CHAPTER 5

Methods For Normal Data

5.1 Introduction

The most common probability model for continuous multivariate data is the multivariate normal distribution. Many standard methods for analyzing multivariate data, including factor analysis, principal components and discriminant analysis, are based upon an assumption of multivariate normality. Moreover, the classical techniques of linear regression and analysis of variance assume conditional normality of the response variables given linear functions of the predictors, which is the conditional distribution implied by a multivariate normal model for all the variables. Because statistical methods motivated by assumptions of normality are in such widespread use, it is natural to seek general techniques for inference from incomplete normal data.

Datasets encountered in the real world often deviate from multivariate normality, but in many cases the normal model will be useful even when the actual data are nonnormal. There are several important reasons for this. First, one can often make the normality assumption more tenable by applying suitable transformations to one or more of the variables. Second, if some variables in a dataset are clearly nonnormal (e.g. discrete) but are completely observed, then the multivariate normal model may still be used for inference provided that (a) it is plausible to model the incomplete variables as conditionally normal given a linear function of the complete ones, and (b) the parameters of inferential interest pertain only to this conditional distribution (Section 2.6.2).

Finally, even if some of the incompletely observed variables are clearly nonnormal, it may still be reasonable to

use the normal model as a convenient device for creating multiple imputations. As pointed out in Section 4.5.4, inference by multiple imputation may be robust to departures from the imputation model if the amounts of missing information are not large, because the imputation model is effectively applied not to the entire dataset but only to its missing part. For example, it may be quite reasonable to use normal model to impute a variable that is ordinal (consisting of small number of ordered categories), provided that the amount of missing data is not extensive and the marginal distribution is not too far from being unimodal and symmetric. When using the normal model to impute categorical data, however, the continuous imputes should be rounded off to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analyst. We have found that the normal model, when used in this fashion, can be an effective tool for imputing ordinal and even binary data in instances where constructing a more elaborate categorical-data model would be impractical (Schafer, Khare and Ezzati-Rice, 1993).

5.2 Relevant properties of the complete-data model

5.2.1 Basic notation

We begin by establishing some notational conventions that will be used throughout the chapter. The dataset, as depicted in Figure 2.1, is assumed to be a matrix of n rows and p columns, with rows corresponding to observational units and columns corresponding to variables. Denote the complete data by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis}, are the observed and missing portions of the matrix, respectively. Let y_{ij} denote an individual element of Y, $i = 1.2,...,n$, $j = 1,2,...,p$. The ith row of Y, expressed as a column vector (all vectors will be regarded as column vectors), is

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

We assume that $y_1, y_2,...,y_n$ are independent realizations of a random vector, denoted symbolically as $(Y_1, Y_2,...,Y_p)^T$ which

has a multivariate normal distribution with mean vector μ and covariance matrix Σ;; that is,

$$y_1, y_2, ..., y_n \mid \theta \sim iidN(\mu, \Sigma),$$

where $\theta=(\mu,\Sigma)$ is the unknown parameter. Throughout the chapter, we assume no prior restrictions on θ other than the positive definiteness of $\Sigma(\Sigma>0)$; 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,

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

Maximum-likelihood estimates

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

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

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

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

where

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

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

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

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

and T_2 is the matrix of columnwise sums of squares and crossproducts,

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

Because the multivariate normal is a regular exponential family and the loglikelihood is linear in the elements of T_1 and T_2, we can maximize the likelihood by equating the realized values of T_1 and T_2 with their expectations, $E(T_1) = n\mu$ and $E(T_2) = n(\Sigma+\mu\mu^T)$. This leads immediately to the well known result that the MLEs for μ and Σ are the sample mean vector

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i, \tag{5.5}$$

and the sample covariance matrix

$$\begin{aligned} S &= n^{-1} Y^T Y - \bar{y}\bar{y}^T \\ &= n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T, \end{aligned} \tag{5.6}$$

respectively. Note that S is a biased estimate of Σ, and in practice it is more common to use the unbiased version n $(n-1)^{-1}$ S. Further details on estimation and frequents inference for the multivariate normal model can be found in standard texts on multivariate analysis (e.g. Anderson, 1984).

5.2.2 Bayesian inference under a conjugate prior

The simplest way to conduct Bayesian inference in the complete-data case is to apply a parametric family or class of prior distributions that is *conjugate* to the likelihood function (5.1). A conjugate class has the property that any prior $\pi(\theta)$ in the class leads to a posterior $P(\theta \mid Y) \propto \pi(\theta) L(\theta \mid Y)$ 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 x p data matrix whose rows are iid $N(0,\Lambda)$,, then the matrix of sums of squares and cross-products $A = X^T X$ is said to have a Wishart distribution, and we write

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

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 frequents theory as the sampling distribution of S. For our purposes it will be more convenient to work with the inverted-Wishart distribution. If $A \sim W(m,\Lambda)$ then $B = A^{-1}$ is said to be inverted-Wishart, and we write

$$B \sim W^{-1}(m,\Lambda).$$

Omitting normalizing constants, the inverted-Wishart density for $m \geq p$ can be shown to be

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

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, $c\chi^{-2}_{m,}$, with $c=\Lambda^{-1}$. These and other well-known properties of the Wishart and inverted-Wishart distributions are discussed in many texts on multivariate analysis; an excellent reference is Muirhead (1982).

For our purposes, it will also be useful to know that the mode of the inverted-Wishart density is

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

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,

$$\mu \mid \Sigma \sim \mathcal{N}\left(\mu_0, T^{-1}\Sigma\right), \tag{5.11}$$

where the hyperparameters $\mu_o \in \Re^p$ and $T>0$ are fixed and known. Moreover, suppose that Σ is inverted-Wishart,

$$\Sigma \sim \mathcal{W}^{-1}(m, \Lambda) \tag{5.12}$$

for fixed hyperparameters $m \geq p$ and $\Lambda > 0$. The prior density for θ is then

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

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

$$\begin{aligned} L(\theta \mid Y) \propto \; & |\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{n}{2} \operatorname{tr} \Sigma^{-1} S\right\} \\ & \times \exp\left\{-\frac{n}{2}(\bar{y}-\mu)^T \Sigma^{-1}(\bar{y}-\mu)\right\} \end{aligned} \tag{5.14}$$

Multiplying this likelihood by (5.13) and performing some algebraic manipulation, it follows that $P(\theta|Y)$ has the same

form as (5.13) but with new values for $(\tau, m, \mu_0, \Lambda)$ that is, the complete-data posterior is normal inverted-Wishart,

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

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

where the updated hyperparameters are

$$\begin{aligned} \tau' &= \tau + n, \\ m' &= m + n, \\ \mu_0' &= \left(\frac{n}{\tau + n}\right)\bar{y} + \left(\frac{\tau}{\tau + n}\right)\mu_o, \end{aligned}$$

and

$$\Lambda' = \left[\Lambda^{-1} + nS + \left(\frac{\tau n}{\tau + n}\right)(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T\right]^{-1}.$$

In the special case of $p = 1$, the posterior becomes

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

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

where

$$c' = c + \sum_{i=1}^{n}(y_i - \bar{y})^2 + \left(\frac{\tau n}{\tau + n}\right)(\bar{y} - \mu_0)^2$$

and $\mathrm{c} = \Lambda^{-1}$ 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 p is available, because it results in a posterior distribution for μ centered about $\bar{y}$. Moreover, in some cases a choice of $m < p$ may be attractive as well: see Section 5.2.3 below.

