapter, we will regard the levels $1, 2, \ldots, d_j$ 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 $Y_1, Y_2, \ldots, Y_p$ 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 $Y_1, Y_2, \ldots, Y_p$. 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 ($Y_1 = 1, Y_2 = 2$) 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 $d = 1, 2, \ldots, D$. Let x_d be the number of sample units that fall into cell d, and let

$$x = (x_1, x_2, \ldots, x_D)$$

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

$$\theta = (\theta_1, \theta_2, \ldots, \theta_D),$$

where θ_d is the probability that a unit falls into cell d. The probability distribution for x is given by

$$P(x|\theta) = \frac{n!}{x_1! x_2! \cdots x_D!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_D^{x_D} \tag{7.1}$$

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 $n - \sum_{d=1}^{D-1} x_d$ 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 $\theta = (\theta_1, \ldots, \theta_D)$ as the unknown parameter, with the understanding that it must lie within the simplex

$$\Theta = \left\{ \theta : \theta_d \geq 0 \text{ for all } d \text{ and } \sum_{d=1}^{D} \theta_d = 1 \right\}, \tag{7.2}$$

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,

$$P(x \mid n) = \frac{n!}{x_1!\,(n-x_1)!}\,\theta_1^{x_1}(1-\theta_1)^{n-x_1}.$$

In this special case, we will sometimes use the notation

$$x_1 \mid \theta_1 \sim B(n, \theta_1)$$

as an alternative to

$$(x_1, n-x_1) \mid \theta \sim M(n, (\theta_1, n-\theta_1)).$$

The first two moments of the multinomial distribution are given by

$$\begin{aligned} E(x_d \mid \theta) &= n\,\theta_d, \\ V(x_d \mid \theta) &= n\,\theta_d(1-\theta_d), \\ \mathrm{Cov}(x_d, x_{d'} \mid \theta) &= -n\,\theta_d\theta_{d'}, \quad d' \neq d. \end{aligned}$$

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 $L(\theta \mid Y)$ rather than $L(\theta \mid x)$. 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 $x = (x_1, \ldots, x_D)$. Therefore, complete-data ML estimates can be obtained simply by equating each observed cell count x_d to its expectation $E(x_d \mid \theta) = n\theta_d$, leading to the well-known result that the ML estimates for the cell probabilities are the observed proportions

$$\hat{\theta}_d = \frac{x_d}{n}, \quad d = 1, \ldots, D. \tag{7.5}$$

7.2.2 Collapsing and partitioning the multinomial

The multinomial distribution has two convenient properties that enable us to factor the probability distribution $P(x \mid \theta)$ and the likelihood $L(\theta \mid Y)$. 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 $x^* = (z, x_3, \ldots, x_D)$ where $z = x_1 + x_2$. Then (a) the distribution of 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,

$$(x_1, x_2) \mid z, \theta \sim M(z, (\theta_1/\xi, \theta_2/\xi)). \tag{7.7}$$

Property (a) is derived by summing the multinomial probabilities for all x-vectors consistent with $x_1 + x_2 = z$,

$$\begin{aligned} P(x^* \mid \theta) &= \sum_{j=0}^{z} P(x_1 = j, x_2 = z-j, x_3, \ldots, x_D) \\ &= \sum_{j=0}^{z} \frac{n!}{j!\,(z-j)!\,x_3! \cdots x_D!}\,\theta_1^j \theta_2^{z-j} \theta_3^{x_3} \cdots \theta_D^{x_D} \end{aligned}$$

$$= \frac{n!}{z!\,x_3!\cdots x_D!}\,\theta_3^{x_3}\cdots\theta_D^{x_D}\sum_{j=0}^{z}\frac{z!}{j!\,(z-j)!}\,\theta_1^{j}\theta_2^{z-j},$$

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

$$(z, n-z) \mid \theta \sim M(n, (\xi, 1-\xi)).$$

Moreover, if we collapse $x_3 + \cdots + x_D = n - z$ but leave x_1 and x_2 intact, then

$$(x_1, x_2, n-z) \mid \theta \sim M(n, (\theta_1, \theta_2, 1-\xi)).$$

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 $\{1, 2, \ldots, D\}$ into subsets $A_1, A_2, \ldots, A_K$ 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 $\{x_{(1)}, x_{(2)}, \ldots, x_{(K)}\}$ 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 kth part by

$$z_k = \sum_{d \in A_k} x_d.$$

The collection $z = (z_1, z_2, \ldots, z_K)$ of these total frequencies will be called the *collapsed table*. Denote the probability that a sample unit falls into the kth 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 kth part by

