The Calculation of Posterior Distributions by Data Augmentation

MARTIN A. TANNER and WING HUNG WONG*

The idea of data augmentation arises naturally in missing value problems, as exemplified by the standard ways of filling in missing cells in balanced two-way tables. Thus data augmentation refers to a scheme of augmenting the observed data so as to make it more easy to analyze. This device is used to great advantage by the EM algorithm (Dempster, Laird, and Rubin 1977) in solving maximum likelihood problems. In situations when the likelihood cannot be approximated closely by the normal likelihood, maximum likelihood estimates and the associated standard errors cannot be relied upon to make valid inferential statements. From the Bayesian point of view, one must now calculate the posterior distribution of parameters of interest. If data augmentation can be used in the calculation of the maximum likelihood estimate, then in the same cases one ought to be able to use it in the computation of the posterior distribution. It is the purpose of this article to explain how this can be done.

The basic idea is quite simple. The observed data y is augmented by the quantity z, which is referred to as the latent data. It is assumed that if y and z are both known, then the problem is straightforward to analyze, that is, the augmented data posterior $p(\theta \mid y, z)$ can be calculated. But the posterior density that we want is $p(\theta \mid y)$, which may be difficult to calculate directly. If, however, one can generate multiple values of z from the predictive distribution $p(z \mid y)$ (i.e., multiple imputations of z), then $p(\theta \mid y)$ can be approximately obtained as the average of $p(\theta \mid y, z)$ over the imputed z's. However, $p(z \mid y)$ depends, in turn, on $p(\theta \mid y)$. Hence if $p(\theta \mid y)$ was known, it could be used to calculate $p(z \mid y)$. This mutual dependency between $p(\theta \mid y)$ and $p(z \mid y)$ leads to an iterative algorithm to calculate $p(\theta \mid y)$. Analytically, this algorithm is essentially the method of successive substitution for solving an operator fixed point equation. We exploit this fact to prove convergence under mild regularity conditions.

Typically, to implement the algorithm, one must be able to sample from two distributions, namely $p(\theta \mid y, z)$ and $p(z \mid \theta, y)$. In many cases, it is straightforward to sample from either distribution. In general, though, either sampling can be difficult, just as either the E or the M step can be difficult to implement in the EM algorithm. For $p(\theta \mid y, z)$ arising from parametric submodels of the multinomial, we develop a primitive but generally applicable way to approximately sample θ. The idea is first to sample from the posterior distribution of the cell probabilities and then to project to the parametric surface that is specified by the submodel, giving more weight to those observations lying closer to the surface. This procedure should cover many of the common models for categorical data.

There are several examples given in this article. First, the algorithm is introduced and motivated in the context of a genetic linkage example. Second, we apply this algorithm to an example of inference from incomplete data regarding the correlation coefficient of the bivariate normal distribution. It is seen that the algorithm recovers the bimodal nature of the posterior distribution. Finally, the algorithm is used in the analysis of the traditional latent-class model as applied to data from the General Social Survey.

KEY WORDS: Bayesian inference; Monte Carlo sampling; Imputation; Correlation coefficient; Latent class analysis; Convergence results; Dirichlet sampling.

* Martin A. Tanner is Assistant Professor, Departments of Statistics and Human Oncology, University of Wisconsin, Madison, WI 53706. Wing Hung Wong is Associate Professor, Department of Statistics, University of Chicago, Chicago, IL 60637. This work was supported by National Science Foundation Research Grant MCS-8301459 and National Institutes of Health Grant R23 CA35464. The authors thank David Wallace for suggesting the covariance matrix example and Stephen Stigler for his helpful comments.

1. INTRODUCTION

This article introduces an iterative method for the computation of posterior distributions. The method applies whenever the data can be augmented in such a way that (a) it becomes easy to analyze the augmented data and (b) it is easy to generate the augmented data given the parameter. Let y denote the observed data whose distribution depends on a parameter vector θ. Suppose that there is a way to augment y with latent data z (unobserved) so that the augmented data, $x = (y, z)$, is straightforward to analyze [i.e., the augmented data posterior density, $p(\theta \mid x)$, is of known form]. The method consists of iterating the following two steps: (a) Given the current guess of the posterior distribution of θ given y, generate a sample of $m > 0$ latent data patterns from the predictive distribution of z given y. (b) Update the posterior of θ, given y, to be the mixture of the m augmented data posteriors.

The sample size m can change from iteration to iteration. If m is always taken to be very large, then the algorithm can be interpreted as the method of successive substitution for solving a fixed point problem characterizing the true posterior distribution. The updated posterior at the end of the iterations can then be taken to be a close approximation of the true posterior distribution. When m is small, however, we will need to pool over the latent data patterns generated near the end of the iterations to get a reasonable approximation to the true posterior distribution.

The plan of the article is as follows. In the remaining part of this introduction, we discuss data augmentation as a general tool for the analysis of data in complex models. At the same time, we will review relevant literature. In Section 2, we motivate and present the basic algorithm and illustrate the steps of the algorithm in the context of a simple example. In Section 3, we apply the method to the problem of inference on the covariance matrix of the multivariate normal distribution with missing values. In Section 4, we introduce the Dirichlet sampling procedure as a way to facilitate the approximate sampling from the posterior distribution in complex models of multinomial data. In Section 5, this procedure is applied to the study of social survey data modeled by a log-linear model with a latent variable. We also use this example to illustrate and discuss issues of identifiability in Bayesian modeling. In Section 6, we return to the study of the basic algorithm. We will discuss the uniqueness of the fixed point characterization that motivates the basic algorithm and will present convergence results for the algorithm. The reader who

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

is interested in applications, rather than theoretical details, may skip Section 6 without loss of continuity. In Section 7, variations of the basic algorithm will be presented and issues in its practical implementation will be discussed.