Inferences about the mean vector

By integrating the normal inverted-Wishart density function (5.13) over Σ, one can show that the marginal prior distribution of μ implied by (5.11)-(5.12) is a multivariate t distribution centered at μ_0 with $v = m - p + 1$ degrees of freedom. The mean of this distribution is μ_0 provided that $v > 1$, and the covariance matrix is $(v-2)^{-1}\tau^{-1}\Lambda^{-1}$ provided that $v > 2$. Other properties of this multivariate t distribution are discussed in many texts on multivariate analysis; a good reference is Press (1982). In particular, the marginal prior distribution of any scalar component or linear function of the components of μ is univariate t. Suppose that $\xi = \alpha_T\mu$, where a is a constant vector of length p. The marginal prior distribution of ξ implied by (5.11)-(5.12) is then $(\xi-\xi_0)/\sigma \sim \tau_v$, where $v=m-p+1, \xi_0 = \alpha^T\mu_0$, and

$$\sigma = \sqrt{\frac{\alpha^{\mathrm{T}}\Delta^{-1}\alpha}{\tau v}}.$$

The marginal prior density is

$$P(\xi) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v\sigma^2}}\left[1 + \frac{(\xi-\xi_o)^2}{v\sigma^2}\right]^{-(v+1)/2} \qquad (5.17)$$

where $\Gamma(\cdot)$ denotes the gamma function. After observing Y we can obtain $P(\xi|Y)$, the marginal posterior distribution of ξ, simply by replacing the hyperparameters $(\tau, m, \mu_o, \Lambda)$ in the above expressions with their updated values $(\tau', m', \mu_0', \Lambda')$.

Inferences about the covariance matrix

In many problems the parameters of interest are functions of μ, and Σ is best regarded as a nuisance parameter. On occasion, however, an estimate of Σ is needed. From a Bayesian standpoint there is no universally accepted `best' estimate of Σ. The optimal estimate depends on the choice of a

loss function, and in practice it tends to be difficult or impossible to choose among the various loss functions. Bayesian estimation of a covariance matrix raises some interesting theoretical problems that have yet to be resolved (Dempster, 1969a). If the current state of knowledge about Σ is described by $\Sigma \sim W^{-1}(m,\Lambda)$, then competing estimates include the mean (5.9) and the mode (5.10). To complicate matters further, suppose that the mean μ and the covariance matrix Σ are both of interest, and the current state of knowledge about $\theta = (\mu,\Sigma)$ is represented by the normal inverted-Wishart distribution

$$\mu \mid \Sigma \sim N\left(\mu_0, \tau^{-1}\Sigma\right),$$
$$\Sigma \sim W^{-1}(m,\Lambda).$$

By a calculation that is essentially equivalent to maximizing the multivariate-normal loglikelihood function, one can then show that the joint mode is achieved at $\mu = \mu_0$ and

$$\Sigma = \frac{1}{m+p+2}\Lambda^{-1}.$$

Note that maximizing the joint density for μ and Σ is not equivalent to maximizing the marginal densities for μ and Σ separately.

When a Bayesian estimate of Σ is needed, we will adopt the following rule-of-thumb: if the current state of knowledge about Σ is described by $\Sigma \sim W^{-1}(m,\Lambda)$ irrespective of μ, then estimate Σ by $m^{-1}\Lambda^{-1}$. This represents a compromise between the mean (5.9) and the marginal mode (5.10).

5.2.3 Choosing the prior hyperparameters

A noninformative prior

When no strong prior information is available about θ, it is customary to apply Bayes's theorem with the improper prior

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

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

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

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

A non-Bayesian justification for the use of this prior is that the posterior distribution of the pivotal quantity

$$T^2 = (n-1)(\bar{y}-\mu)^T S^{-1}(\bar{y}-\mu)$$

becomes $(n-1)p(n-p)^{-1}F_{p,n-p}$, the same as its sampling distribution conditionally upon θ (DeGroot, 1970). The ellipsoidal $(1-\alpha)$ 100% HPD region for μ under this prior is identical to the classical $(1-\alpha)100\%$ confidence region for μ from sampling theory, and for inferences about μ the Bayesian and frequents answers coincide. The improper prior (5.18) also arises by applying the Jeffreys invariance principle to μ and Σ (Box and Tiao, 1992).

If our primary interest is not in μ but in Σ, then the frequents justification for using (5.18) as a noninformative prior is not as strong because of the ambiguities involved in estimation of Σ. Notice, however, that if we use our rule-of-thumb that a reasonable estimate for $\Sigma \sim \mathcal{W}^{-1}(m,\Lambda)$, 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 (5.1l)-(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

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

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

which is proper provided that $m + n \geq p$ and $(\Lambda_{-1}+nS) > 0$. Notice that this posterior is very similar to the posterior distribution (5.19)-(5.20) obtained under the standard noninformative prior, except that the covariance matrix Σ has been 'smoothed' toward a matrix proportional to Λ^{-1}. If we take $m = \epsilon$ for some $\epsilon > 0$ and $\Lambda^{-1} = \epsilon S^*$ for some covariance matrix S^*, then our rule-of-thumb estimate of Σ is

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

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,

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

The resulting `prior', which is not really a prior in the Bayesian sense because it is partly determined by the data, has the practical effect of allowing the means and variances to be estimated from the data alone, but smooths the correlation matrix slightly toward the identity. The degree of smoothing is determined by the relative sizes of 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 $z^{\mathrm{T}} = \left(z_1^{\mathrm{T}}, z_2^{\mathrm{T}}\right)$ 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 $\mu T=\left(\mu_1^{\mathrm{T}}, \mu_2^{\mathrm{T}}\right)$ and

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

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

$$\begin{aligned} \mathrm{E}\left(z_2 \mid z_1\right) &= \mu_2 + B_{2.1}\left(z_1 - \mu_1\right) \\ &= \alpha_{2.1} + B_{2.1} z_1 \end{aligned}$$

and covariance matrix $\Sigma_{22Æ1}$, where

$$\begin{aligned} \alpha_{2.1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \\ \mathrm{B}_{2.1} &= -\Sigma_{21}\Sigma_{11}^{-1}, \\ \Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{5.24}$$

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 $\phi_2=(\alpha_{2.1},B_{2.1},\Sigma_{22.1})$ It is easy to show that the transformation $\phi=\phi(\theta)$ is one-to-one, with the inverse transformation $\theta=\phi^{-1}(\phi)$ given by

$$\begin{aligned} \mu_2 &= \alpha_{2.1} + B_{2.1}\mu_1 \\ \Sigma_{12} &= \Sigma_{11}B_{2.1}^{T}, \\ \Sigma_{22} &= \Sigma_{22.1} + B_{2.1}\Sigma_{11}B_{2.1}^{T}. \end{aligned} \tag{5.25}$$

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 $\alpha_{2.1}, B_{2.1}$ and $\Sigma_{22.1} > 0$ will produce a valid $\theta = (\mu, \Sigma)$ with $\Sigma > 0$.

When a probability distribution is applied to $\theta=(\mu,\Sigma)$ .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

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

where J is the Jacobian or first-derivative matrix of the transformation from θ to ϕ, and $\|J\|$means the absolute value of the determinant of J. Notice that α_{21}, $B_{2.1}$ and $\Sigma_{22.1}$ are of the same dimension as μ_2, Σ_{21} and Σ_{22}, respectively, so J can be partitioned as

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

where the submatrices along the diagonal are square. By inspection of (5.24), we see that this matrix has the pattern

$$J = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \times & \times & I & \times & 0 \\ 0 & \times & 0 & \times & 0 \\ 0 & \times & 0 & \times & I \end{bmatrix},$$

where I denotes an identity matrix, 0 denotes a zero matrix and x denotes a matrix that is neither I nor 0. It is a well-known property of determinants that

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

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

$$|J| = \left| \frac{\partial B_{2\cdot 1}}{\partial \Sigma_{21}} \right|. \tag{5.27}$$

With Σ_{11} held fixed, $B_{2.1}=\Sigma_{21}\Sigma^{-1}{}_{11}$ is a linear transformation of Σ_{21}. It can be shown that the Jacobian of the linear transformation from W ($p \times q$) to $Z = WB$ for nonsingular B ($q \times q$) is $|B|^p$ (e.g. Mardia, Kent and Bibby, 1979, Table 2.4.1), and thus

$$\| J \| = \left| \Sigma_{11} \right|^{-p2}. \tag{5.28}$$

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

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

where the elements of H are given by

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

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_{jl}=g_{jl} in row and column k with $h_{jk} = g_j l$-$g_{jl}h_{jk}$. and finally, replace the remaining elements $g_{jl} = g_{lj}$ in the other rows and columns by $h_{jl} = g_{jl} - g_{kl}h_{jk}$. This method is efficient both in terms of computation time and memory, because no storage locations other than the matrix itself are necessary. Because both G and H are symmetric, further savings can be achieved by computing and retaining only the upper-triangular portion of the matrix.

