95227

DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

A BAYESIAN TREATMENT OF MULTIVARIATE NORMAL DATA WITH OBSERVATIONS MISSING AT RANDOM

by

Joseph B. Kadane[1]

Carnegie-Mellon University

and

Ramona L. Trader[2]

University of Maryland

Technical Report No. 308

Department of Statistics

Carnegie-Mellon University

June, 1983

[1] Research supported in part by National Institutes of Justice Grant 81-IJ-CX-0087 and in part by the Office of Naval Research, Contract N00014-82-K-0622.

[2] Research supported in part by a Research Board Award from the Graduate School of the University of Maryland, College Park.

ABSTRACT

This paper provides a Bayesian formulation of the multivariate normal missing data problem when observations are missing at random. Conjugate prior distributions are considered both when the covariance matrix is known and when it is unknown. When the covariance matrix is known, the posterior distribution of the mean is multivariate normal. When the covariance matrix is unknown, the conditional distribution of the mean given the covariance matrix is again multivariate normal. The kernal of the distribution of the covariance matrix is given and is seen to be difficult to integrate analytically. The paper concludes with some comments on the bivariate case, which is partially tractable for nested missing data.

1. Introduction

The problem of missing observations in multivariate normal data is important both because of the central role that the normal distribution plays in statistics and because it provides a base from which to test a variety of heuristics for handling missing data. In this paper a Bayesian approach is employed in order to provide a conjugate analysis for the missing data problem.

Attention is restricted to the special case in which the data are missing at random (as discussed by Rubin, 1976), not because this is necessarily the most important case empirically, but because it is a relatively simple, tractable case. It is assumed that the parameters of the process causing the data to be missing are distinct from those of the normal distribution which generates the data.

Let $\underline{x}$ be a p-variate random vector having a normal distribution with mean μ and covariance matrix Σ; Σ may be known with certainty or unknown. Conjugate prior families for both cases are considered in Section 2. In Section 3, the case when Σ is known is discussed in detail. Remarkably, the conjugate prior and posterior distributions take very simple forms in this case. In Section 4, the more general and difficult case when Σ is unknown is considered. The conditional distribution of μ given Σ is normal, but only the kernel of the marginal distribution of Σ can be given because the resulting integral is difficult to evaluate. Comments on the case when p=2 are given in Section 5.

An important point of comparison for a Bayesian analysis is maximum likelihood estimation, since, asymptotically, posterior distributions are normally distributed with mean given by the maximum likelihood estimate and precision matrix equal to the Fisher information matrix. (See Walker, 1969, for a simple exposition.) Anderson (1957) gives a treatment of maximum likelihood estimation for both mean and covariance matrix unknown in the special case of nested missing data. In the nested case, the p variables are divided into k blocks, where block i has p_i variables and $\sum_{i=1}^{k} p_i = p$. The missing data form a nested pattern if, whenever data are missing in block j, they are missing in blocks j+1,...,k as well. Thus the variables can be reordered so that the missing data take the following form:

$\underline{x}_1$	1	1	1	.	.	.	1	Block 1
.				.				
.				.				
.				.				
$\underline{x}_{p+}$	1	1	1	.	.	.	1	
$\underline{x}_{p_1+1}$	1	1	1	.	.	1	0	Block 2
.				.				
.				.				
.				.				
$x_{p_1+p_2}$	1	1	1	.	.	1	0	
$x_{p_1+p_2+1}$	1	1	0	.	.	.	0	Block 3
.				.				
.				.				
.	1	1	0	.	.	.	0	
.				.				.
.				.				.

"1" indicates presence of data
"0" indicates absence of data

Further work on nested data is given by Rubin (1974 and 1976), and a Bayesian analysis based on a full data conjugate prior distribution is given by Chen (1984).

In the non-nested case, iterative methods for finding maximum likelihood estimates have been given by Hartley and Hocking (1971), Orchard and Woodbury (1972), and Beale and Little (1975). See also Dempster, Laird and Rubin (1977). Since the likelihood surface may have multiple maxima (Murray, 1977), there may be difficulties in using numerical methods based on maximizing the posterior density (Stewart and Sorensen, 1981).

2. Conjugate Opinions When Normal Data Are Missing at Random

Let $\underline{d}$ be a vector of length p comprised of 1's and 0's. Let $\mathcal{D}$ be the set of all such vectors, except for the vector consisting solely of 0's. The $\mathcal{D}$ has 2^p-1 elements. Let $|\underline{d}|$ represent the number of 1's in $\underline{d}$. It will now be useful to introduce some convenient notation from the APL programming language (Gilman and Rose, 1976).