We now turn to the idea of data augmentation. In well-designed experiments, it often happens that, if not for the presence of missing values, the estimation of parameters will be straightforward. In currently popular terminology, the observed data are called the incomplete data. The complete data refer to the set of missing and observed values. Through the work of many authors, a large body of iterative techniques for maximum likelihood estimation from incomplete data has recently emerged, all of which exploit the simple structure of the complete data problem. This area is elegantly synthesized and further developed in the influential paper of Dempster, Laird, and Rubin (1977), in which references to earlier research can be found. Briefly, based on a current estimate of the parameter value, the method seeks to compute the expected value of the log-likelihood of the complete data and then maximizes the log-likelihood to obtain the updated parameter value. Dempster et al. called this approach the EM algorithm because of the expectation and maximization calculations involved. Although the details of the EM algorithm are not of direct interest for the present article, the aspect of Dempster et al. (1977) that is most important for our purpose is the impressive list of examples, which includes missing data problems, mixture problems, factor analysis, iteratively reweighted least squares, and many others. In each example, enough detail is presented to show how the EM algorithm can be applied. By these examples, the authors make it clear that even in cases that at first sight may not appear to be an incomplete data problem, one may sometimes still profit by artificially formulating it as such to facilitate the maximum likelihood estimation.

It seems that the potential usefulness of this problem formulation is still not fully appreciated by some practitioners, possibly because their problems appear to have little to do with missing values or incomplete data. For this reason, we will use the terms observed data (denoted by y) and augmented data (denoted by x), instead of incomplete data (y) and complete data (x). We will also use the term latent data (z) to denote the unobserved supplementary data needed for the augmentation of y so that the augmented data, $x = (y, z)$, is straightforward to analyze.

In general, this data augmentation scheme is used for the calculation of maximum likelihood estimates or posterior modes. For making inferential statements, the validity of the normal approximation is assumed and the precision of the estimate is given by the observed Fisher information. In most cases, however, it is not possible to obtain the Fisher information directly from the basic EM calculations and one must do further calculations to obtain standard errors [see the discussion following Dempster et al. (1977); see also Louis (1982)]. Except in simple cases, it is difficult to obtain an indication to the validity of the normal approximation.

In the present article, we are interested in the entire likelihood or posterior distribution, not just the maximizer and the curvature at the maximizer. The method we propose exploits the simplicity of the posterior distribution of the parameter given the augmented data, just as the EM algorithm exploits the simplicity of maximum likelihood estimation given the complete data. Even in large sample situations, when the normal approximation is expected to be valid, it would still be comforting to note that the obtained posterior is consistent with the picture given by the maximum likelihood analysis. In small sample situations, the pitfalls of maximum likelihood estimation are well known, and the present method will provide a way of improving inference based on the entire posterior distribution (or the entire likelihood). The examples presented in this article will illustrate that a few steps of the iterative algorithm will provide a diagnostic for the adequacy of the normal approximation for the maximum likelihood estimate.

In practice, one is often interested in the marginal distribution of various parameters of interest. Even if one can evaluate the joint posterior distribution, obtaining the marginal distribution can be difficult and is a topic of current interest (Smith, Skene, Shaw, Naylor, and Dransfield 1985; Tierney and Kadane 1985; Zellner and Rossi 1984). In the data augmentation setup, one is faced with the additional complication that the posterior distribution given the observed data may not be expressible in closed form. Ideally, one would want to choose the augmentation such that the posterior given the augmented data can be sampled from with ease. In cases where this cannot be done, one would have to resort to approximate sampling methods. The Dirichlet sampling scheme discussed in Section 4 provides a simple approach for approximate sampling in the case of multinomial data. Moreover, the recent works on marginalization referred to previously may potentially be helpful in this regard.

We wish to draw the reader's attention to the concurrent and independent work of K. H. Li (1985a,b), who has devised an algorithm for doing multiple imputation of missing values that is very similar in its formal structure to our method. Whereas the main goal in the present article is to exploit the data augmentation formulation in the Bayesian inference of parameters, in Li's work, the initial focus, as well as sources of examples, have been the imputation of missing values. Thus the essential difference is that Li's method exploits the simplicity of the distribution of one component of the missing values given both the observed data and the remainder of the missing values, whereas our method relies on the simplicity of the posterior distribution of the parameter given the augmented data. Upon completion of both works, it was realized that when one identifies the unknown parameters as part of the missing values, then the two algorithms become essentially the same.

Our present results are, to a considerable extent, anticipated in the work of Rubin. In particular, the two key concepts of data augmentation and multiple imputation have been advocated and studied by Rubin in a series of

papers on inference in the presence of incomplete data (Dempster et al. 1977; Rubin 1978, 1980).

2. THE BASIC ALGORITHM

The algorithm is motivated by the following simple representation of the desired posterior density:

$$p(\theta \mid y) = \int_Z p(\theta \mid z, y)p(z \mid y)\, dz, \qquad (2.1)$$

where $p(\theta \mid y)$ denotes the posterior density of the parameter θ given the data y, $p(z \mid y)$ denotes the predictive density of the latent data z given y, and $p(\theta \mid z, y)$ denotes the conditional density of θ given the augmented data $x = (z, y)$. The predictive density of z can, in turn, be related to the desired posterior density by

$$p(z \mid y) = \int_\Theta p(z \mid \phi, y)p(\phi \mid y)\, d\phi. \qquad (2.2)$$

In the above equations, the sample space for the latent data z is denoted by Z and the parameter space for θ is denoted by Θ. (From this point on the range of integration will be omitted from the expressions, as it will be specified implicitly by the differentials dz or $d\phi$.) Substituting (2.2) into (2.1) and interchanging the order of integration, we see that $p(\theta \mid y)$ must satisfy the integral equation

$$g(\theta) = \int K(\theta, \phi)g(\phi)\, d\phi,$$
$$\text{where } K(\theta, \phi) = \int p(\theta \mid z, y)p(z \mid \phi, y)\, dz. \qquad (2.3)$$