Suppose that a $p \times p$ matrix G is partitioned as

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where G_{11} is $p_1 \times p_1$. After sweeping on positions 1, 2,..., p_1, the matrix becomes

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

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

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

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,..., 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,..., $k-1$,

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

Then

$$|G| = \prod_{k=1}^{p} \gamma_k, \tag{5.30}$$

where γ_1 is taken to be g_{11}, the first element of G. Thus the determinant can be found by computing the product of the

pivots (i.e. the diagonal elements of the matrix) as they appear immediately before the matrix is swept on them (Dempster, 1969b).

It is also convenient to define a *reverse-sweep* operator that returns a swept matrix to its original form. The reverse-sweep operator, denoted by

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

replaces the elements of G with

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

Notice that reverse sweep is remarkably similar to sweep, with the only difference being a minus sign in the calculation of $h_{jk} = h_{kj}$. It is easy to verify that reverse sweep is indeed the inverse of sweep,

$$\mathrm{RSW}[k]\ \mathrm{SWP}[k]\ G = G$$

and that reverse sweep is commutative,

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

Computing alternative parameterizations

From a statistical viewpoint, the sweep operator is highly useful for the following reason: when applied to the parameters of the multivariate normal model, sweep converts a variable from a response to a predictor. Suppose that z is a $p \times 1$ random vector distributed as $N(\mu,\Sigma)$, and we partition it as $z^T = \left(z_1^T, z_2^T\right)$ where z_1 has length p_1. Let us arrange the parameters $\theta = (\mu,\Sigma)$ as a $(p+1) \times (p+1)$ matrix in the following manner,

$$\theta = \begin{bmatrix} -1 & \mu^T \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} -1 & \mu_1^T & \mu_2^T \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{5.32}$$

The reason for placing -1 in the upper-left corner will be explained shortly. To simplify book-keeping, we will allow the row and column indices to run from 0 to p rather than from 1 to $p + 1$, so that the parameters pertaining to the jth variable

will appear in row and column j. Suppose that we sweep this θ-matrix on positions 1, 2,...,p_1; the result will be, by the matrix analogue of (5.29),

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

Comparing this to (5.24), we see that the last p - p_1 rows and columns contain $\alpha_{2.1}$, $B_{2.1}$, and $\Sigma_{22.1}$, the parameters of the conditional distribution of z_2 given z_1,

$$\mathrm{SWP}[1,\ldots p_1]\theta = \begin{bmatrix} -1-\mu_1^{T}\Sigma_{11}^{-1}\mu_1 & \mu_1^{T}\Sigma_{11}^{-1} & \alpha_{2\cdot 1}^{T} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & B_{2\cdot 1}^{T} \\ \alpha_{2\cdot 1} & B_{2\cdot 1} & \Sigma_{22\cdot 1} \end{bmatrix}$$

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,

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

We have thus shown that ϕ=(μ_1,Σ_{11},$\alpha_{2.1}$,$B_{2.1}$,$\Sigma_{22.1}$), expressed in matrix form as

$$\phi = \begin{bmatrix} -1 & \mu_1^{T} & \alpha_{2\cdot 1}^{T} \\ \mu_1 & \Sigma_{11} & B_{2\cdot 1}^{T} \\ \alpha_{2\cdot 1} & B_{2\cdot 1} & \Sigma_{22\cdot 1} \end{bmatrix}. \qquad (5.33)$$

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

$$\mathrm{RSW}[0]\begin{bmatrix} -1 & \mu^{T} \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} 1 & \mu^{T} \\ \mu & \Sigma+\mu\mu^{T} \end{bmatrix}. \qquad (5.34)$$

the parameters of the multivariate normal distribution expressed in terms of the first two moments of z about the

origin. This unswept version of θ is quite useful because it is the natural representation for computing ML estimates. Suppose that Y is an $n \times p$ data matrix whose rows are independent realizations of the random vector z. If we arrange the sufficient statistics $T_1 = Y^T 1$ and $T_2 = Y^T Y$ into a $(p + 1) \times (p + 1)$ matrix

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

then the moment equations for ML estimation set (5.34) equal to $n^{-1}T$. Hence the ML estimate of θ may be computed from the sufficient statistics by

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

Because ML estimates are invariant under transformations of the parameter, the MLE for an alternative parameterization ϕ can be obtained by sweeping $\hat{\theta}$ on the appropriate positions.

	Y_1	Y_2	Y_3	...	Y_p
patterns $s = 1$	1	1	1		1
2	0	1	1		1
·	1	0	1		1
·	0	0	1		1
·	1	1	0		1
·	·	·	·		·
·	·	·	·		·
·	·	·	·		·
·	0	1	0		0
S	1	0	0		0

Figure 5.1. *Matrix of missingness patterns associated with Y with 1 denoting an observed variable and 0 denoting a missing variable.*

5.3 The EM algorithm

When portions of the data matrix Y are missing, ML estimates cannot in general be obtained in closed form; we must resort to iterative computation. The EM algorithm for a multivariate

normal data matrix with an arbitrary pattern of missing values was described by Orchard and Woodbury (1972); Beale and Little (1975); Dempster, Laird and Rubin (1977); and Little and Rubin (1987). Because of its usefulness and its similarities to the simulation algorithms that follow, we describe in detail one possible implementation of EM for incomplete multivariate normal data.

5.3.1 Preliminary manipulations

To simplify notation and facilitate computations, it is helpful at the outset to group the rows of Y by their missingness patterns. A matrix of missingness patterns corresponding to Y is shown in Figure 5.1. We will index the missingness patterns by $s = 1, 2,..., S$, where S is the number of unique patterns appearing in the data matrix. The trivial pattern with all variables missing should be omitted from consideration. Rows of Y that are completely missing contribute nothing to the observed-data likelihood and would only slow the convergence of EM by increasing the fractions of missing information (Section 3.3.2).

For book-keeping purposes it will be helpful to define the following quantities. Let R be an $S \times p$ matrix of binary indicators with typical element rsj, where

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

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

$$O(s) = \{j : r_{sj} = 1\},$$

$$M(s) = \{j : r_{sj} = 0\}.$$

Finally, let I(s) denote the subset of $\{1, 2,..., n\}$ corresponding to the rows of Y that exhibit pattern s. For example, suppose that the data matrix has ten rows with no missing values, and

after sorting these rows are labeled 1,..., 10; the first row of R is then (1, 1,..., 1), and

$$O(1) = \{1, 2, ..., p\},$$
$$M(1) = \varnothing,$$
$$I(1) = \{1, 2, ..10\}.$$

5.3.2 The E-step

Recall that in the E-step of EM, one calculates the expectation of the complete-data sufficient statistics over $P(Y_{mis}|Y_{obs},\theta)$ for an assumed value of θ. These statistics are of the form $\Sigma_i y_{ij}$ and $\Sigma_i y_{ij}y_{ik}$, so to perform the E-step we need to find the expectations of y_{ij} and $y_{ij}y_{ik}$ over $P(Y_{mis}|Y_{obs},\theta)$.