The symbol / represents the reduction operator. If $\underline{d}$ and μ are both of length p, and $\underline{d} \in \mathcal{D}$, then

$$\mu_{\underline{d}} = \underline{d}/\mu \tag{1}$$

is the $\underline{d}$-reduction of μ. It is a vector whose length is $|\underline{d}|$ and whose elements are the

elements of μ corresponding to the 1's in $\underline{d}$. For example, if $\mu' = (1,2,3)$ and $\underline{d}' = (1,0,1)$, then $\mu'_{\underline{d}} = (1,3)$.

Note that $\underline{d}$ can be viewed as a pattern of missing data, where a "1" corresponds to the presence of a data element and a "0" corresponds to its absence. Hence, if μ is the vector of means for all p variables, $\mu_{\underline{d}}$ is the mean vector for the observed variables.

A similar operation can be performed on a matrix Σ. The matrix

$$\underline{d} \; / \; \Sigma$$

is the column reduction of Σ, so the resulting matrix will have as many rows as does Σ, and $|\underline{d}|$ columns. Also,

$$\underline{d} \,⌿\, \Sigma$$

is the row reduction of Σ, yielding a matrix which has as many columns as Σ and $|\underline{d}|$ rows. If Σ is square and symmetric, it is easily verified that

$$\underline{d} \; / \; \underline{d} \,⌿\, \Sigma = \underline{d} \,⌿\, \underline{d} \; / \; \Sigma \tag{2}$$

is again square and symmetric, and has $|\underline{d}|$ rows and columns.

If Σ is the covariance matrix of a set of random variables and $\underline{d}$ represents a pattern of missing data, then

$$\Sigma_{\underline{d}} = \underline{d} \; / \; \underline{d} \,⌿\, \Sigma \tag{3}$$

is the covariance matrix of the observed variables.

Let $\underline{x}_{1,\underline{d}},\ldots,\underline{x}_{N_{\underline{d}},\underline{d}}$ be a sample of $N_{\underline{d}}$ vectors of size $|\underline{d}| \times 1$ that are mutually independent, normally distributed, and each having missing data pattern $\underline{d}$. Then their joint distribution, conditional on knowing $\mu_{\underline{d}}$ and $\Sigma_{\underline{d}}$, is given by

$$f(\underline{x}_{1,\underline{d}},\ldots,\underline{x}_{N_{\underline{d}},\underline{d}} \mid \mu_{\underline{d}},\Sigma_{\underline{d}}) = \frac{\exp[-\frac{1}{2} \sum_{j=1}^{N_{\underline{d}}} (\underline{x}_{j,\underline{d}}-\mu_{\underline{d}})'\Sigma_{\underline{d}}^{-1}(\underline{x}_{j,\underline{d}}-\mu_{\underline{d}})]}{(2\pi)^{N_{\underline{d}}|\underline{d}|/2} \; |\Sigma_{\underline{d}}|^{N_{\underline{d}}/2}} \tag{4}$$

It is well known (see Press, 1982, p. 186) that this likelihood function may be rewritten in terms of the sufficient statistics

$$\bar{\underset{\sim}{x}}_{\underline{d}} = \frac{1}{N_{\underline{d}}} \sum_{j=1}^{N_{\underline{d}}} \underset{\sim}{x}_{j,\underline{d}}$$

$$V_{\underline{d}} = \Sigma_{j=1}^{N_{\underline{d}}} (\underset{\sim}{x}_{j,\underline{d}} - \bar{\underset{\sim}{x}}_{\underline{d}})(\underset{\sim}{x}_{j,\underline{d}} - \bar{x}_{\underline{d}})' \tag{5}$$

$$n_{\underline{d}} = N_{\underline{d}} - 1,$$

so that

$$f(\bar{\underset{\sim}{x}}_{\underline{d}}, V_{\underline{d}} \mid \mu_{\underline{d}}, \Sigma_{\underline{d}}) \propto \frac{|V_{\underline{d}}|^{(n_{\underline{d}} - |\underline{d}| - 1)/2} \exp \{-\tfrac{1}{2} \operatorname{tr} \Sigma_{\underline{d}}^{-1} [V_{\underline{d}} + N_{\underline{d}}(\bar{\underset{\sim}{x}}_{\underline{d}} - \mu_{\underline{d}})(\bar{\underset{\sim}{x}}_{\underline{d}} - \mu_{\underline{d}})']}{|\Sigma_{\underline{d}}|^{N_{\underline{d}}/2}} \tag{6}$$

The conjugate distribution for $\mu_{\underline{d}}$ given $\Sigma_{\underline{d}}$ is

$$f(\underline{u}_{\underline{d}} \mid \Sigma_{\underline{d}}) = \frac{|\Sigma_{\underline{d}}|^{-|\underline{d}|/2}}{2\pi^{|\underline{d}|/2}} \exp [((-b_{\underline{d}})/2)\ (\underline{u}_{\underline{d}} - \underset{\sim}{a}_{\underline{d}})' \Sigma_{\underline{d}}^{-1} (\mu_{\underline{d}} - \underset{\sim}{a}_{\underline{d}})] \tag{7}$$