$$\phi_{kd} = \theta_d / \xi_k \quad \text{for all } d \in A_k. \tag{7.10}$$

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

$$\psi = (\xi, \phi_1, \ldots, \phi_K),$$

are a one-to-one function of $\theta = (\theta_1, \ldots, \theta_d)$; the forward transformation $\psi = \psi(\theta)$ is defined by (7.9)–(7.10), and the back transformation $\theta = \psi^{-1}(\psi)$ is

$$\theta_d = \xi_k \phi_{kd} \quad \text{for all } d \in A_k, \tag{7.13}$$

$k = 1, 2, \ldots, K$. Moreover, the parameters for the collapsed table and each part of the partitioned table are mutually distinct; any values of $\xi, \phi_1, \ldots, \phi_K$ 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,

$$L(\psi|x) = L(\xi|z)\, L(\phi_1|x_{(1)}) \cdots L(\phi_K|x_{(K)}).$$

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 $\xi, \phi_1, \ldots, \phi_K$ can be calculated independently; they are

$$\hat{\xi}_k = \frac{z_k}{n} \quad \text{and} \quad \hat{\phi}_{kd} = \frac{x_d}{z_k} \quad \text{for all } d \in A_k.$$

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 non-fixed 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

$$P(\theta|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2) \cdots \Gamma(\alpha_D)}\, \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1} \tag{7.14}$$

over the simplex Θ, where $\alpha_0 = \sum_{d=1}^{D} \alpha_d$ and $\Gamma(\cdot)$ 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 $\alpha_d > 0$ for $d = 1, \ldots, D$.

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

$$\pi(\theta) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1}, \tag{7.15}$$

where $\alpha_1, \ldots, \alpha_D$ 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 $\theta_1, \ldots, \theta_{D-1}$. 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

$$(\theta_1, \theta_2) \mid \alpha \sim D(\alpha).$$

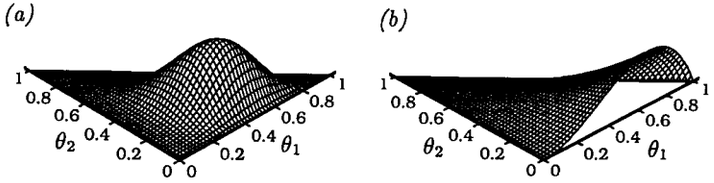

Figure 7.1. *Dirichlet densities for (a)* $\alpha = (5, 3, 4)$ *and (b)* $\alpha = (3, 1, 2)$, *plotted as functions of* θ_1 *and* θ_2.

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

$$\begin{aligned} E(\theta_d) &= \frac{\alpha_d}{\alpha_0}, \\ V(\theta_d) &= \frac{\alpha_d(\alpha_0 - \alpha_d)}{\alpha_0^2(\alpha_0 + 1)}, \\ \mathrm{Cov}(\alpha_d, \alpha_{d'}) &= -\frac{\alpha_d \alpha_{d'}}{\alpha_0^2(\alpha_0 + 1)}, \quad d' \neq d. \end{aligned}$$

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 $x_d = \alpha_d - 1$, $d = 1, \ldots, D$. 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 α_d 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 α 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 $v_1, v_2, \ldots, v_D$ 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 $\alpha = (\alpha_1, \ldots, \alpha_D)$,

$$\begin{aligned} x \mid \theta &\sim M(n, \theta), && (7.17) \\ \theta &\sim D(\alpha). && (7.18) \end{aligned}$$

Multiplying the Dirichlet density (7.15) by the multinomial likelihood (7.3) produces

$$P(\theta|Y) \propto \theta_1^{\alpha_1+x_1-1}\theta_2^{\alpha_2+x_2-1}\ldots\theta_D^{\alpha_D+x_D-1}, \qquad (7.19)$$

which is a Dirichlet density with parameters

$$\begin{aligned} \alpha' &= (\alpha'_1, \alpha'_2, \ldots, \alpha'_D) \\ &= (\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_D + x_D) \\ &= \alpha + x. \end{aligned} \qquad (7.20)$$

The posterior distribution of θ under (7.17)–(7.18) is thus

$$\theta \mid Y \sim D(\alpha').$$

The posterior mean is

$$E(\theta|Y) = \left(\frac{\alpha'_1}{\alpha'_0}, \frac{\alpha'_2}{\alpha'_0}, \ldots, \frac{\alpha'_D}{\alpha'_0}\right)$$

where $\alpha'_0 = \sum_{d=1}^{D}(\alpha_d + x_d) = \alpha_0 + n$, and the posterior mode is

$$\text{mode}(\theta|Y) = \left(\frac{\alpha'_1 - 1}{\alpha'_0 - D}, \frac{\alpha'_2 - 1}{\alpha'_0 - D}, \ldots, \frac{\alpha'_D - 1}{\alpha'_0 - D}\right)$$

provided that every $\alpha'_d \geq 1$.