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

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

where $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing subvectors of y_i, respectively. The distribution $P(y_{i(mis)}|y_{i(obs)},\theta)$ is a multivariate normal linear regression of $y_{i(mis)}$ on $y_{i(obs)}$, and the parameters of this regression can be calculated by sweeping the θ–matrix on the positions corresponding to the variables in $y_{i(obs)}$. If row i is in missingness pattern s, then the parameters of $P(y_{i(mis)}|y_{i(obs)},\theta)$ are contained in $\mathrm{SWP}[O(s)]\theta$ in the rows and columns labeled $M(s)$ Let A denote the swept parameter matrix

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

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

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = a_{oj} + \sum_{\kappa \in O(s)} a_{kj} y_{ik},$$

$$\mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) = a_{jk}$$

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

$$E\left(y_{ij} \mid Y_{obs}, \theta\right) = y_{ij},$$

$$\mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) = 0,$$

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

$$\begin{aligned} E\left(y_{ij} y_{i\kappa} \mid Y_{obs}, \theta\right) = \; & \mathrm{Cov}\left(y_{ij}, y_{ik} \mid Y_{obs}, \theta\right) \\ & + E\left(y_{ij} \mid Y_{obs}, \theta\right) E\left(y_{i\kappa} \mid Y_{obs}, \theta\right) \end{aligned}$$

it follows that

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

and

$$E\left(y_{ij} y_{i\kappa} \mid Y_{obs}, \theta\right) = \begin{cases} y_{ij} y_{ik} \text{ for } j, k \in O(s), \\ y_{ij}^{*} y_{ik} \text{ for } j \in M(s), k \in \mathrm{O}(s), \\ \alpha_{jk} + y_{ij}^{*} y_{ik}^{*} \text{ for } j, k \in M(s), \end{cases}$$

where

$$y_{ij}^{*} = \alpha_{oj} + \sum_{k \in O(s)} \alpha_{kj} y_{ik}. \tag{5.36}$$

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 $E(T|Y_{obs},\theta)$ where T is the matrix of complete-data sufficient statistics

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

The elements below the diagonal are not shown and may be omitted from the calculations because they are redundant.

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

5.3.3 Implementation of the algorithm

Once $E(T|Y_{obs},\theta)$ has been found, carrying out the M-step is relatively trivial. For a given value of T the complete-data MLE is $\hat{\theta} = \text{SWP}[0]n^{-1}T$, and the M-step merely carries out this same operation on $E(T|Y_{obs},\theta)$ rather than T. A single iteration of EM can thus be written succinctly as

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

In principle the EM algorithm for incomplete multivariate normal data is completely defined by (5.37), but from a practical standpoint we should still consider how to implement the algorithm in an efficient manner. It is beneficial to keep both processing time and memory usage down, but trade-offs between the two are inevitable; one can always reduce processing time at the expense of additional memory by storing rather than recomputing quantities that must be used repeatedly. The implementation suggested here stores rather than recomputes the portions of $E(T|Y_{obs},\theta)$ that do not depend on θ and thus remain the same for every E-step. This method may not be optimal for any particular dataset, but it is not difficult to program and seems to perform well in a wide variety of situations.

Observed and missing parts of the sufficient statistics

We can express the matrix T as the sum of matrices corresponding to the individual missingness patterns. Let

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

where all sums are taken over $i \in I(s)$, and $n_s = \sum_{i \in I(s)}$ is the sample size in missingness pattern s; then

$$T = \sum_{s=1}^{S} T(s).$$

Each $T(s)$ can be further partitioned into an observed part and a missing part. Notice that the elements of $T(s)$ in the rows and columns labeled $M(s)$ are functions of Y_{mis} and perhaps Y_{obs} whereas the remaining elements of $T(s)$ are functions of Y_{obs} only. Define a new matrix $T_{mis}(s)$ which has the same elements as $T(s)$ in the rows and columns labeled $M(s)$, but with all other elements set to zero, and define $T_{obs}(s)$ to be $T(s) - T_{mis}(s)$. For example, consider a dataset with $p = 3$ variables, and suppose that missingness pattern s has Y_1 and Y_3 observed but Y_2 missing; then

$$T_{obs}(s)=\begin{bmatrix} n_s & \Sigma y_{i1} & 0 & \Sigma y_{i3} \\ & \Sigma y_{i1}^2 & 0 & \Sigma y_{i1}y_{i3} \\ & & 0 & 0 \\ & & & \Sigma y_{13}^2 \end{bmatrix},$$

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

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

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

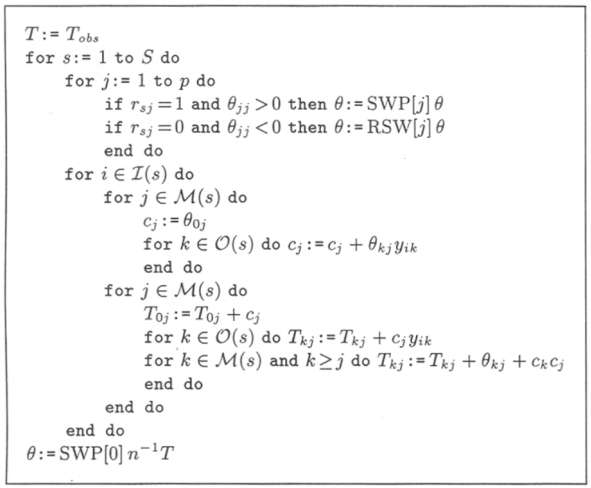


Figure 5.2. *Single iteration of EM for incomplete multivariate normal data, written in pseudocode*

so that $T = T_{obs} + T_{mis}$. The E-step may then be written

$$E(T \mid Y_{obs}, \theta) = T_{obs} + E(T_{mis} \mid Y_{obs}, \theta)$$
$$= \sum_{s=1}^{S} T_{obs}(s) + \sum_{s=1}^{S} E(T_{mis}(s) \mid Y_{obs}, \theta).$$

The elements of T_{obs} can be calculated once at the outset of the program and stored for all future iterations of EM.

An implementation in pseudocode

One possible implementation of an iteration of EM is shown in Figure 5.2. It is written in *pseudocode*, a shorthand language that can be understood by anyone with programming experience and is easily converted into standard languages like Fortran or C. In this pseudocode, the symbol `: =' indicates the operation of assignment; for example, `a: = b' means `set a equal to b.' This implementation requires two $(p + 1) \times (p + 1)$ matrix workspaces: T, into which the expected sufficient statistics are accumulated, and θ, which holds the current estimate of the parameter. For simplicity, the rows and columns of these matrices are labeled from 0 to p rather than

from 1 to $p + 1$. In addition, a single vector of length p, denoted by $c = (c_1,...,c_p)$, is needed as a temporary workspace to hold the values of y_{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[O(s)] condition; for this, we use the convenient book-keeping device that a diagonal element θ_{jj} is negative if and only if θ has been swept on position j. Finally, after the expected sufficient statistics are fully accumulated into T, the new parameter estimate is calculated and stored in θ in preparation for the next iteration.

For efficiency, the code in Figure 5.2 does not calculate the off-diagonal elements of T more than once. If θ and T are stored as two-dimensional arrays, then only the upper-triangular portions should be used, and T_{jk} or θ_{jk} should be interpreted as the (j, k)th element if $\leq k$ or the (k, j)th element if $j > k$. Memory requirements can be reduced by retaining only the upper-triangular parts of T and 0 in packed storage. To reduce the impact of rounding errors, T, θ, and c should be stored in double precision. Rounding errors can also be reduced by centering and scaling the columns of Y at the outset; for example, we could transform the observed data in each column of Y to have mean zero and unit variance before running EM. If the data are centered and scaled, however, we should remember that θ will be expressed on this transformed scale, and for interpretability we may need to transform the estimate of θ back to the original scale at the end of the program.