Let T be the integral transformation that transforms any integrable function f into another integrable function Tf by the equation

$$Tf(\theta) = \int K(\theta, \phi)f(\phi)\, d\phi. \qquad (2.4)$$

The method of successive substitution for solving (2.3) thereby suggests an iterative method for the calculation of $p(\theta \mid y)$. Namely, start with any initial approximation $g_0(\theta)$ to $p(\theta \mid y)$, and successively calculate

$$g_{i+1}(\theta) = (Tg_i)(\theta). \qquad (2.5)$$

In Section 6 we will show that under mild conditions the g_i's calculated this way will always converge to the desired posterior $p(\theta \mid y)$.

If the integral transform (2.5) can be calculated analytically, then the implementation of this method is straightforward. Unfortunately, this is seldom the case. In typical cases, the integration in (2.1), (2.2), and (2.5) is difficult to perform analytically. It is often possible, however, by the Monte Carlo method, to perform the integration. Equation (2.1) then motivates the following iterative scheme: Given the current approximation g_i to $p(\theta \mid y)$,

(a) generate a sample $z^{(1)}, \ldots, z^{(m)}$ from the current approximation to the predictive density $p(z \mid y)$
(b) update the current approximation to $p(\theta \mid y)$ to be the mixture of conditional densities of θ given the augmented data patterns generated in (a), that is,

$$g_{i+1}(\theta) = m^{-1} \sum_{j=1}^{m} p(\theta \mid z^{(j)}, y).$$

In the above, we must either be able to calculate $p(\theta \mid z, y)$ for any augmented data (z, y) or we must be able to sample numerically from this distribution. This is a prerequisite for the data augmentation scheme, and we will assume that it is true for the remainder of this discussion. Now consider step (a), that is, the generation of the latent data from $p(z \mid y)$. Given that the current approximation to $p(\theta \mid y)$ is $g_i(\theta)$, (2.2) then suggests that z can be generated from the current predictive distribution in two steps:

(a1) generate θ from $g_i(\theta)$.
(a2) generate z from $p(z \mid \phi, y)$, where ϕ is the value obtained in (a1).

Clearly, when m is large, the two steps (a) and (b), where (a) may be implemented by (a1) and (a2), will provide a close approximation to one iteration of (2.5). Furthermore, as we will see in Sections 6 and 7, even when m is as small as 1, the iteration is still "in the right direction" in the sense that the average of $p(\theta \mid x)$ over the augmented data patterns generated across iterations will converge to the $p(\theta \mid y)$. It is noted that m need not be held fixed from iteration to iteration, and in Section 7 comments on how m should be adaptively varied are presented.

Step (a) requires the generation of multiple values of the latent data z by sampling from the conditional density of z given y. This process is termed *multiple imputation* by Rubin (1980), who first introduced it as a method for handling nonresponse in sample surveys and in censuses. Thus step (a) can be referred to as the "imputation" step. Step (b) requires the computation (or sampling) of the posterior distribution of θ based on the augmented data sets. We will call this step the "posterior" step. The algorithm consists of iterating between the imputation and posterior steps.

The usefulness of the algorithm depends to a large extent on the ease of implementation of the imputation and posterior steps. In general, neither step is guaranteed to be easy. There is a parallel limitation on the EM algorithm; namely, that in general both the E and M steps may be difficult to implement. There remains, however, a rich class of problems, especially those connected with exponential families, for which there are natural ways to carry out these steps. This is illustrated by the examples here and the examples in Dempster et al. (1977).

Linkage Example

To illustrate the basic algorithm, we consider an example that was presented in Rao (1973) and reexamined in Dempster et al. (1977) and Louis (1982). In particular, from a genetic linkage model, it is believed that 197 animals are distributed multinomially into four categories,

$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$, with cell probabilities specified by

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{(1-\theta)}{4}, \frac{(1-\theta)}{4}, \frac{\theta}{4}\right).$$

To illustrate the algorithm, y is augmented by splitting the first cell into two cells, one of which having cell probability $\frac{1}{2}$, the other having cell probability $\theta/4$. Thus the augmented data set is given by $x = (x_1, x_2, x_3, x_4, x_5)$, where $x_1 + x_2 = 125$, $x_3 = y_2$, $x_4 = y_3$, and $x_5 = y_4$. The likelihood is of the form

$$p(y \mid \theta) \propto (2 + \theta)^{y_1}(1 - \theta)^{y_2 + y_3}\theta^{y_4},$$

and the augmented likelihood is of the form

$$p(x \mid \theta) \propto \theta^{x_2 + x_5}(1 - \theta)^{x_3 + x_4}.$$

Thus, in this example, the augmented likelihood has a very simple form.

The implementation of our algorithm is then given as follows:

(a) *I Step* (*Imputation Step*).

(a1) Draw θ from the current estimate of $p(\theta \mid y)$.

(a2) Generate x_2 by drawing from the binomial distribution with parameters $(125, \theta/(\theta + 2))$.

Repeat steps (a1) and (a2) m times.

(b) *P Step* (*Posterior Step*). Set the posterior density of θ equal to the mixture of beta distributions, mixed over the m imputed values of x_2; that is,

$$p(\theta \mid y) = \frac{1}{m}\sum_{i=1}^{m} Be(v_1^{(i)}, v_2^{(i)})(\theta),$$

where

$$v_1^{(i)} = x_2^{(i)} + x_5 + 1, \quad v_2^{(i)} = x_3 + x_4 + 1,$$

and

$$Be(v_1, v_2)(\theta) = \frac{\Gamma(v_1 + v_2)}{\Gamma(v_1)\,\Gamma(v_2)}\theta^{v_1 - 1}(1 - \theta)^{v_2 - 1}.$$

In this step, the prior for θ is assumed to be uniform in $(0, 1)$.