The Dirichlet prior (7.17) is a proper probability distribution if $\alpha_1, \alpha_2, \ldots, \alpha_D$ 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

$$\pi(\theta) \propto \theta_1^{-1}\theta_2^{-1}\cdots\theta_D^{-1}, \qquad (7.21)$$

which is the limiting form of the $D(\alpha)$ density as α approaches $(0, 0, \ldots, 0)$, 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 $\alpha = (0, 0, \ldots, 0)$; 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 $\alpha = (\alpha_1, \ldots, \alpha_D)$, 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 θ, it may be sensible to take $\alpha_1, \alpha_2, \ldots, \alpha_D$ equal to a common value—that is, to set $\alpha = (c, c, \ldots, c)$ 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

$$\hat{\theta} = \left(\frac{x_1}{n}, \frac{x_2}{n}, \ldots, \frac{x_D}{n}\right) \qquad (7.22)$$

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 $x = (x_1, \ldots, x_D)$ 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 $x = (x_1, \ldots, x_D)$ represents a cross-classification by discrete variables $Y_1, Y_2, \ldots, Y_p$, 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 $\alpha = (c, c, \ldots, c)$ 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 $\epsilon = c$ 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 ϵ imaginary counts to every cell introduces information equivalent to $D\epsilon$ 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 $1, 2, \ldots, d_j$ 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

$$x_{(k)} = \{x_d : d \in A_k\}$$

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

$$\phi_{kd} = \theta_d / \xi_k \quad \text{for all } d \in A_k.$$

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 $\phi \doteq (\phi_1, \phi_2, \ldots, \phi_K)$ given ξ is a set of K independent Dirichlet distributions,

$$\begin{aligned} \phi_1 \mid \xi, \alpha &\sim D(\alpha_{(1)}), \\ \phi_2 \mid \xi, \alpha &\sim D(\alpha_{(2)}), \\ &\vdots \\ \phi_K \mid \xi, \alpha &\sim D(\alpha_{(K)}), \end{aligned} \tag{7.26}$$

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 $Y_1, Y_2, \ldots, Y_p$. 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 $x = (x_1, x_2, \ldots, x_D)$ is actually a cross-classification by the levels of $Y_1, Y_2, \ldots, Y_p$, and as such can be regarded as a p-dimensional array. Suppose that variable Y_j takes possible values $1, 2, \ldots, d_j$. Let x_y, where $y = (y_1, y_2, \ldots, y_p)$, be the total number of units in the sample for which the event $Y_1 = y_1, Y_2 = y_2, \ldots, Y_p = y_p$ 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 $(Y_1, Y_2, \ldots, Y_p)$ 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 $\mathcal{Y}$. Assuming for the moment that there are no structural zeroes, $\mathcal{Y}$ is the Cartesian cross-product of the sets $\{1, 2, \ldots, d_j\}$ for $j = 1, 2, \ldots, p$. When a cell count or probability appears with the vector subscript $y = (y_1, y_2, \ldots, y_p)$ 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,

$$x = (x_1, x_2, \ldots, x_D), \quad \theta = (\theta_1, \theta_2, \ldots, \theta_D),$$

and at other times as p-dimensional arrays,

$$x = \{x_y : y \in \mathcal{Y}\}, \quad \theta = \{\theta_y : y \in \mathcal{Y}\}.$$

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

$$x^{(s)} = \{x_y^{(s)} : y \in \mathcal{Y}\}$$

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 $\mathcal{O}_s$ and $\mathcal{M}_s$ 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,

$$\begin{aligned} \mathcal{O}_s(y) &= \{y_j : r_{sj} = 1\}, \\ \mathcal{M}_s(y) &= \{y_j : r_{sj} = 0\}. \end{aligned}$$

Also, let O_s and M_s be, respectively, the sets of all possible values of $\mathcal{O}_s(y)$ and $\mathcal{M}_s(y)$. 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 $\mathcal{O}_s(y) = (y_1, y_4)$, $\mathcal{M}_s(y) = (y_2, y_3)$,

$$\begin{aligned} O_s &= \{(y_1, y_4) : y_1 = 1, 2, \ldots, d_1;\ y_4 = 1, 2, \ldots, d_4\}, \\ M_s &= \{(y_2, y_3) : y_2 = 1, 2, \ldots, d_2;\ y_3 = 1, 2, \ldots, d_3\}. \end{aligned}$$

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

$$z_{\mathcal{O}_s(y)}^{(s)} = \sum_{\mathcal{M}_s(y) \in M_s} x_y^{(s)} \quad \text{for all } \mathcal{O}_s(y) \in O_s. \tag{7.27}$$

The marginal probability that an observation falls within cell $\mathcal{O}_s(y)$ of this table will be called

$$\beta_{\mathcal{O}_s(y)} = \sum_{\mathcal{M}_s(y) \in M_s} \theta_y. \tag{7.28}$$