Starting values

EM requires a starting value $\theta^{(0)}=(\mu^{(0)},\Sigma^{(0)})$for the first iteration. Any starting value may be used provided that $\Sigma^{(0)}$ is

positive definite, but in practice it helps to choose a value that is likely to be close to the mode. Several choices for starting values are described by Little and Rubin (1987). The mean vector and covariance matrix calculated only from the completely observed rows of Y may work well, provided that there are at least $p + 1$ such rows. Another easy method is to use the observed data from each variable to supply starting values for the means and variances, and set the initial correlations to zero; if the columns of Y have been centered and scaled at the outset to have mean 0 and variance 1, then this corresponds to taking $\mu^{(0)}=(0,0,...,0)^T$ and $\Sigma^{(0)}=I$.

Unless the fractions of missing information for some components of θ are very high, the choice of starting value is usually not crucial; when the missing information is low to moderate, the first few iterations of EM tend to bring θ to the vicinity of the mode from any sensible starting value. When writing a program for general use, it is helpful to give the user the option of supplying a starting value, because restarting EM from a variety of locations helps to diagnose unusual features of the observed-data likelihood, such as ridges and multiple modes.

Estimates on the boundary

It sometimes happens, particularly with sparse datasets, that the observed-data likelihood function increases without limit as θ approaches the boundary of the parameter space (i.e. as Σ approaches a singular matrix). When this occurs, the EM algorithm may behave in a variety of ways. In some problems, the elements of θ stabilize and EM appears to converge to a solution on the boundary. In other problems, the program halts due to numeric overflow or attempted division by zero. In yet other problems, the sweeps required for the E-step become numerically unstable as the iterates approach the boundary, and substantial rounding errors are introduced. We have found that these rounding errors sometimes `deflect' θ away from the boundary, causing a sudden large drop in likelihood from one iteration to the next. The iterates may approach the boundary

for a number of steps, deflect away, approach again, and deflect away again in a recurring fashion. If the elements of θ do not appear to have converged after a large number of iterations, then it is advisable to monitor both the loglikelihood (Section 5.3.5) and some aspect of Σ (e.g. the determinant, or the ratio of the largest eigenvalue to the smallest) to determine whether the iterates are approaching the boundary.

When an ML estimate falls on the boundary, it is often helpful to apply a ridge prior and use EM to find the posterior mode as described below.

5.3.4 EM for posterior modes

This EM algorithm can be easily altered to compute a mode of the observed-data posterior distribution rather than an MLE. As discussed in Section 3.2.3, the E-step is no different; only the M-step needs to be modified. The exact form of this modification will depend on the prior distribution applied to θ.

Priors for incomplete data

At this point, it is worthwhile to consider what prior distributions may be appropriate for an incomplete dataset. Because a prior distribution by definition reflects one's state of knowledge about θ before any data are observed, the fact that some data are missing should from a strictly Bayesian viewpoint have no effect whatsoever on the choice of a prior. To the Bayesian purist, any prior that is appropriate for complete data will be equally appropriate for incomplete data. Most statisticians would agree, however, that choosing a prior distribution (including its analytic form) purely by introspection can be difficult, and in practice most priors are chosen at least partly for computational convenience. The normal inverted-Wishart family of prior distributions, described in Sections 5.2.2 and 5.2.3, is computationally convenient for the EM and data augmentation algorithms in this chapter. In general, this family is not conjugate when data are incomplete; the observed data posterior $P(\theta \mid Y_{obs})$ under a

normal inverted-Wishart prior is tractable only in special cases. Yet EM and data augmentation are both easy to implement under this family of priors, because the simplicity of these algorithms depends upon the tractability of the complete-data problem.

When prior information about θ is scanty, we suggest that the customary diffuse prior for complete data,

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

may also be reasonable when some data are missing. Recall from Section 5.2.3 that one important justification for this prior with complete data is that Bayesian and frequents inferences about p coincide. This result does not immediately generalize to incomplete data, but limited experience suggests that Bayesian inferences under this prior may also be approximately valid from a frequents point of view. Little (1988) reports that in the case of bivariate datasets with missing values on one variable generated by an ignorable mechanism, this prior leads to Bayesian inferences about μ that are well-calibrated; the HPD regions tend to have frequency coverage close to the nominal levels. Because this prior treats the variables Y_1, Y_2,..., Y_p in a symmetric fashion, we conjecture that similar results may hold for more complicated multivariate scenarios as well.

When data are sparse and certain aspects of 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 $\Lambda^{-1}=\epsilon \text{Diag } S$, 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 $\Lambda^{-1}=\epsilon I$.

Modications to the M-step

The joint mode of the normal inverted-Wishart distribution,

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

is achieved at μ_0 and $(m+p+2)^{-1}\Lambda^{-1}$ for μ and Σ, respectively (Section 5.2.2). Thus the complete-data posterior mode for $\theta = (\mu, \Sigma)$ under the normal inverted-Wishart prior with hyperparameters $(\tau, m, \mu_0, \Lambda)$, denoted by $\tilde{\theta} = \left(\tilde{\mu}, \tilde{\Sigma}\right)$, is

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

where μ_0, m' and Λ' are the updated versions of the hyperparameters given in Section 5.2.2. By reverse-sweeping the mode on position 0 and equating the result to a matrix of modified sufficient statistics,

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

the mode can be computed as if it were an ML estimate based on $\tilde{T}_1$ and $\tilde{T}_2$ rather than T_1 and T_2. Solving for $\tilde{T}_1$ and $\tilde{T}_2$ and substituting expressions for the updated hyperparameters gives

$$\tilde{T}_1 = \left(\frac{n}{n+\tau}\right)T_1 + \left(\frac{\tau}{n+\tau}\right)n\mu_0$$

and

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

as the modified sufficient statistics, where

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

To modify the EM algorithm shown in Figure 5.2 to compute a posterior mode rather than an MLE, we need only to replace the expected sufficient statistics T_1 and T_2 in the workspace T by the modified versions $\tilde{T}_1$ and $\tilde{T}_2$ immediately before executing the final step $\theta := \text{SWP}[0]n^{-1}T$.

5.3.5 Calculating the observed-data loglikelihood

One of the great advantages of the EM algorithm is that it never requires calculation of the observed-data loglikelihood function or its derivatives. The observed-data likelihood for this problem, discussed in Example 3 of Section 2.3.2, or its logarithm $l(\theta|Y_{obs})$, would be very tedious to differentiate or maximize by gradient-based methods. Evaluation of $l(\theta|Y_{obs})$, at a specific value of θ, however, is not overwhelmingly difficult; the computations required for a single evaluation are comparable to those needed for a single iteration of EM.

It follows from (2.10) that the observed data-loglikelihood function may be written as

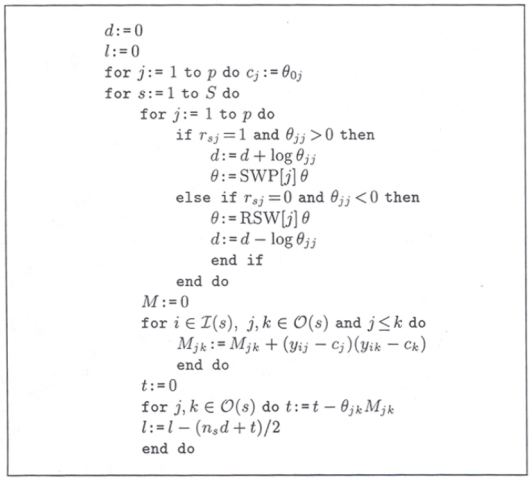
$$\sum_{s=1}^{S}\sum_{i\in I(s)}\left\{-\tfrac{1}{2}\log\left|\Sigma_s^*\right|-\tfrac{1}{2}\left(y_{i(obs)}-\mu_s^*\right)^T\Sigma_s^{*-1}\left(y_{i(obs)}-\mu_s^*\right)\right\},$$

where $y_{i(obs)}$ denotes the observed part of y_i and μ_s^* and Σ_s^* denote the subvector of μ and the submatrix of Σ, respectively, that pertain to the variables that are observed in pattern s. An equivalent but computationally more convenient expression is