Figure 1 presents the posterior density estimates of θ for this example. In particular, the normal approximation with $\hat{\mu} = .63$ and $\hat{\sigma} = .05$ (solid line) is plotted along with the true posterior distribution (dotted line)

$$p(\theta \mid y) \propto (2 + \theta)^{y_1}(1 - \theta)^{y_2 + y_3}\theta^{y_4},$$

and the estimated posterior (dashed line) obtained by plotting the mixture of the beta distributions at the final iteration in which $m = 1{,}600$. In the density scale, all three estimates are congruent. In the log-scale, however, even in this large sample situation a departure of the true posterior from the quadratic approximation at the mode is evident (Fig. 2).

Alternatively, we consider a second version of the data in which the sample size is reduced by a factor of 10, though the cell proportions are approximately unchanged; that is, $y = (13, 2, 2, 3)$. The resulting posterior density estimates are plotted in Figure 3. In this case, although the true posterior density and the estimated posterior density are congruent, the validity of the normal approximation may be in doubt, even when viewed on the density scale. An even more dramatic illustration is given in Figure 4, where $y = (14, 0, 1, 5)$. In cases with such a dramatic departure from normality, one or two iterations of our

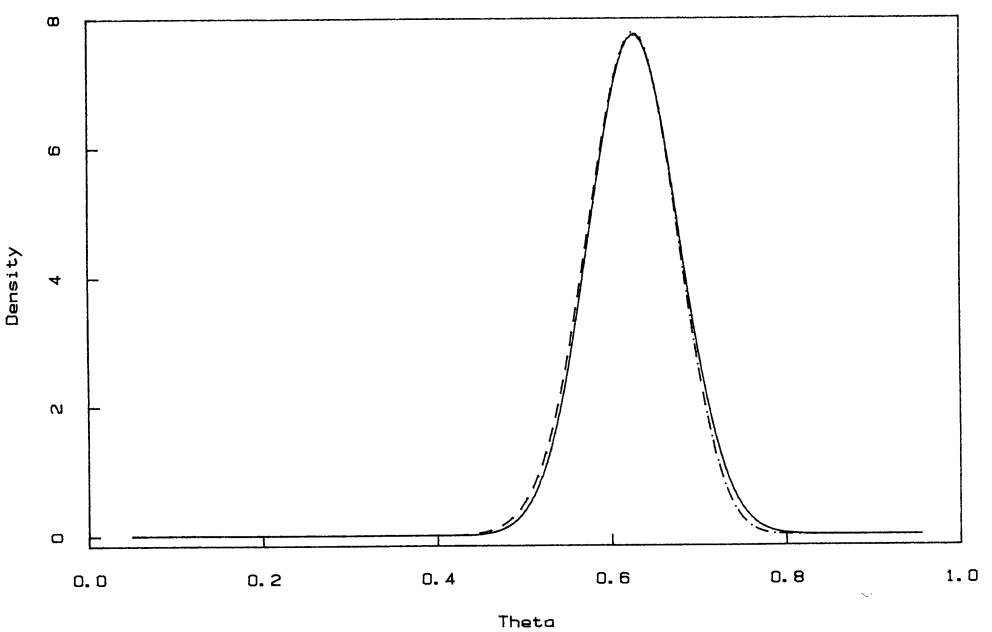

Figure 1. Posterior Density of θ for Data (125, 18, 20, 34). The solid, dashed, and dotted lines represent the normal approximation, the estimated posterior distribution, and the true posterior, respectively. The dashed and dotted lines are superimposed.

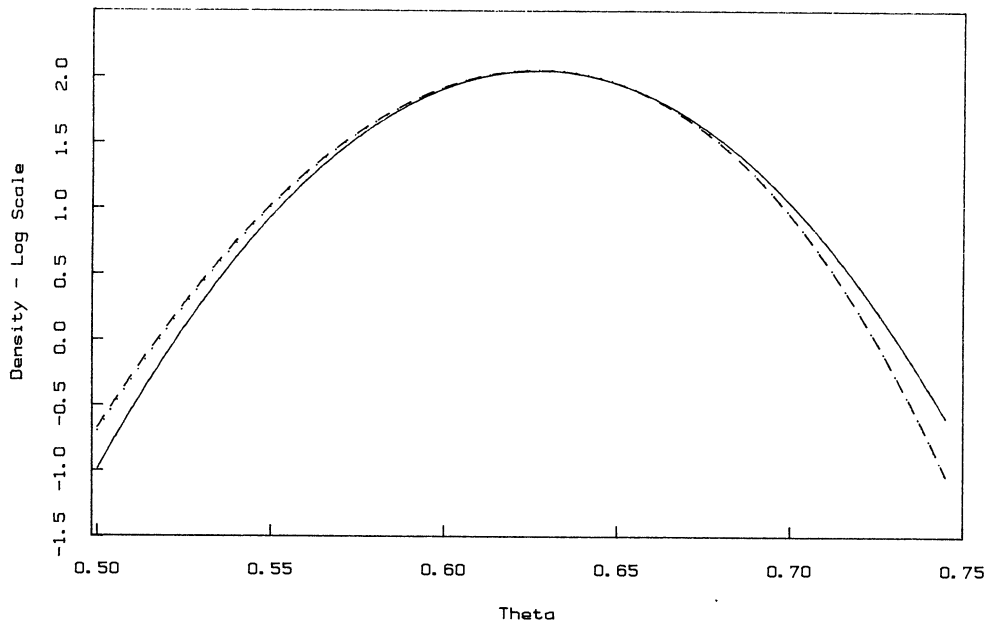

Figure 2. Log-Posterior Density of θ (same data and legend as in Fig. 1).

algorithm would indicate the inadequacy of the normal approximation.