Observed-data likelihood

When $x = (x_1, x_2, \ldots, x_D)$ has a multinomial distribution with parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$, then the complete-data loglikelihood function for θ is

$$l(\theta | Y) = \sum_{d=1}^{D} x_d \log \theta_d$$

over the simplex Θ (Section 7.2.1). Equivalently, viewing x and θ as p-dimensional arrays, we can write the loglikelihood as

$$l(\theta | Y) = \sum_{y \in \mathcal{Y}} x_y \log \theta_y. \tag{7.29}$$

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

$$z^{(s)} = \{z_{\mathcal{O}_s(y)}^{(s)} : \mathcal{O}_s(y) \in O_s\}. \tag{7.30}$$

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

$$\beta^{(s)} = \{\beta_{\mathcal{O}_s(y)} : \mathcal{O}_s(y) \in O_s\}. \tag{7.31}$$

That is, the contribution of $z^{(s)}$ to the observed-data loglikelihood is

$$\sum_{\mathcal{O}_s(y) \in O_s} z_{\mathcal{O}_s(y)}^{(s)} \log \beta_{\mathcal{O}_s(y)}.$$

The observed-data loglikelihood is the sum of these contributions for missingness patterns $s = 1, 2, \ldots, S$,

$$l(\theta | Y_{obs}) = \sum_{s=1}^{S} \sum_{\mathcal{O}_s(y) \in O_s} z_{\mathcal{O}_s(y)}^{(s)} \log \beta_{\mathcal{O}_s(y)}. \tag{7.32}$$

Despite the concise appearance of (7.32), it is a rather complicated function of the individual elements of θ. Evaluating $l(\theta \mid Y_{obs})$ 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 $l(\theta \mid Y_{obs})$ 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 $x = \{x_y : y \in \mathcal{Y}\}$, the unobserved p-dimensional table

that cross-classifies all sample units by their values of $Y_1, Y_2, \ldots, Y_p$. 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 $1, \ldots, S$. 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

$$x^{(s)}_{\mathcal{O}_s(y)} = \{x^{(s)}_y : \mathcal{M}_s(y) \in M_s\} \tag{7.33}$$

denote the portion of $x^{(s)}$ that is obtained by fixing $\mathcal{O}_s(y)$ at a specific value but varying $\mathcal{M}_s(y)$ over M_s; that is, $x^{(s)}_{\mathcal{O}_s(y)}$ is simply the set of all cell counts in $x^{(s)}$ that contribute to the observed count $z^{(s)}_{\mathcal{O}_s(y)}$. By the partitioning rules, $x^{(s)}_{\mathcal{O}_s(y)}$ has, given $z^{(s)}_{\mathcal{O}_s(y)}$, a multinomial distribution with index $z^{(s)}_{\mathcal{O}_s(y)}$ and parameters

$$\gamma_{\mathcal{O}_s(y)} = \{\theta_y / \beta_{\mathcal{O}_s(y)} : \mathcal{M}_s(y) \in M_s\}; \tag{7.34}$$

that is,

$$x^{(s)}_{\mathcal{O}_s(y)} \mid z^{(s)}_{\mathcal{O}_s(y)}, \theta \sim M(z^{(s)}_{\mathcal{O}_s(y)}, \gamma_{\mathcal{O}_s(y)}). \tag{7.35}$$

Notice that (7.34) is simply the portion of θ corresponding to $x^{(s)}_{\mathcal{O}_s(y)}$, rescaled so that its elements sum to one. It follows that the conditional expectation of an element of $x^{(s)}$ is

$$E(x^{(s)}_y \mid z^{(s)}, \theta) = z^{(s)}_{\mathcal{O}_s(y)} \theta_y / \beta_{\mathcal{O}_s(y)}. \tag{7.36}$$

The E-step consists of calculating (7.36) for every $s = 1, \ldots, S$ and summing the results,

$$E(x_y \mid Y_{obs}, \theta) = \sum_{s=1}^{S} z^{(s)}_{\mathcal{O}_s(y)} \theta_y / \beta_{\mathcal{O}_s(y)}. \tag{7.37}$$

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

$$\text{estimate } \theta_y \text{ by } E(x_y \mid Y_{obs}, \theta)/n \tag{7.38}$$

for all $y \in \mathcal{Y}$.