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

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

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



```
d := 0
l := 0
for j := 1 to p do c_j := 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 $l(\theta|Y_{obs})$ is shown in Figure 5.3. This algorithm requires a $p \times p$ matrix workspace M to hold values of $M_{<i>s}$</i>>, and a $p \times 1$ vector c for temporary storage of μ. The constants d and t hold $\log\left|\Sigma_s^*\right|$ and $tr\Sigma_s^{*-1}M_s$, respectively, and after execution the loglikelihood value is contained in 1. This program modifies the parameter matrix θ; if necessary, however, the single line

$$\theta := RSW[O(S)\theta$$

may be added at the end of the program, which will return θ to its original state except for rounding errors.

Notice that the algorithm for evaluating $L(\theta|Y_{obs})$ bears a strong resemblance to a single step of EM. An obvious question to ask is whether the two sets of code can be combined, so that an evaluation of the loglikelihood is efficiently woven into EM itself This is certainly possible, but subject to the following caveats. First, the loglikelihood would have to be evaluated at the parameter estimate from the *previous* iteration; that is, we would have to evaluate $l(\theta^{(t)}|Y_{obs})$ as we computed $\theta^{(t+1)}$. Second, notice that a

loglikelihood evaluation requires accumulation of the observed parts of the complete-data sufficient statistics, rather than the expected values of the missing parts. Recall that the EM code in Figure 5.2 assumes that T_{obs} the portion of the expected value of T that does not change over the iterations, has already been computed and stored at the outset of the program. Evaluation of the observed-data loglikelihood, however, requires access to the individual matrices $T_{obs}(s)$ for $s = 1,2,\ldots, S$ which could be very cumbersome to store. If, as in Figure 5.3, the matrices $T_{obs}(s)$ are not stored but effectively recomputed at each iteration, then the proportionate reductions in computing time achieved by combining the two algorithms over running them separately would not be overwhelming.

When EM is used to find a posterior mode rather than an MLE, the function that is guaranteed to be non-decreasing at each iteration is no longer the observed-data likelihood but the observed-data posterior density. The logarithm of the observed-data posterior density is

$$\log P(\theta \mid Y_{obs}) = L(\theta \mid Y_{obs}) + \log \pi(\theta),$$

where unnecessary normalizing constants have been omitted. Thus the log-posterior density may be evaluated by adding $\log \pi(\theta)$ to the result of the algorithm in Figure 5.3. Under a normal inverted-Wishart prior with hyperparameters $(\tau, m, \mu_0, \Lambda)$, this additional term is

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

where

$$M_0 = \Lambda^{-1} + \tau(\mu - \mu_0)(\mu - \mu_0)^T,$$

and unnecessary constants have again been omitted.

5.3.6 Example: serum-cholesterol levels of heart-attack patients

Ryan and Joiner (1994, Table 9.1) report serum-cholesterol levels for $n = 28$ patients treated for heart attacks at a Pennsylvania medical center. For all patients in the sample, cholesterol levels were measured 2 days and 4 days after the attack. For 19 of the 28 patients, an additional measurement

was taken 14 days after the attack. The data are displayed in Table 5.1 (a), with readings at 2, 4 and 14 days denoted by Y_1, Y_2 and Y_3, respectively.

Regarding the complete data as a random sample from a trivariate normal distribution, we applied EM to find the observed-data

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 $\Sigma^{(0)}=(50)^2I$, convergence within four significant digits to

$$\hat{\mu}=\begin{bmatrix}253.9\\230.6\\222.2\end{bmatrix},\ \hat{\Sigma}=\begin{bmatrix}2195 & 1455 & 835.4\\ & 2127 & 1515\\ & & 1953\end{bmatrix}$$

was achieved in just three iterations. Because no data, *e* missing for Y_1 or Y_2, the five parameters $(\mu_1,\mu_2,\sigma_{11},\sigma_{22},\rho_{12})$ converge in a single step regardless of the starting value. Iterates of the four remaining parameters, expressed as μ_3, $\sigma_3=\sqrt{\sigma_{33}}$, ρ_{13} and ρ_{23}, are displayed to six significant digits in Table 5.1 (b).

For estimation of θ, the iterations beyond t = 4 are superfluous because precision beyond three or four digits is rarely necessary. As discussed in Section 3.3.4, however, these additional iterations can be used to estimate elementwise rates of convergence, which are typically equal to the largest fraction of missing information. Elementwise rates of convergence for the four parameters that do not converge in one step, estimated using (3.27), are displayed in Table 5.1 (c). These estimates, which are all close to 47%, do not measure the individual rates of missing information for the four parameters μ_3, σ_3, ρ_{13} and ρ_{23}; rather, they pertain to the function of θ for which the rate of missing information is highest.

Notice that the 47% rate of missing information is somewhat higher than the 9/28 = 32% rate of missing observations for Y_3. Because we know that the parameters pertaining to the joint distribution of (Y_1, Y_2) have no missing information, the 47% rate must pertain to some function of the parameters of the regression of Y_3 on Y_1 and Y_2. It is instructive to consider why the largest rate of missing information exceeds the rate of missing observations for Y_3. A hint is provided by the scatterplot of Y_1 versus Y_2 displayed in

Figure 5.4 (a). The cases having missing values for Y_3 tend to be slightly farther, on average, from the center of the (Y_1, Y_2) distribution than do the cases for which Y_3 is observed. Because they are farther from the center, they exert more influence on the estimates of the regression parameters. A well known measure of influence in linear regression models is provided by the *leverage values*, the diagonal elements of the hat matrix (e.g. Draper and Smith, 1981).

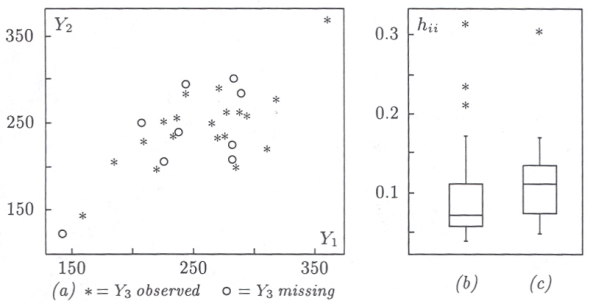


Figure 5.4. *(a) Scatterplot of Y_1 versus Y_2 for all cases, and boxplots of leverage values h_{ii} for cases having (b) Y_3 observed and (c) Y_3 missing.*

The hat matrix for linear regression is defined to be

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

where X is the matrix of predictor variables, in this case a 28×3 matrix containing the observed values of Y_1 and Y_2 and the column vector $1=(1,1,...,)^T$. Boxplots of the diagonal elements hii of H for the cases having Y_3 observed and the cases having Y_3 missing are shown in Figures 5.4 (b) and (c), respectively. The incomplete cases tend to have slightly higher values of h_{ii} and thus exert greater influence on an average, per-case basis over the estimates of the regression parameters.

The parameters of greatest interest in this problem appear to be functions of μ, such as comparisons or contrasts among μ_1, μ_2 and μ_3. Although the rate of missing observations for Y_3 is 32%, we might conjecture that the rate of missing information for μ_3 or a contrast involving μ_3 is substantially lower, because of the high correlations between Y_3 and the completely

observed variables Y_1 and Y_2. The rate of missing information for μ_3, a contrast involving μ_3 or any other function of θ may be estimated in a straightforward manner by multiple imputation; see Section 6.2.1.

5.3.7 Example: changes in heart rate due to marijuana use

Weil et al. (1968) describe a pilot study to investigate the clinical and psychological effects of marijuana use in human subjects. Nine

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 example?

Table 5.3. *ML estimates of means, standard deviations and correlations for the columns of Table 5.2, with eigenvalues of the estimated correlation matrix*

(a) Means

7.38	16.90	14.00	10.60	7.56	−2.58

(b) Standard deviations

8.47	7.72	15.90	21.50	8.98	11.50

(c) Correlation matrix

1.000	−0.301	−0.565	0.385	−0.083	0.211
	1.000	0.620	−0.545	−0.558	0.150
		1.000	−0.860	−0.707	0.199
			1.000	0.705	0.024
				1.000	−0.059
					1.000

(d) Eigenvalues

3.186	1.262	0.890	0.498	0.165	0.000

There are two primary reasons. First, the incomplete cases appear to be very influential. A comparison of the ML estimates of the means in Table 5.3 (a) with the means of the observed data in the columns of Table 5.2 is quite revealing. The large discrepancy for the fourth column (10.6 versus 1.0) demonstrates that a disproportionate amount of information about the mean for that column is provided by subjects 4 and 5. Further examination of Table 5.2 reveals that these two

subjects have rather extreme values in some of the other columns, which gives them high leverage. When these two subjects are deleted, EM converges rapidly and the estimated largest fraction of missing information drops to 45%.

A second reason why this example is problematic is that the complete-data estimation problem is poorly conditioned. The number of subjects $n = 9$ is not much greater than the number of variables $p = 6$. When n and p are nearly equal, it becomes likely that certain linear combinations of the columns of Y will show little or no variability, particularly when the columns are correlated. The multivariate normal model for this example has 27 parameters, too many to be estimated well from a dataset of this size even with complete data. Although certain aspects of θ are poorly estimated, however, we can still make reasonable inferences about the parameters of interest; see Section 5.4.4.

5.4 Data augmentation

5.4.1 The I-step

Data augmentation for incomplete multivariate normal data is remarkably similar to the EM algorithm. The deterministic E- and M-steps are replaced by stochastic I- and P-steps, respectively, where the I-step simulates

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

and the P-step simulates

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

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

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

independently for $i = 1, 2,..., n$. As discussed in Section 5.3.2, if row i is in missingness pattern s then the conditional

distribution of $y_{i(mis)}$ given $y_{i(obs)}$ and θ is multivariate normal with means

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

and covariances

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

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

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

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* $A := \mathrm{Chol}_S A$.

on square submatrices of (5.41). The Cholesky factor of a positive definite matrix A, denoted by

$$C = \mathrm{Chol}A,$$

is an upper-triangular matrix of the same dimension of A having the property that $C^{\mathrm{T}}C = A$. To simulate a random vector z from $N(b, A)$, we may take

$$z = b + (\mathrm{Chol}A)^T z_0,$$

where z_0 is a vector of the same length as z containing independent standard normal variates. A typical Cholesky factorization routine operates on the upper-triangular portion of a symmetric matrix, overwriting it with its Cholesky factor, To draw from the distribution of $y_{i(mis)}$ given $y_{i(obs)}$ and θ,

however, we need to calculate the Cholesky factor of only the square submatrix of (5.41) corresponding to the rows and columns in M(s). For a set S of row labels of a matrix A, let us use

$$A := \text{Chol}_S A \tag{5.42}$$

to indicate the operation that overwrites (the upper triangular portion of) the square submatrix $\{a_{jk}: j,k \in S\}$ with its Cholesky factor, while leaving the remaining elements of A unchanged. A simple algorithm for this operation, adapted from pseudocode given by Thisted (1988, p. 83), is shown in Figure 5.5.

Once the Cholesky factorization is available, the I-step becomes a simple matter of cycling through the missingness patterns $s = 1,\ldots, S$, calculating

$$\text{Chol}_{M(s)}\text{SWP}[O(s)]\theta$$

for each s, and simulating $y_{i(mis)}$ for each $i \in I(s)$. An implementation of the I-step is shown in Figure 5.6. The code simulates the