(Press, 1982, p.83), and the conjugate distribution for $\mu_{\underline{d}}$ and $\Sigma_{\underline{d}}$ jointly is

$$f(\mu_{\underline{d}}, \Sigma_{\underline{d}}) \propto |\Sigma_{\underline{d}}^{-1}|^{(\nu_{\underline{d}}+1)/2} \exp \left\{-\tfrac{1}{2} [(\mu_{\underline{d}} - \underset{\sim}{a}_{\underline{d}})' \Sigma_{\underline{d}}^{-1} (\mu_{\underline{d}} - \underset{\sim}{a}_{\underline{d}}) b_{\underline{d}} + \operatorname{tr} \Sigma_{\underline{d}}^{-1} W_{\underline{d}}]\right\}. \tag{8}$$

In general, let $\theta_{\underline{d}} = \mu_{\underline{d}}$ when Σ, and consequently $\Sigma_{\underline{d}}$, is known, and let $\theta_{\underline{d}} = (\mu_{\underline{d}}, \Sigma_{\underline{d}})$ when Σ is unknown. Thus $\theta_{\underline{d}}$ is the parameter pertaining to the missing data pattern $\underline{d}$. Similarly, let the sufficient statistic $\underline{s}_{\underline{d}} = \bar{\underset{\sim}{x}}_{\underline{d}}$ when Σ is known, and $\underline{s}_{\underline{d}} = (\bar{\underset{\sim}{x}}_{\underline{d}}, V_{\underline{d}})$ when Σ is unknown. Finally, let $\underline{t}_{\underline{d}}$ be the hyperparameter: it is $(\underline{a}_{\underline{d}}, b_{\underline{d}})$ when Σ is known, and $(\underline{a}_{\underline{d}}, b_{\underline{d}}, W_{\underline{d}}, \nu_{\underline{d}})$ when Σ is unknown. The property of conjugacy can be written as

$$\frac{f(\underline{s}_{\underline{d}} \mid \underline{\theta}_{\underline{d}}) f(\underline{\theta}_{\underline{d}} \mid \underline{t}_{\underline{d}})}{\int f(\underline{s}_{\underline{d}} \mid \underline{\theta}_{\underline{d}}) f(\underline{\theta}_{\underline{d}} \mid \underline{t}_{\underline{d}}) d\underline{\theta}_{\underline{d}}} = f(\underline{\theta}_{\underline{d}} \mid \underline{t}_{\underline{d}}, \underline{s}_{\underline{d}}) = f(\underline{\theta}_{\underline{d}} \mid \underline{t}'_{\underline{d}}), \tag{9}$$

where $\underline{t}'_{\underline{d}} = \underline{t}'_{\underline{d}}(\underline{s}_{\underline{d}})$ is the posterior value of the hyperparameter after observing data having sufficient statistic $\underline{s}_{\underline{d}}$.

Now suppose that data are available on various patterns of missing data in $\mathcal{D}$. The missing data are missing at random, and the parameters governing the process by which data are missed are assumed to be independent of θ. Suppose in particular that $N_{\underline{d}} \geq 1$ for each $\underline{d} \in \mathcal{D}_1 \subseteq \mathcal{D}$. Then the likelihood function has the form

$$\prod_{\underline{d} \in \mathcal{D}_1} f(\underline{s}_{\underline{d}} \mid \underline{\theta}_{\underline{d}}). \tag{10}$$

One set of prior distributions that might be considered for this problem is

$$\prod_{\underline{d} \in \mathcal{D}_1} f(\underline{\theta}_{\underline{d}} \mid \underline{t}_{\underline{d}}). \tag{11}$$

With this prior distribution, the posterior distribution is proportional to

$$\prod_{\underline{d} \in \mathcal{D}_1} f(\underline{s}_{\underline{d}} \mid \underline{\theta}_{\underline{d}}) f(\underline{\theta}_{\underline{d}} \mid \underline{t}_{\underline{d}}), \tag{12}$$

i.e., proportional to

$$\prod_{\underline{d} \in \mathcal{D}_1} f(\underline{\theta}_{\underline{d}} \mid \underline{t}'_{\underline{d}}) \tag{13}$$

which is in the same family as the prior distribution. Hence the family in (11) is conjugate to the likelihood given by (10). This analysis holds for missing data from any family having sufficient statistics of fixed dimension.

The next two sections explore the consequences of this theory for the two special cases of interest in this paper: multivariate normal data with known covariance matrix but unknown mean vector, and multivariate normal data with mean vector and covariance matrix both unknown.