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

```
for y ∈ 𝒴 do x_y := 0
for s := 1 to S do
    for 𝒪_s(y) ∈ O_s do
        if z^(s)_{𝒪_s(y)} ≠ 0 then
            if M_s = ∅ then
                x_y := x_y + z^(s)_{𝒪_s(y)}
            else
                sum := 0
                for ℳ_s(y) ∈ M_s do sum := sum + θ_y
                for ℳ_s(y) ∈ M_s do x_i := x_y + z^(s)_{𝒪_s(y)} θ_y/sum
                end if
            end if
        end do
    end do
for y ∈ 𝒴 do θ_y := x_y/n
```

Figure 7.2. *Single iteration of EM for the saturated multinomial model.*

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 $\mathcal{M}_s(y)$ is nonempty). If not, then the observed counts for pattern s are added into the 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 $\alpha = \{\alpha_y : y \in \mathcal{Y}\}$, then the last line in Figure 7.2 should be changed to

$$\texttt{for } y \in \mathcal{Y} \texttt{ do } \theta_y := (x_y + \alpha_y - 1)/(n + \alpha_0 - D), \qquad (7.39)$$

where $\alpha_0 = \sum_{y \in \mathcal{Y}} \alpha_y$ and D is the total number of cells in θ. Taking $\alpha_y = 1$ for all $y \in \mathcal{Y}$ 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

$$\texttt{for } y \in \mathcal{Y}^* \texttt{ do } \theta_y := (x_y + \alpha_y - 1)/(n + \alpha_0^* - D^*), \qquad (7.40)$$

where $\mathcal{Y}^*$ is the set of all possible values of y excluding the structural zeroes, $\alpha_0^* = \sum_{y \in \mathcal{Y}^*} \alpha_y$ is the sum of the prior hyperparameters and D^* is the number of elements in $\mathcal{Y}^*$ (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

$$\log P(\theta \,|\, Y_{obs}) \;=\; l(\theta \,|\, Y_{obs}) \;+\; \log \pi(\theta),$$

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 $x^{(s)}$


```
l := 0
for s := 1 to S do
    for O_s(y) ∈ O_s do
        if z^(s)_{O_s(y)} ≠ 0 then
            if M_s = ∅ then
                l := l + z^(s)_{O_s(y)} log θ_y
            else
                sum := 0
                for M_s(y) ∈ M_s do sum := sum + θ_y
                l := l + z^(s)_{O_s(y)} log (sum)
                end if
            end if
        end do
    end do
```

Figure 7.3. *Evaluation of the observed-data loglikelihood function.*

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 $D(\alpha + x)$.

In the pseudocode of Figure 7.2, the line

$$\texttt{for } \mathcal{M}_s(y) \in M_s \texttt{ do } x_y := x_y + z^{(s)}_{\mathcal{O}_s(y)} \theta_y / \text{sum} \qquad (7.41)$$

allocates an observed count $z^{(s)}_{\mathcal{O}_s(y)}$ 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

$$x^{(s)}_{\mathcal{O}_s(y)} \;\sim\; M(z^{(s)}_{\mathcal{O}_s(y)}, \gamma_{\mathcal{O}_s(y)})$$

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 $\gamma_{\mathcal{O}_s(y)}$. 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 multi-

```
        for m:= 1 to z^(s)_{O_s(y)} do
            draw u ~ U(0,1)
            k:= 0
            for M_s(y) ∈ M_s do
                if k + θ_y/sum > u then
                    x_y:= x_y + 1
                    goto 1
                else
                    k:= k + θ_y/sum
                    end if
                end do
1           continue
            end do
```

Figure 7.4. *Table sampling for the data augmentation I-step.*

nomial draws in that situation, which relies on a Poisson variate generator, is described by Brown and Bromberg (1984).

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 $z^{(1)}, \ldots, z^{(S)}$ and the improper Dirichlet prior with $\alpha = (0, 0, \ldots, 0)$ is being used, then depending on the pattern of the zeroes the P-step could be undefined, because some elements of $\alpha + x$ could be zero. For this reason, the prior $\alpha = (0, 0, \ldots, 0)$ should be avoided whenever random zeroes are present. A proper prior, e.g. a flattening prior with $\alpha = (c, c, \ldots, c)$ 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 $P(x \mid Y_{obs})$. 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 $P(Y_{mis} \mid Y_{obs})$. 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 $\mathcal{I}(s)$ denotes the rows of Y that exhibit missingness pattern s. The vector workspace $y = (y_1, y_2, \ldots, y_p)$ serves as a counter, indexing the cells of the p-dimensional contingency table. For any row i in missingness pattern s, the subvector $\mathcal{O}_s(y)$ of y is first set equal to the observed data in $y_{i(obs)}$, so that the remaining portion $\mathcal{M}_s(y)$ 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 ≠ ∅ then
        for i ∈ I(s) do
            O_s(y):= y_{i(obs)}
            sum:= 0
            for M_s(y) ∈ M_s do sum:= sum + θ_y
            draw u ~ U(0,1)
            k:= 0
            for M_s(y) ∈ M_s do
                if k + θ_y/sum > u then
                    y_{i(mis)}:= M_s(y)
                    goto 1
                else
                    k:= k + θ_y/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 ($\theta_{2+} = \theta_{21} + \theta_{22}$),

$$\begin{aligned} \delta &= \theta_{+2} - \theta_{2+} \\ &= \theta_{12} - \theta_{21}. \end{aligned} \tag{7.43}$$