```
C  T := T_obs
   for s := 1 to S do
       for j := 1 to p do
           if r_sj = 1 and 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 $P(\theta | Y_{obs}, Y_{mis})$ is a normal inverted-Wishart distribution. The P-step of data augmentation, therefore, is merely a simulation of the normal inverted-Wishart distribution,

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

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 $\Sigma \sim \mathcal{W}^{-1}(m, \Lambda)$ is to take $\Sigma = (X^{\mathrm{T}}X)^{-1}$, where X is an $m \times p$ random matrix whose rows are independent draws from $N(0, \Lambda)$. This method cannot be used for non-integer values of m, however, and may be cumbersome for large m because it requires mp random variates. More efficient methods for generating random Wishart matrices are available that require simulation of only $p\ (p + 1)/2$ random variates. One such method relies on a characterization of the Wishart distribution known as the

Bartlett decomposition (e.g. Muirhead, 1982). If $A \sim W(m,I)$ where I is a $p \times p$ identity matrix and $m \geq p$, then we can write $A = B^T B$ where B is an upper triangular matrix whose elements are independently distributed as

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

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

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

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

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

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

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

$$\mu = \mu_0 + \tau^{-1/2} M^{\mathrm{T}} z.$$

where $z \sim N(0,I)$ is a $p \times 1$ vector of independent standard normal variates, results in $\mu|\Sigma \sim N(\mu_0, \tau^{-1}\Sigma)$This method requires the inversion of only the triangular matrix B^T, which can be accomplished via a simple backsolving operation. Note that with the exception of M, all matrices used here are either symmetric or triangular, so memory requirements can be reduced by retaining only their upper-triangular portions in packed storage.

5.4.3 Example: cholesterol levels of heart-attack patients

Recall the example of Section 5.3.6 in which cholesterol measurements were recorded for patients 2, 4 and 14 days after heart attack. The EM algorithm converged rapidly with an estimated largest fraction of missing information equal to 47%. We applied data augmentation to this example under the

noninformative prior (5.18). Output analysis from preliminary runs suggested that the data augmentation algorithm also converged rapidly. For illustration, we ran a single chain for 1100 iterations starting from the ML estimate of θ, discarded the first 100 iterations, and estimated ACFs for a variety of scalar functions of θ over the remaining 1000 iterations. We deliberately chose functions of θ for which the rates of missing information were thought to be high, including:

1. μ_3 and σ_3, the mean and standard deviation of Y_3, respectively;

2. the parameters of the linear regression of Y_3 on Y_1 and Y_2, including the slopes

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

the intercept

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

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

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

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;

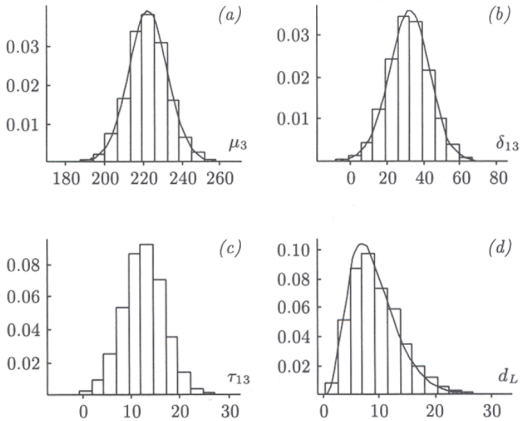


Figure 5.7. *Histograms of sample values of (a)*μ_3*, (b)* δ_{13}*, (C)* τ_{13} *and (d)* d_L *from 5000 consecutive iterations of data augmentation.*

$\delta_{13} = \mu_1 - \mu_3$, 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

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

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

$$k = \frac{\Gamma\left(\frac{n-p+1}{2}\right)}{\Gamma\left(\frac{n-p}{2}\right)\sqrt{\pi(n-p)\sigma^2}}.$$

Rao-Blackwellized density estimates for $\mu_3=(0,0,1)\mu$ and $\delta_{13}=(1,0,-1)\mu$ estimated from the first 1000 iterations after the τ_{13} initial burn-in period are shown superimposed over the histograms in Figure 5.7 (a) and (b). Because τ_{13} is nonlinear its density is somewhat less easy to find, and Rao-Blackwellized estimates for this quantity are not shown.