3. Normal Data Missing at Random With Unknown Mean and Known Covariance Matrix

It is first helpful to introduce additional APL notation. The symbol \ is the expansion operator. If $\underline{x}$ is a vector of length $|\underline{d}|$, then

$$\underline{d} \setminus \underline{x}$$

is the $\underline{d}$-expansion of $\underline{x}$; it is a vector of length p (the length of $\underline{d}$) having elements of $\underline{x}$ where the corresponding element of $\underline{d}$ is 1, and having zeros elsewhere. Thus if $\underline{d}' = (1,0,1,0,1)$ and $\underline{x}' = (1,2,3)$, then $(\underline{d} \setminus \underline{x})' = (1,0,2,0,3)$. In an analogous way,

$$\underline{d} \setminus \Sigma$$

is the column expansion of Σ,

$$\underline{d} \nmid \Sigma$$

is the row expansion of Σ, and so forth.

The conjugate family given in (11), with factors as in (7), becomes

$$f(\mu) = \prod_{\underline{d} \in \mathcal{D}_1} \frac{|\Sigma_d|^{-|d|/2}}{(2\pi)^{|d|/2}} \exp\left[\frac{-b_d}{2} (\mu_d - \underline{a}_d)' \Sigma_d^{-1} (\mu_d - \underline{a}_d)\right]. \tag{14}$$

Since μ_d appears only in the exponential function, (14) may be rewritten as

$$f(\mu) \propto \exp\left\{\sum_{\underline{d} \in \mathcal{D}_1} \left[\frac{-b_d}{2} (\mu_d - \underline{a}_d)' \Sigma_d^{-1} (\mu_d - \underline{a}_d)\right]\right\}. \tag{15}$$

The sum in (15) is awkward because the quadratic forms which are being added are of different dimensions. They can be made to have the same dimension with the help of the expansion operator.

Let $\underline{a}_d^* = \underline{d} \setminus \underline{a}_d$, and let $H_d = \underline{d} \setminus \underline{d} \setminus b_d \Sigma_d^{-1}$. Then

$$(\mu_d - \underline{a}_d)' \Sigma_d^{-1} (\mu_d - \underline{a}_d) = (\mu - \underline{a}_d^*)' H_d (\mu - \underline{a}_d^*), \tag{16}$$

since zeros have been introduced in H_d corresponding to the pattern of missing data in $\underline{d}$. Anything can appear in the vector that multiplies with H_d in these positions without altering the value of the quadratic form.

Substituting (16) into (15) gives

$$f(\mu) \propto \exp\left\{ \sum_{\underline{d} \in \mathcal{D}_1} [-\tfrac{1}{2} (\mu - \underline{a}_d^*)' H_d (\mu - \underline{a}_d^*)]\right\}. \tag{17}$$

This kernel has the form of the product of a number of normal distribution kernels for μ, with means $\underline{a}_d$ and precision matrices H_d. The product is also a normal kernel, with precision matrix $H = \sum_{\underline{d} \in \mathcal{D}_1} H_d$ and mean $\underline{a} = H^{-1} \sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*$, as long as H is invertible. (H will be invertible if each variable is observed at least once, and this can be assumed without loss of generality.)

Thus, remarkably, the prior family (11) is no more complicated than the full data conjugate prior as long as Σ is known. The nonmissing data combine in the likelihood function and prior distribution to yield a posterior distribution that is a p-dimensional normal distribution.

4. Normal Data Missing at Random With Unknown Mean and Covariance Matrix

In this section the conjugate prior family in (11), with factors given in (8), is studied. In this case,

$$f(\mu, \Sigma) \propto \prod_{\underline{d} \in \mathcal{D}_1} |\Sigma_d^{-1}|^{(\nu_d+1)/2} \exp\left\{-\tfrac{1}{2} [(\mu_d - \underline{a}_d)' \Sigma_d^{-1} (\mu_d - \underline{a}_d) b_d + \operatorname{tr} \Sigma_d^{-1} W_d]\right\}. \tag{18}$$

Since this distribution is the product of normal kernels for μ given Σ, and inverse Wishart kernels for Σ_d, it is natural to think of this joint density as the product of a density on μ given Σ, and a density on Σ.

Using the results of Section 3, $f(\mu|\Sigma)$ is again normal with mean $\underline{a}$ and precision matrix H. Hence,

$$f(\Sigma) \propto \frac{1}{|H|^{1/2}} \left[\prod_{\underline{d} \in \mathcal{D}_1} |\Sigma_d^{-1}|^{(\nu_d+1)/2} \exp\left\{-\tfrac{1}{2}(\operatorname{tr} \Sigma_d^{-1} W_d + \underline{a}_d^{*\prime} H_d \underline{a}_d^*)\right\}\right]$$
$$\exp\{\tfrac{1}{2}(\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*)' H^{-1} (\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*)\}. \tag{19}$$