3. FUNCTIONALS OF THE MULTIVARIATE NORMAL COVARIANCE MATRIX

In this section, the posterior distribution of the correlation coefficient from the bivariate normal distribution will be investigated. To illustrate, suppose that the data in Table 1 (Murray 1977) represent 12 observations from the bivariate normal distribution with $\mu_1 = \mu_2 = 0$, correlation coefficient ρ, and variances σ_1^2 and σ_2^2. Before proceeding to the formal analysis, we note that in the four pairs of observations, two pairs have correlation 1 and the remaining two pairs have correlation -1. Thus we can expect a nonunimodal posterior distribution for ρ in this data set. In such a case, the maximum likelihood estimate and the associated standard error will clearly be misleading. Furthermore, we point out that the information regarding σ_1^2 and σ_2^2 in the eight incomplete observations cannot be ignored because information regarding σ_1^2 and σ_2^2 is of use in making inference regarding ρ.

The implementation of the algorithm in this problem is straightforward. Given the covariance matrix Σ, the unobserved data is generated as follows:

1. If x_1 is known, then generate the unobserved observation from

$$N\left(\rho \frac{\sigma_2}{\sigma_1} x_1, \sigma_2^2(1 - \rho^2)\right).$$

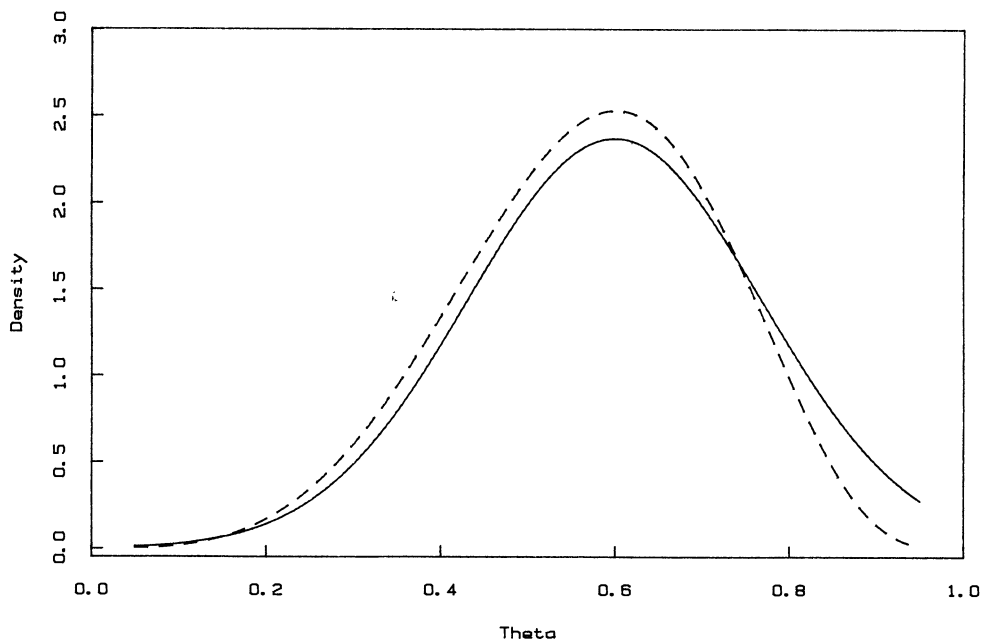

Figure 3. Posterior Density of θ for Data (13, 2, 2, 3) (same legend as in Fig. 1). The dashed lines and dotted lines are superimposed.

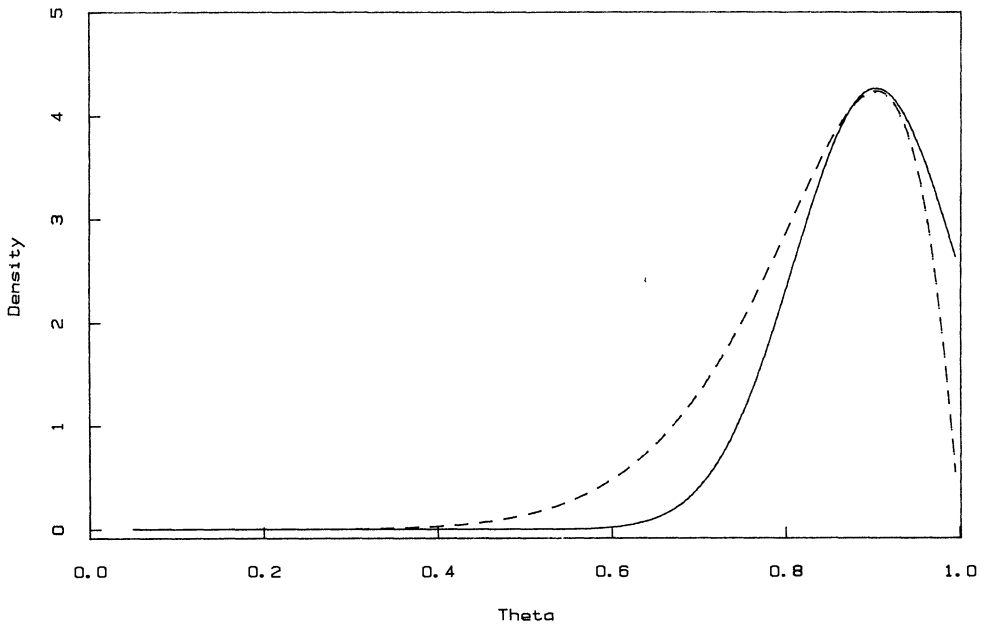

Figure 4. Posterior Density of θ for Data (14, 0, 1, 5) (same legend as in Fig. 1). The dashed and dotted lines are superimposed.