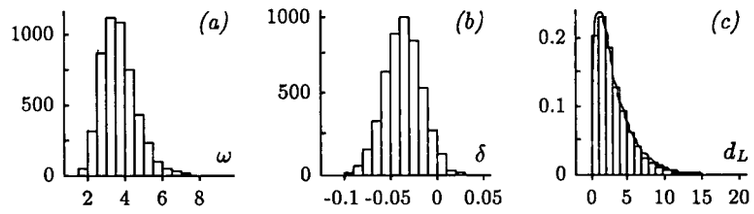

Figure 7.6. *Histograms of (a) $\omega = (\theta_{11}\theta_{22})/(\theta_{12}\theta_{21})$ and (b) $\delta = \theta_{12} - \theta_{21}$ over 5000 iterations of data augmentation, and of (c) the likelihood-ratio statistic d_L with the χ^2_3 density superimposed.*

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 $\alpha_i = 1/2$), 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 ω 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

$$d_L = 2[l(\hat{\theta}|Y_{obs}) - l(\theta|Y_{obs})],$$

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^2_3 density function superimposed; the two are nearly indistinguishable.

By averaging the 5000 iterates of ω and δ, we obtain simulated posterior means

$$\hat{E}(\omega|Y_{obs}) = 3.67 \quad \text{and} \quad \hat{E}(\delta|Y_{obs}) = -0.036.$$

Comparing these to the ML estimates obtained in Section 3.2.2,

$$\hat{\omega} = 3.57 \quad \text{and} \quad \hat{\delta} = -0.037,$$

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

$$\hat{V}(\hat{\delta}) = \frac{1}{n}\left[\hat{\theta}_{12}(1-\hat{\theta}_{12}) + \hat{\theta}_{21}(1-\hat{\theta}_{21}) + 2\hat{\theta}_{12}\hat{\theta}_{21}\right]$$

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 $\alpha = (c, c, \ldots, c)$ 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,

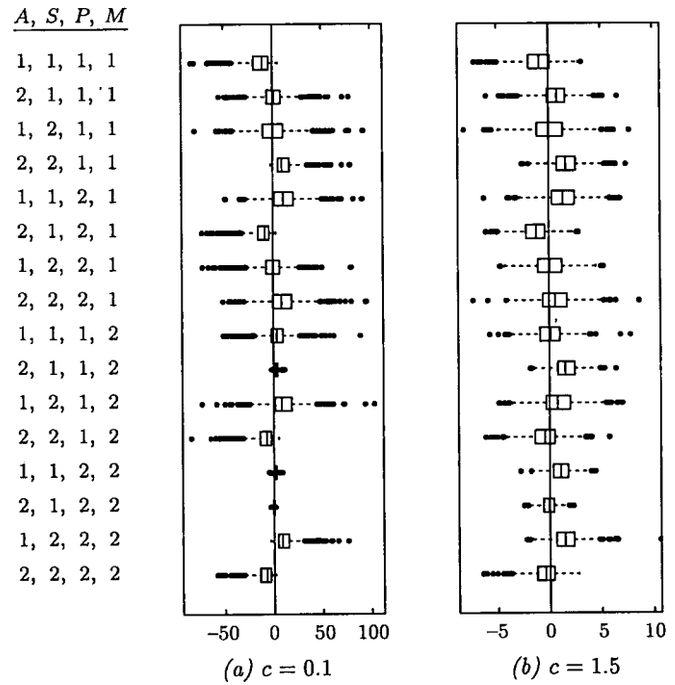

Figure 7.7. *Boxplots of simulated log-odds ratios from 1000 iterations of data augmentation under two flattening priors.*

because with $c = 1.5$ we have already added the equivalent of $1.5 \times 64 = 96$ 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

$$\begin{aligned} P(Y_1,\ldots,Y_p|\theta) &= P(Y_1|\phi_1)\,P(Y_2|Y_1,\phi_2) \\ &\quad \cdots P(Y_p|Y_1,\ldots,Y_{p-1},\phi_p), \end{aligned} \tag{7.44}$$