To simplify the exponential term, first note that

$$\underline{a}_d^{*\prime} H_d \underline{a}_d^* = \underline{a}_d' \Sigma_d^{-1} b_d \underline{a}_d$$
$$= \operatorname{tr} b_d \Sigma_d^{-1} \underline{a}_d \underline{a}_d'.$$

Hence,

$$\operatorname{tr} \Sigma_d^{-1} W_d + \underline{a}_d^{*\prime} H_d \underline{a}_d^* = \operatorname{tr} \Sigma_d^{-1} (W_d + b_d \underline{a}_d \underline{a}_d'). \tag{20}$$

Also, the term

$$(\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*)' H^{-1} (\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*), \text{ where } H = \sum_{\underline{d} \in \mathcal{D}_1} H_d \tag{21}$$

needs to be simplified.

Consider the block-partitioned matrix $\tilde{H}$ which has the form

$$\tilde{H} = \begin{pmatrix} H_1 & 0 & \cdot & \cdot & 0 \\ 0 & H_2 & & & \\ \cdot & & \ddots & & \\ \cdot & & & & \\ 0 & & & & \cdot \end{pmatrix},$$

with the H_d down the main diagonal and matrices of zeros off the diagonal. Let $\tilde{I} = (I, I, \ldots, I)$ be a matrix containing as many $p \times p$ identity matrices as there are matrices down the diagonal of $\tilde{H}$. Then $H = \tilde{I} \tilde{H} \tilde{I}'$. Also, $\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^* = \tilde{I} \tilde{H} a^*$, where $a^{*\prime} = (\underline{a}_1, \underline{a}_2, \ldots)$. Then (21) can be rewritten as

$$(\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*)' H^{-1} (\sum_{\underline{d} \in \mathcal{D}_1} H_d \underline{a}_d^*) = a^{*\prime} \tilde{H} \tilde{I}' (\tilde{I} \tilde{H} \tilde{I}')^{-1} \tilde{I} \tilde{H} a^*. \tag{22}$$

Now rewrite (20) so that it may be combined with (22):

$$\sum_{\underset{\sim}{d} \in \mathcal{D}_1} \operatorname{tr} \Sigma_{\underset{\sim}{d}}^{-1} (W_{\underset{\sim}{d}} + b_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}}') = \operatorname{tr} \sum_{\underset{\sim}{d} \in \mathcal{D}_1} \Sigma_{\underset{\sim}{d}}^{-1} (W_{\underset{\sim}{d}} + b_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}}') = \operatorname{tr} \tilde{I} \tilde{W} \tilde{H} \tilde{I}, \tag{23}$$

where $\tilde{W}$ is a block-diagonal matrix with $\underset{\sim}{d}^{th}$ block given by

$$\underset{\sim}{d} \setminus \underset{\sim}{d} \nrightarrow \operatorname{tr} (W_{\underset{\sim}{d}} + b_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}} \underset{\sim}{a}_{\underset{\sim}{d}}'). \tag{24}$$

Hence the argument of the exponential function in (19) can be written as

$$\begin{aligned} & a^{*\prime} \tilde{H} \tilde{I}' (\tilde{I} \tilde{H} \tilde{I}')^{-1} \tilde{I} \tilde{H} a^* - \operatorname{tr} \tilde{I} \tilde{W} \tilde{H} \tilde{I}' \\ &= \operatorname{tr} a^{*\prime} \tilde{H} \tilde{I}' (\tilde{I} \tilde{H} \tilde{I})^{-1} \tilde{I} \tilde{H} a^* - \operatorname{tr} \tilde{I} \tilde{W} \tilde{H} \tilde{I}' \\ &= \operatorname{tr} \left[(\tilde{I} \tilde{H} \tilde{I}')^{-1} \tilde{I} \tilde{H} a^* a^{*\prime} \tilde{H} \tilde{I}' \right] - \operatorname{tr} \left[(\tilde{I} \tilde{H} \tilde{I}')^{-1} (\tilde{I} \tilde{H} \tilde{I}') \tilde{I} \tilde{W} \tilde{H} \tilde{I}' \right] \\ &= \operatorname{tr} \left[(\tilde{I} \tilde{H} \tilde{I}')^{-1} \tilde{I} \tilde{H} \left[a^* a^{*\prime} - \tilde{I}' \tilde{I} \tilde{W} \right] \tilde{H} \tilde{I}' \right]. \end{aligned} \tag{25}$$

Let $\tilde{W}^* = a^* a^{*\prime} - \tilde{I}' \tilde{I} \tilde{W}$ Then

$$f(\Sigma) \propto \frac{1}{|H|^{1/2}} \sum_{\underset{\sim}{d} \in \mathcal{D}_1} |\Sigma_{\underset{\sim}{d}}^{-1}|^{(\nu_{\underset{\sim}{d}}+1)/2} \exp \left\{ -\tfrac{1}{2} \operatorname{tr} \left[(\tilde{I} \tilde{H} \tilde{I}')^{-1} \tilde{I} \tilde{H} \tilde{W}^* \tilde{H} \tilde{I}' \right] \right\}. \tag{26}$$

The distribution whose kernel appears in (26) may be regarded as a generalization of the inverse Wishart distribution. This kernel apparently cannot be integrated in closed form. And unlike the case of Section 3, the kernel does not reduce to the full-data conjugate case, which would be an inverted Wishart distribution.