2. If x_2 is known, then generate the unobserved observation from

$$N\left(\rho \frac{\sigma_1}{\sigma_2} x_2, \sigma_1^2(1 - \rho^2)\right).$$

The covariance matrix Σ is then generated from the current guess of the posterior distribution $p(\Sigma \mid y)$. At the first iteration, ρ can be generated from $U[-1, 1]$ and σ_1^2 and σ_2^2 can be generated from weighted χ_7^2 distributions. At succeeding iterations, the updated posterior $p(\Sigma \mid y)$ is a mixture of inverted Wishart distributions. This last point follows from the fact that $p(\Sigma \mid x)$ is an inverted Wishart distribution (Box and Tiao 1973, p. 428) when the prior of Σ is given as

$$p(\Sigma) \propto |\Sigma|^{-(p+1)/2},$$

where p is the dimension of the multivariate normal distribution. Thus, in the second step of the algorithm, we generate m observations from this mixture of inverted Wishart distributions and compute the associated correlation coefficient for each observation.

Regarding the implementation of the algorithm, it is noted that the algorithm of Odell and Feiveson (1966) can be used to generate observations from the inverted Wishart distribution. The amount of computation in this algorithm is not extensive, since the computation is of order $p(p + 1)/2$, which does not depend on the sample size.

In Figure 5, we plot the histogram of the imputed correlation coefficients based on pooling the tenth through fifteenth iterations (m = 6,400). In addition, the true posterior of the correlation coefficient, which is proportional to $[(1 - \rho^2)^{4.5}]/[(1.25 - \rho^2)^8]$, is also plotted. As is evident from the plot, the estimated posterior distribution recovers the bimodal nature of the true distribution.

Finally, it is noted that the algorithm presented in this article can be used to examine the posterior distribution of any functional of the covariance matrix. For example, the posterior distribution of the largest eigenvalue of the covariance matrix (Tiao and Fienberg 1969) may be examined by simply computing the largest eigenvalue of each of the observations from the inverted Wishart distribution computed in the second step of the algorithm.

Table 1. Twelve Observations From a Bivariate Normal Distribution

1	1	−1	−1	2	2	−2	−2	*	*	*	*
1	−1	1	−1	*	*	*	*	2	2	−2	−2

* Value not observed (missing at random).

4. THE DIRICHLET SAMPLING PROCESS

In the linkage example of Section 2, the augmented posterior distribution $p(\theta \mid x)$ is a beta distribution. Thus it is a trivial matter to carry out the P step. In more complicated models, the sampling of θ from $p(\theta \mid x)$ may not be so simple. We now present a primitive but generally applicable procedure, based on a Dirichlet sampling process, which can be used to approximately sample from the posterior distribution of parametric models for multinomial data. In this section, we develop and illustrate the procedure using the linkage example. Further uses will be illustrated in Section 5.

In the linkage example, conditional on the augmented data, the distribution of the last four cell probabilities (P_2, P_3, P_4, P_5) is equal in distribution to that of ($v_2/2$, $v_3/2$, $v_4/2$, $v_5/2$), where (v_2, v_3, v_4, v_5) has the Dirichlet distribution

$$\frac{\Gamma(x_2 + x_3 + x_4 + x_5 + 4)}{\Gamma(x_2 + 1)\Gamma(x_3 + 1)\Gamma(x_4 + 1)\Gamma(x_5 + 1)} v_2^{x_2} v_3^{x_3} v_4^{x_4} v_5^{x_5},$$

$$\sum_{i=2}^{5} v_i = 1, \quad (4.1)$$

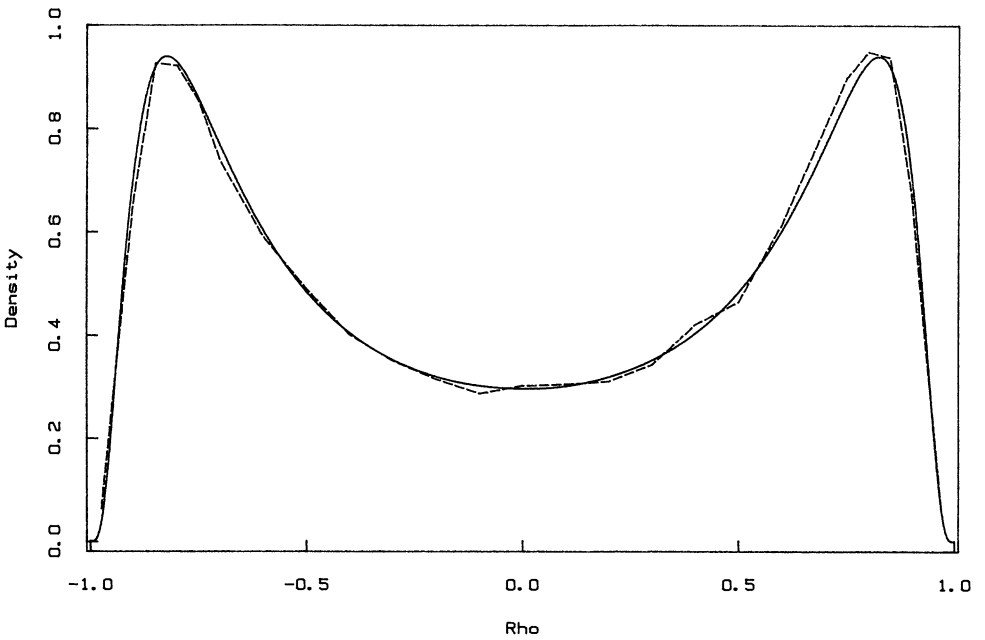

Figure 5. Posterior Density of the Correlation Coefficient. The solid and dashed lines represent the true and estimated posterior, respectively.