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 $\mathcal{F}_j$ be a function that extracts from y the first j elements,

$$\mathcal{F}_j(y) = (y_1,\ldots,y_j),$$

and let $\mathcal{L}_j$ extract the last $p-j$ elements,

$$\mathcal{L}_j(y) = (y_{j+1},\ldots,y_p).$$

Let F_j and L_j, respectively, be the sets over which $\mathcal{F}_j(y)$ and $\mathcal{L}_j(y)$ 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

$$\xi_{\mathcal{F}_j(y)} = \sum_{\mathcal{L}_j(y)\in L_j} \theta_y,$$

and the full set of parameters governing the marginal distribution of $(Y_1,Y_2,\ldots,Y_j)$ as

$$\xi_j = \{\xi_{\mathcal{F}_j(y)} : \mathcal{F}_j(y)\in F_j\}.$$

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

$$\phi_{\mathcal{F}_j(y)} = \xi_{\mathcal{F}_j(y)}/\xi_{\mathcal{F}_{j-1}(y)}, \tag{7.45}$$

and the full set of parameters governing the conditional distribution of Y_j given $(Y_1,Y_2,\ldots,Y_{j-1})$ is

$$\phi_j = \{\phi_{\mathcal{F}_j(y)} : \mathcal{F}_j(y)\in F_j\}.$$

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

$$z_{\mathcal{F}_j(y)} = \sum_{\mathcal{L}_j(y)\in L_j} x_y,$$

and the entire j-dimensional table by

$$z_j = \{z_{\mathcal{F}_j(y)} : \mathcal{F}_j(y)\in F_j\}.$$

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

$$z_{j;\mathcal{F}_{j-1}(y)} = \{z_{\mathcal{F}_j(y)} : y_j = 1,2,\ldots,d_j\},$$

which is simply the portion of z_j obtained by fixing $(y_1,\ldots,y_{j-1})$ at a specific value but letting y_j vary over $\{1,2,\ldots,d_j\}$. The table z_j is then the collection of these vectors,

$$z_j = \{z_{j;\mathcal{F}_{j-1}(y)} : \mathcal{F}_{j-1}(y)\in F_{j-1}\}.$$

If we partition ϕ_j in the same fashion, as

$$\phi_j = \{\phi_{j;\mathcal{F}_{j-1}(y)} : \mathcal{F}_{j-1}(y)\in F_{j-1}\}$$

where

$$\phi_{j;\mathcal{F}_{j-1}(y)} = \{\phi_{\mathcal{F}_j(y)} : y_j = 1,2,\ldots,d_j\},$$

then the conditional distribution of z_j given z_{j-1} is

$$z_{j;\mathcal{F}_{j-1}(y)} \mid z_{j-1},\phi_j \sim M(z_{\mathcal{F}_{j-1}(y)},\phi_{j;\mathcal{F}_{j-1}(y)}) \tag{7.46}$$

independently for all $\mathcal{F}_{j-1}(y)\in F_{j-1}$.

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 product-multinomial 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 $L(\phi_j \mid z_j)$ is a product-multinomial likelihood. The full set of parameters $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$ forms a one-to-one transformation of θ, and it follows from (7.45) that the back-transformation is

$$\theta_y = \phi_{\mathcal{F}_1(y)}\, \phi_{\mathcal{F}_2(y)} \cdots \phi_{\mathcal{F}_p(y)}. \tag{7.47}$$

Factoring the prior

Just as the likelihood function factors into independent pieces for $\phi_1, \phi_2, \ldots, \phi_p$, 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,

$$\alpha = \{\alpha_y : y \in \mathcal{Y}\}.$$

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

$$\beta_j = \{\beta_{\mathcal{F}_j(y)} : \mathcal{F}_j(y) \in F_j\},$$

are obtained by summing the elements of α in the same way the elements of θ were summed to produce ξ_j,

$$\beta_{\mathcal{F}_j(y)} = \sum_{\mathcal{L}_j(y) \in L_j} \alpha_y.$$

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

$$\beta_j = \{\beta_{j;\mathcal{F}_{j-1}(y)} : \mathcal{F}_{j-1}(y) \in F_{j-1}\}$$

where

$$\beta_{j;\mathcal{F}_{j-1}(y)} = \{\beta_{\mathcal{F}_j(y)} : y_j = 1, 2, \ldots, d_j\},$$

the conditional distribution of ϕ_j given ξ_{j-1} is

$$\phi_{j;\mathcal{F}_{j-1}(y)} \mid \xi_{j-1} \sim D(\beta_{j;\mathcal{F}_{j-1}(y)}) \tag{7.49}$$

independently for all $\mathcal{F}_{j-1}(y) \in F_{j-1}$.

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 $\phi = (\phi_1, \ldots, \phi_p)$ 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,

$$\begin{aligned} x \mid \theta &\sim M(n, \theta), \\ \theta &\sim D(\alpha), \end{aligned}$$

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

$$\phi_{j;\mathcal{F}_{j-1}(y)} \mid z_j \sim D(\beta_{j;\mathcal{F}_{j-1}(y)} + z_{j;\mathcal{F}_{j-1}(y)}) \tag{7.52}$$

independently for all $\mathcal{F}_{j-1}(y) \in F_{j-1}$.