5. Data Missing at Random When p = 2

Surprisingly, the bivariate normal case of data missing at random does not, in general, lead to a mathematically tractable form for (26). An examination of the bivariate kernel reveals the complexity inherent in the missing data problem.

Since it is the inverses of the covariance matrices $\Sigma_{\underset{\sim}{d}}$ which appear in (26), it is helpful to reparametrize in terms of $\Sigma_{\underset{\sim}{d}}^{-1}$. Let $\Sigma_{(1,0)}^{-1} = Z_1^*$, $\Sigma_{(0,1)}^{-1} = Z_2^*$, and

$$\Sigma_{(1,1)}^{-1} = \Sigma^{-1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} = \begin{matrix} Z_1 & Z_{12} \\ Z_{12} & Z_2 \end{matrix} = Z .$$

Then it is easy to see that

$$Z_1^* = \frac{1}{(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\, Z_2} = \frac{(Z_1 Z_2 - Z_{12}^2)}{Z_2}$$

and

$$Z_2^* = \frac{1}{(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\, Z_1} = \frac{(Z_1 Z_2 - Z_{12}^2)}{Z_1} . \tag{27}$$

In other words, the precisions are defined differently in terms of σ_1^2, σ_2^2, and σ_{12} in different dimensions.

Also, let (a_1, b_1, w_1, ν_1) be the hyperparameters corresponding to $\underset{\sim}{d}' = (1,0)$ and a_2, b_2, w_2, ν_2 be the hyperparameters corresponding to $\underset{\sim}{d}' = (0,1)$. For the totally observed data, represented by $\underset{\sim}{d}' = (1,1)$, let the hyperparameters be $(\underset{\sim}{a}_3, b_{12}, W_3, \nu_{12})$, where

$$\underset{\sim}{a}_3' = (a_{31}, a_{32})$$

and

$$W_3 = \begin{pmatrix} W_{31} & W_{12} \\ W_{12} & W_{32} \end{pmatrix}$$

Then, since the Jacobian of the transformation from Σ to Σ^{-1} is $|\Sigma^{-1}|^{-(p+1)} = |Z|^{-3}$, (26) may be rewritten as

$$\begin{aligned} f(Z) \propto & \frac{1}{\begin{vmatrix} b_1 Z_1^* + b_{12} Z_1 & b_{12} Z_{12} \\ b_{12} Z_{12} & b_2 Z_2^* + b_{12} Z_2 \end{vmatrix}^{1/2}} Z_1^{*(\nu_1+1)/2} \; Z_2^{*(\nu_2+1)/2} \\ & \cdot (Z_1 Z_2 - Z_{12}^2)^{[(\nu_{12}+1)/2] - 3} \exp \Big[-\tfrac{1}{2} \Big\{ (w_{31} + b_{12} a_{31}^2) z_1 \\ & + 2(w_{12} + b_{12} a_{31} a_{32}) Z_{12} + (w_{32} + b_{12} a_{32}^2) Z_2 + (w_1 + b_1 a_1^2) Z_1^* \end{aligned}$$

$$+ (w_2+b_2a_2^2)Z_2^{*} - \frac{Z_1Z_2}{(Z_1Z_2-Z_{12}^2)\left[\left\{b_1b_2+(b_1+b_2+b_{12})b_{12}\right\}Z_1Z_2-b_1b_2Z_{12}^2\right]}$$

$$\cdot \Big[b_1^2b_{12}a_1a_{31}Z_1^{*}Z_2+b_2^2b_{12}a_2a_{32}Z_1Z_2^{*}+b_1^2b_2a_1^2Z_1^{*2}Z_2^{*}$$

$$+ b_1b_2^2a_2^2Z_1^{*}Z_2^{*2}+b_1b_2b_{12}Z_1^{*}Z_2^{*}f_1(Z_1,Z_2,Z_{12})$$

$$+ b_1b_{12}^2Z_1^{*}f_2(Z_1,Z_2,Z_{12}) + b_2b_{12}^2Z_2^{*}f_3(z_1,z_2,z_{12})$$

$$+ b_{12}^3(Z_1Z_2-Z_{12}^3)\left[a_{31}^2Z_1+2a_{31}a_{32}Z_{12}+a_{32}^2Z_2\right]\Big]\Big\}\Big], \qquad (28)$$