which will be denoted by $D(x_2, x_3, x_4; x_5)$. It is a trivial matter to generate observations from such a Dirichlet distribution. Our model, however, is not a saturated multinomial model. In fact, the linkage model specifies that (P_2, P_3, P_4, P_5) must lie on a linear parametric curve,

$$C = \left\{\left(\frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{\theta}{4}\right): \theta \in [0, 1]\right\}.$$

The posterior distribution $p(\theta \mid x)$ will only induce a distribution of (P_2, P_3, P_4, P_5) on the curve C. How is this induced distribution related to the Dirichlet distribution (4.1)? The answer is simple:

Lemma. The distribution induced by $p(\theta \mid x)$ on the curve C is the same as the conditional distribution induced by the Dirichlet distribution (4.1) on C (through the relationship $\mathbf{P} = \frac{1}{2}\boldsymbol{\nu}$).

Proof. To verify the lemma, it is sufficient to check that the ratio of the densities evaluated at any two points on C is identical under either distribution.

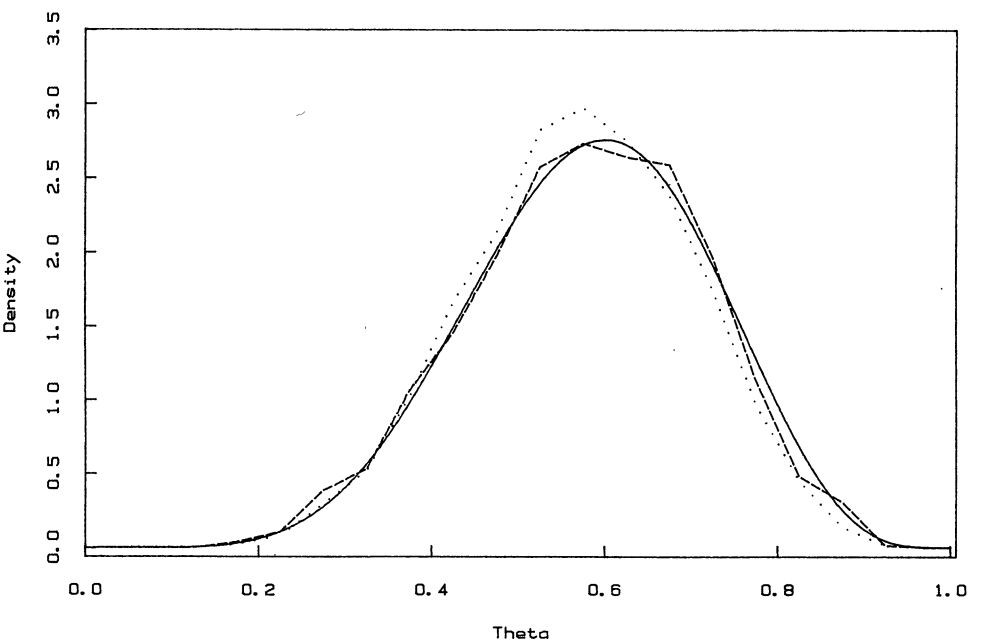

Figure 6. Posterior Density of θ for Data (3, 2, 2, 3). The dotted, dashed, and solid lines represent the estimate based on 10,000 values, the estimate based on 3,000 values, and the true posterior distribution, respectively.

This lemma suggests a simple two-stage algorithm: (a) generate observations from the Dirichlet distribution (4.1) and (b) accept only those points lying relatively close to the parametric curve C.

To be specific, observations are drawn from $D(x_2, x_3, x_4; x_5)$ and for each of these observations, we find the $\hat{\theta}$ that gives cell probabilities $(\hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)$ closest to the observed Dirichlet observation (p_2, p_3, p_4, p_5). Given the functional dependencies of each of the probabilities on θ: $P_2 = \theta/4$, $P_3 = 1/4 - \theta/4$, $P_4 = 1/4 - \theta/4$, and $P_5 = \theta/4$, the least squares solution yields $\hat{\theta} = 2(p_2 + p_5)$. The approximate posterior distribution for θ is then obtained by forming the histogram of those $\hat{\theta}$ values whose corresponding $(\hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)$ vector is within an ε-neighborhood of (p_2, p_3, p_4, p_5), that is, such that

$$\left(\sum_{i=2}^{5} (p_i - \hat{p}_i)^2\right)^{1/2} < \varepsilon.$$

According to the above lemma, if ε is sufficiently small, then the $\hat{\theta}$ values obtained in this way will have a distribution approximately equal to $p(\theta \mid x)$.

In practice, the value of ε is selected by plotting a sequence of estimated posterior distributions of θ corresponding to a sequence of decreasing ε values. The curves tend to converge as the value of ε is decreased. The aforementioned procedure is generally applicable to parametric models for multinomial data if the cell probabilities are linear in θ *or* if the posterior distribution is relatively concentrated in comparison with the curvature of the parametric surface. Otherwise, the raw histogram of $\hat{\theta}$ must be multiplied by some adjustment factor.

To test the procedure in the linkage example, assume that the augmented data vector is given by (3, 2, 2, 3). To obtain the posterior distribution of θ, we begin by drawing 10,000 observations from the Dirichlet distribution corresponding to this data vector. For each of these Dirichlet observations, the value of θ that gives the closest $(\hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)$ vector is found using least squares. The resulting histograms of the $\hat{\theta}$ values (using 10,000 initial values and 3,000 accepted values) and the true posterior distribution are presented in Figure 6. An examination of this figure reveals that the estimated distribution of θ based on the restricted set of $\hat{\theta}$ values is quite similar to the true distribution.