In addition to μ_3, δ_{13} and τ_{13}, we also calculated and stored values of the likelihood-ratio statistic

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

over the 5000 iterations, where $\hat{\theta}$ is the ML estimate. For large samples, the posterior distribution of d_L is approximately 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 $\delta_{jk}=\mu_j-\mu_k, j,k=1,...,6$. Following the original article be Weil et al. (1968), we will focus attention on the six treatment comparisons below.

15 minutes		90 minutes	
Low *vs.* Placebo	δ_{21}	Low *vs.* Placebo	δ_{54}
High *vs.* Placebo	δ_{31}	High *vs.* Placebo	δ_{64}
High *vs.* Low	δ_{32}	High *vs.* Low	δ_{65}

Data augmentation under the usual noninformative prior (5.18) does not work for this problem; the iterates of θ quickly wander to the boundary of the parameter space, causing numeric overflow. This pathological behavior suggests that the posterior is not proper. To stabilize the inference, we applied a ridge prior as described in Sections 5.2.3 and 5.3.4. After centering and scaling the columns of Y so that the observed data in each column have mean zero and unit variance, we set the hyperparameters of the normal inverted-Wishart prior to $\tau=0$, $m=\epsilon$, and $\Lambda^{-1}=\epsilon I$ for $\epsilon=0.5$. Under this weak prior, EM converges slowly but reliably to a posterior mode in the interior of the parameter space, with the largest fraction of missing information estimated at 95%.

The slow convergence of EM in this example suggests that data augmentation will also converge slowly, and output analysis from a preliminary run confirmed this. Using the same ridge prior, we simulated a single chain beginning at the

posterior mode and monitored a variety of scalar summaries of θ. Time-series plots for δ_{21} and δ_{54} (on the original scale) from the first 100 iterations are shown in Figure 5.8 (a) and (b), respectively. The iterates of δ_{21} appear to approach stationarity quickly, whereas the series for δ_{54} shows long-range dependence. This is not surprising, because δ_{54} is a function of μ_4, and our earlier analysis led us to conjecture that the rate of missing information for μ_4 was very high. Sample ACFs for δ_{21} and δ_{54} estimated from 10 000 iterations are displayed in Figure 5.8 (c) and (d), respectively. Figure 5.8 (d) is typical of the ACFs for other slowly converging functions of θ. For all the functions we examined, the serial correlations effectively died out by lag 50.

The slow convergence in this example should lead us to use extra caution in designing the simulation experiment. Running independent chains from overdispersed starting values would be attractive,

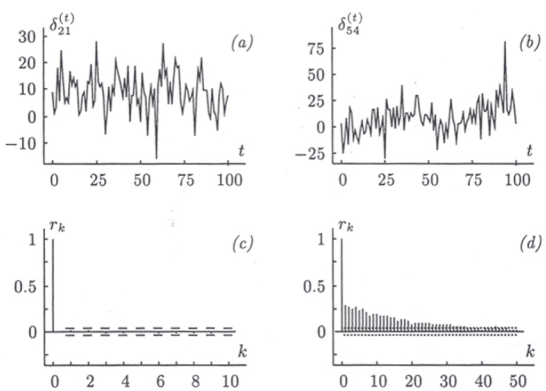


Figure 5.8. *Time-series plots of (a) δ_{21} and (b) δ_{54} over the first 100 iterations of data augmentation, and sample AM for (C) δ_{21} and (d) δ_{54} estimated from 10 000 iterations, with dashes indicating approximate 0.05-level critical values for testing $\rho_k=\rho_{k+1}=\ldots=0$.*

but obtaining overdispersed starting values is not easy. Bootstrap resampling is unlikely to work well, because n is not much larger than p, so the distribution of θ over bootstrap samples will probably bear little resemblance to the observed-data posterior. Sampling from the prior is not possible, because the prior is not a proper probability distribution. Because convergence to stationarity tends to be fastest when the starting value is near the center of the observed-data posterior, we decided to run ten independent chains of 5500 iterations each, starting each chain at the posterior mode. After discarding the first 500 values from each chain, the pth sample quantile for each contrast 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

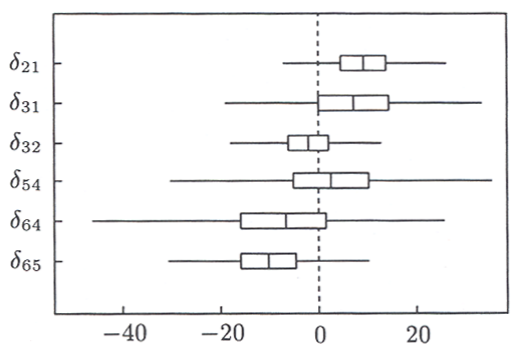


Figure 5.9. *Simulated posterior medians, quantiles and 951% equal-tailed intervals for six contrasts.*

ranged from 0.02 to 0.72, which is quite small relative to the width of the intervals displayed in Figure 5.9, so these simulation results are sharp enough for our purposes.

One could very well argue that the unrestricted multivariate normal model has too many parameters to be estimated from a dataset of this size, and that the unnecessarily large number of nuisance parameters hinders us from making clear inferences about the parameters of interest. Indeed, the long tails exhibited in the marginal posteriors of Figure 5.9, particularly for the two contrasts involving μ_4, suggest that some of the nuisance parameters are very poorly estimated, and we might do well to simplify the model. One possible simplification is to reduce the number of free parameters by applying a priori constraints to Σ. For example, we could require Σ to satisfy the condition of *compound symmetry* (i.e. equal diagonal elements and equal off-diagonal elements). Simulation algorithms for incomplete multivariate normal data with constrained covariance structure are possible, but they are beyond the scope of this book. A slightly different approach would be to specify fixed, additive effects for the rows and columns of the data matrix, and define the parameters of interest to be contrasts among the column effects (Chapter 9).

Yet another possibility is to perform a simple bivariate analysis for each contrast, making inferences about $\delta_{j\kappa}$ using only the data in columns j and k. Under this bivariate approach, it is no longer possible to make joint inferences about the contrasts. Moreover, ignoring the data in columns other than j and k when making inferences about δ_{jk} may tend to introduce nonresponse biases;

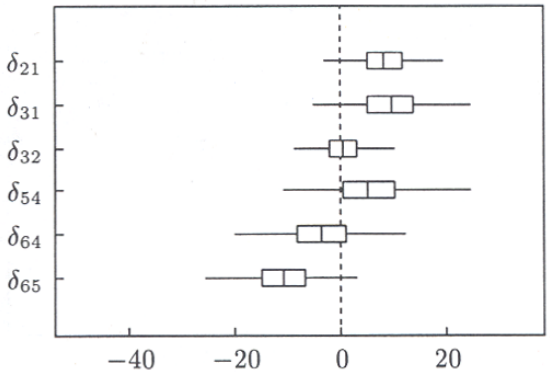


Figure 5.10. *Simulated posterior medians, quantiles and 951% equal-tailed intervals for six contrasts using a bivariate approach.*

the MAR assumption tends to be less plausible for the bivariate dataset than for the one with six variables. The decision whether to include additional variables in an analysis is not always an easy one, particularly for small datasets, and is an important topic worthy of further research.

Simulated posterior quantiles from a bivariate analysis are shown in Figure 5.10. For each contrast, data augmentation was applied to the bivariate dataset under the standard noninformative prior (5.18). Output analyses suggested that convergence to stationarity was rapid. For each contrast, 10 100 steps of a single Markov chain were simulated, beginning from the ML estimate. The first 100 values of the simulated contrast were discarded, and sample quantiles were calculated from the remaining 10 000. The distributions in Figure 5.10 are much narrower than those in Figure 5.9, and there is now a fair amount of evidence that the three contrasts δ_{21}, δ_{31}, and δ_{65} are nonzero.