where

$$f_1(Z_1,Z_2,Z_{12}) = 2a_1a_{31}Z_1 + 2a_2a_{32}Z_2 + 2a_1a_{32}Z_{12} + 2a_2a_{31}Z_{12} - a_1a_2Z_{12}$$

$$f_2(Z_1,Z_2,Z_{12}) = a_1a_{31}(Z_1Z_2-Z_{12}^2) + (a_{31}^{\;2}-a_1a_{31})Z_{12}^{\;2} + 2a_{31}a_{32}Z_2Z_{12} + a_1a_{31}Z_1Z_{12} + a_{32}^2Z_2^2$$

$$f_3(Z_1,Z_2,Z_{12}) = a_2a_{32}(Z_1Z_2-Z_{12}^2) + (a_{32}^2-a_2a_{32})\, Z_{12}^2 + 2a_{31}a_{32}Z_1Z_{12} + a_2a_{32}Z_2Z_{12} + a_{31}^2 + Z_1^2 .$$

Unfortunately, substitution for Z_1^{*} and Z_2^{*} does not allow (28) to be written in a simple form, and there is no apparent way to integrate (28) with respect to Z_1, Z_2 and Z_{12} over the region given by $Z_1 > 0$, $Z_2 > 0$, and $(Z_1Z_2-Z_{12}^2) > 0$ except by using numerical methods.

If the missing data are assumed to form a nested pattern, the results of Chen (1984) may be used to find the integrating constant for the kernel in (18). Assume that data are missing on X_2 but not on X_1. The kernel can be reparametrized from $(\mu_1,\sigma_2,\sigma_1^2,\sigma_2^2,\sigma_{12})$ to the distinct parameters $(\mu_1,\sigma_1^2,\alpha,\beta,\sigma_{22\cdot1}^2)$, where

$$\alpha = \mu_2 - \beta\mu_1,$$

$$\beta = \sigma_{12}/\sigma_1^2,$$

and

$$\sigma_{22\cdot1}^2 = \sigma_2^2 - \sigma_{12}^2/\sigma_1^2 .$$

Chen demonstrates that these distinct parameters are independent when the missing data form a nested pattern. (Chen actually assumes a complete data conjugate prior instead of the more general conjugate prior described herein. However, in the case of nested missing data, his posterior kernel has the same form as the general prior kernel given in (18).)

In our notation, Chen's results imply that

(a) $\sigma_1^{-2} \sim \text{Wishart}\left(\frac{(b_1+b_{12})(w_1+w_{31})+b_1b_{12}(a_1+a_{31})^2}{b_1+b_{12}},\ \nu_1+\nu_{12}+1\right)$

(b) $\mu_1 \mid \sigma_1^2 \sim \text{Normal}\left(\frac{b_1a_1+b_{12}a_{31}}{b_1+b_{12}},\ (b_1+b_{12})^{-1}\ \sigma_1^2\right)$

(c) $(\alpha,\beta) \mid \sigma_{22\cdot1}^2 \sim \text{matrix Normal}\left(\left(a_{32} - \frac{w_{12}}{w_{31}}a_{31},\ \frac{w_{12}}{w_{31}}\right),\ \sigma_{22\cdot1}^2\begin{pmatrix} b_{12}^{-1} + a_{31}^2/w_{31} & -a_{31}/w_{31} \\ -a_{31}/w_{31} & 1/w_{31}\end{pmatrix}\right)$

(d) $\sigma_{22\cdot1}^2 \sim \text{Wishart}\ (w_{32} - w_{12}^2/w_{31},\ \nu_{12}+1).$

Therefore the integrating constant for the kernel in (18) is known in the special case of nested missing data. By pulling out the conditional normal density on (μ_1,μ_2) given $(\sigma_1^{\,2},\sigma_2^{\,2},\sigma_{12})$, the complete marginal density for $(\sigma_1^{\,2},\sigma_2^{\,2},\sigma_{12})$ can also be obtained, but as indicated by (19), its form is quite cumbersome.

The density for (z_1,z_2,z_{12}) has a simpler form in this special case and appears to be an interesting generalization of the Wishart density:

$$f(z_1,z_2,z_{12}) = A^{\frac12(\nu_1+\nu_{12}+1)}\,[w_{32} - w_{12}^2/w_{31}]^{\frac12(\nu_{12}+1)}$$

$$\cdot\, 2^{-\frac12(\nu_1+2\nu_{12}+2)}\frac{\sqrt{(b_1+b_{12})}\sqrt{w_{31}}}{\sqrt{2\pi}}\ \frac{1}{\Gamma((\nu_1+\nu_{12}+1)/2)\ \Gamma((\nu_{12}+1)/2)}$$