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

$$L(\phi|Y_{obs}) = \prod_{j=1}^{p} L(\phi_j|z_j^*),$$

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 $z^*_{\mathcal{F}_j(y)}$ and let

$$z^*_{j;\mathcal{F}_{j-1}(y)} = \{z^*_{\mathcal{F}_j(y)} : y_j = 1, 2, \ldots, d_j\}$$

be a subvector within this table, $L(\phi_j \mid z_j^*)$ will be the likelihood that arises from the product-multinomial distribution

$$z^*_{j;\mathcal{F}_{j-1}(y)} \mid z^*_{j-1}, \phi_j \sim M(z_{\mathcal{F}_{j-1}(y)}, \phi_{j;\mathcal{F}_{j-1}(y)})$$

for all $\mathcal{F}_{j-1}(y) \in F_{j-1}$. Combining this new likelihood with the prior (7.49) leads to the observed-data posterior

$$P(\phi|Y_{obs}) = \prod_{j=1}^{p} P(\phi_j|z_j^*), \tag{7.53}$$

where $P(\phi_j \mid z_j^*)$ is given by

$$\phi_{j;\mathcal{F}_{j-1}(y)} \mid z_j^* \sim D(\beta_{j;\mathcal{F}_{j-1}(y)} + z^*_{j;\mathcal{F}_{j-1}(y)}) \tag{7.54}$$

for all $\mathcal{F}_{j-1}(y) \in F_{j-1}$.

Monotone data augmentation capitalizes on (7.53) to create an efficient simulation algorithm for situations where Y_{obs} is nonmonotone. 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,

$$Y_{mis^*}^{(t+1)} \sim P(Y_{mis^*}|Y_{obs}, \theta^{(t)}).$$

2. P-step: Draw a new value of θ from its posterior distribution given Y_{obs} and the new value of Y_{mis^*},

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis^*}^{(t+1)}).$$

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 $\phi_1, \ldots, \phi_p$ from the factored posterior (7.53), and then numerically transforming the resulting value of $\phi = (\phi_1, \ldots, \phi_p)$ 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 $d_1 \times d_2 \times \cdots d_p$. 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 $j := p, p-1, \ldots, 1$, perform the following steps:

1. If $s_j > s_{j+1}$, impute the missing data for variables $Y_1, \ldots, Y_j$ 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 $Y_1, \ldots, Y_j$, 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 $Y_1, \ldots, Y_{j-1}$) 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 $\{\theta^{(t)} : t = 1, 2, \ldots\}$ which converges in distribution to the correct observed-data posterior,

$$P(\theta^{(t)}|Y_{obs}, \theta^{(0)}) \rightarrow P(\theta|Y_{obs}) \text{ as } t \rightarrow \infty.$$

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 7.5. *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)

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 81 880 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 initial 80 084 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

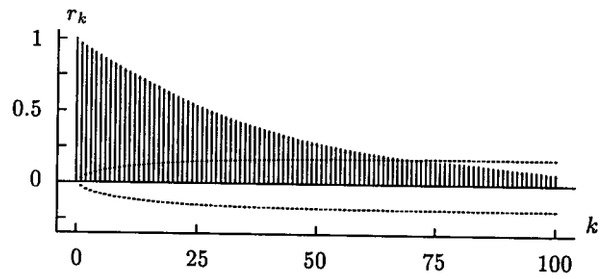

Figure 7.8. *Sample ACFs for the worst linear function of θ, estimated from 5000 iterations of ordinary data augmentation, with dashed lines indicating approximate critical values for testing $\rho_k = \rho_{k+1} = \cdots = 0$.*

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 θ values generated by monotone data augmentation will be 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

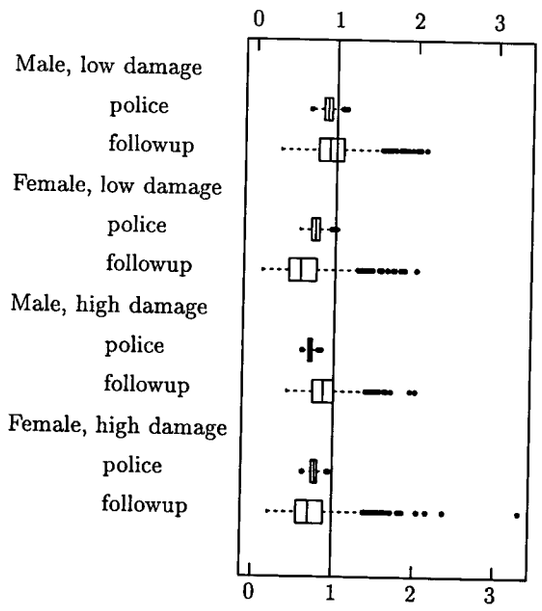

Figure 7.9. *Boxplots of 1000 simulated odds ratios showing the relationship of seatbelt use and injury within classes of damage and sex, both from police reports and from followup data.*

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