$$\cdot\, z_2^{-(\nu_1+1)/2}\,(z_1z_2-z_{12}^2)^{(\nu_1+\nu_{12}-4)/2}\exp\left\{\tfrac12\left[\beta_1z_1+\beta_{12}z_{12} + \beta_2z_2 + \beta_3\frac{z_{12}^{\;2}}{z_2} + \beta_4\frac{1}{z_2}\right]\right\}$$

where

$$A = \left[\frac{(b_1+b_{12})(w_1+w_{31})+b_1b_{12}(a_1+a_{31})^2}{(b_1+b_{12})}\right]^{\frac12(\nu_1+\nu_{12}+1)}$$

$$\beta_1 = \frac{(a_1-a_{31})^2 b_1 b_{12} - a_1^2 b_1^2 - 2w_1(b_1+b_{12})}{b_1+b_{12}}$$

$$\beta_{12} = 2a_{31}a_{32}(b_{12}-1)-2w_{12}$$

$$\beta_2 = a_{32}^2(b_{12}-1)-w_{32}$$

$$\beta_3 = \frac{(a_{31}-a_1)^2 b_1 b_{12} + a_1^2 b_1^2 + w_1(b_1+b_{12})}{b_1+b_{12}}$$

and

$$\beta_4 = \frac{a_1^2 b_1^2}{b_1+b_{12}} .$$

Chen's results apply to any nested configuration of missing data, so the complete density for (18) may always be found in the nested missing data case. Chen's method does not generalize to the non-nested case, however, since distinct parameters can no longer be identified. A direct method of integration is still required in the general case.

6. Conclusion

We have shown that the conjugate-prior analysis of multivariate normal data missing at random is tractable when the covariance matrix can be assumed to be known, and difficult otherwise. The kernel for a useful generalization of the inverse Wishart distribution is given, but we do not know how to integrate it in non-nested cases. Perhaps stimulated by this notice of an unsolved problem, others will take up the challenge.

REFERENCES

Anderson, T. W. (1957) "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing." Journal of the American Statistical Association, **52**, 200-203.

Beale, E. M. L. and Little, R. J. A. (1975) "Missing values in multivariate analysis," Journal of the Royal Statistical Society, Series B, **37**, 129-145.

Chen, C. F. (1984) "Factorization of Normal-Wishart Priors and Bayesian Multivariate Analysis with Nested Missing Data," to appear in P. Goel and A Zellner, eds., Essays in Honor of Bruno DeFinetti.

Dempster, A.P., N. Laird and D. Rubin (1977). "Maximum likelihood estimation from incomplete data with the EM algorithm" (with discussion). Journal of the Royal Statistical Society, Series B, **39**, 1-39.

Gilman, L. and Rose, A. J. (1976) APL: An Interactive Approach. (Second Edition, Revised Reprinting). New: John Wiley and Sons.

Hartley, H. O. and Hocking, R. R. (1971) "Incomplete data analysis." Biometrics, **27**, 783-823.

Murray, G. D. (1977) "Contribution to discussion of paper by A. P. Dempster, N. M. Laird, and D. B. Rubin." Journal of the Royal Statistical Society, Series B, **39**, 27-28.

Orchard, T. and Woodbury, M. A. (1972) "A missing information principle. Theory and applications." Proceedings of the Sixth Berkeley Symposium, **1**, 697-715.

Press, S. J. (1982) Applied Multivariate Analysis (Second Edition) Malabar, Fla: Robert E. Krieger Publishing Company.

Rubin, D. B. (1974) "Characterizing the estimation of parameters in incomplete data problems." Journal of the American Statistical Association, **69**, 467-474.

Rubin, D. B. (1976) "Inference and missing data." Biometrika, **63**, 581-592.

Stewart, W. E. and Sorensen, J. P. (1981) "Bayesian estimation of common parameters from multiresponse data with missing observations." Technometrics, **23**, 131-141.

Walker, A. M. F. (1969) "On the asymptotic behavior of posterior distributions." Journal of the Royal Statistical Society, Series B, **31**, 80-88.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 308	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Bayesian Treatment of Multivariate Normal Data with Observations Missing at Random		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Joseph B. Kadane Ramona L. Trader		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0622
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE June, 1983 13. NUMBER OF PAGES 14
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayesian Analysis; Missing Normal Data; Missing at Random; Conjugate Prior Distributions; Inverse Wishart Distribution

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper provides a Bayesian formulation of the multivariate normal missing data problem when observations are missing at random. Conjugate prior distributions are considered both when the covariance matrix is known and when it is unknown. When the covariance matrix is known, the posterior distribution of the mean is multivariate normal. When the covariance matrix is unknown, the conditional distribution of the mean given the covariance matrix is again multivariate normal. The kernal of the distribution of the

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (continued)

covariance matrix is given and is seen to be difficult to integrate analytically. The paper concludes with some comments on the bivariate case, which is partially tractable for nested missing data.

END