5. THE TRADITIONAL LATENT-CLASS MODEL

The data in Table 2 represent the responses of 3,181 participants in the 1972, 1973, and 1974 General Social Surveys, as presented in Haberman (1979). The participants in these surveys are cross-classified by the year of the survey and their responses to each of three questions regarding abortion. Thus the cell entry n_{abcd} represents the number of subjects who in year $D = d$ give responses a to question A, b to question B, and c to question C. Regarding question A, subjects are asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion *if she is married and does not want any more children.*" In question B, the italicized phrase is replaced with "if the family has a very low income and cannot afford any more children," and in question C it is replaced with "if she is not married and does not want to marry the man." For these data, Haberman (1979) considered several models, one of which is the traditional latent-class model. [See Goodman (1974a,b), Haberman (1979), or Clogg (1977) for an exposition of this model.] In this example, the traditional latent-class model assumes that the manifest variables (A, B, C, D) are conditionally independent, given a dichotomous latent variable (X). In other words, if the value of the dichotomous latent variable is known for a given participant, then knowledge of the response to a given question provides no further information regarding the responses to either of the other two questions. Haberman used the EM and scoring algorithms to obtain maximum likelihood estimates of the cell probabilities.

Table 2. White Christian Subjects in the 1972–1974 General Social Surveys, Cross-Classified by Year of Survey and Responses to Three Questions on Abortion Attitudes

Year (D)	Response to A	Response to B	Response to C	Observed count
1972	Yes	Yes	Yes	334
	Yes	Yes	No	34
	Yes	No	Yes	12
	Yes	No	No	15
	No	Yes	Yes	53
	No	Yes	No	63
	No	No	Yes	43
	No	No	No	501
1973	Yes	Yes	Yes	428
	Yes	Yes	No	29
	Yes	No	Yes	13
	Yes	No	No	17
	No	Yes	Yes	42
	No	Yes	No	53
	No	No	Yes	31
	No	No	No	453
1974	Yes	Yes	Yes	413
	Yes	Yes	No	29
	Yes	No	Yes	16
	Yes	No	No	18
	No	Yes	Yes	60
	No	Yes	No	57
	No	No	Yes	37
	No	No	No	430

Source: Haberman (1979, p. 559).

One parameter of interest associated with this model is the conditional probability of a response a to question A, given that $X = 1$ (which will be denoted as π_{a1}^{AX}). In conjunction with π_{a2}^{AX}, the magnitude of this conditional probability indicates the accuracy of the response a to question A in identifying the latent classification $X = 1$, since the ratio $\pi_{a1}^{AX}/\pi_{a2}^{AX}$ is the likelihood ratio for identifying X based on an observation of A. In the present example, Haberman estimated π_{11}^{AX} to be .892. The estimated standard error can also be obtained using the delta method, though Haberman did not include this value in his presentation.

To obtain the posterior distribution of π_{11}^{AX}, the IP al-

gorithm is implemented as follows. In the initial iteration, the odds of being in the latent class $X = 2$ (which will be denoted as $\theta_{abcd.}$) is taken to be $\frac{1}{2}$ for all values of a, b, c, and d. The unobserved cell counts (n_{abcdx}) are imputed by noticing that conditional on both $\theta_{abcd.}$ and the observed cell counts n_{abcd}, the posterior distribution of n_{abcd1} follows a binomial distribution with parameters n_{abcd} and $1/(1 + \theta_{abcd.})$. The posterior distribution of π_{11}^{AX} is then obtained by drawing from the mixture of augmented posterior distributions. In particular, for a given augmented data set, a vector of probabilities $\{P_{abcdx}\}$ is drawn from the Dirichlet distribution $D(n_{11111}, \ldots, n_{22231}; n_{22232})$ and some of the observations are discarded using the Euclidean distance criterion, as discussed in the previous section. The odds of being in the latent class $X = 2$ given that $A = a$, $B = b$, $C = c$, and $D = d$ is updated using the maximum likelihood estimate (under the conditional independence model)

$$\left(\frac{\sum_{b,c,d} p_{abcd2}}{\sum_{b,c,d} p_{abcd1}}\right)\left(\frac{\sum_{a,c,d} p_{abcd2}}{\sum_{a,c,d} p_{abcd1}}\right)\left(\frac{\sum_{a,b,d} p_{abcd2}}{\sum_{a,b,d} p_{abcd1}}\right)\left(\frac{\sum_{a,b,c} p_{abcd2}}{\sum_{a,b,c} p_{abcd1}}\right)\cdot\left(\frac{\sum_{a,b,c,d} p_{abcd1}}{\sum_{a,b,c,d} p_{abcd2}}\right)^3$$

and the algorithm cycles until convergence is achieved. For each augmented data set, the conditional probability of interest is calculated from the equation

$$\pi_{11}^{AX} = \frac{\sum_{b,c,d} p_{1bcd1}}{\sum_{a,b,c,d} p_{abcd1}}.$$

In Figures 7a and 7b, the estimated posterior distribution of π_{11}^{AX} is presented, where the values from the fifteenth through the twentieth iteration are pooled (m = 1,600) to form the histogram in these figures. As can be seen from the figures, the posterior distribution appears to be bimodal, with one mode occurring at about .039 and the other mode occurring at about .886. The reason for this bimodality stems from the unidentifiability inherent in the problem. In the latent-class model, the data analyst has the choice of identifying a positive attitude toward abortion with the condition that $X = 1$ or with the condition that $X = 2$. The mode occurring at .039 occurs if one identifies a positive attitude with $X = 2$; the second mode occurs if a positive attitude is identified with $X = 1$. In this regard, it is important to note that the modes are well separated. Thus, for the present data set, the conditional probability is, in the Bayesian sense, *locally identifiable*.

Conditioning on the identification of a positive attitude toward abortion with $X = 1$, that is, examining the right mode, we find that our point estimate for π_{11}^{AX} is close to the maximum likelihood estimate (.886 versus .892). (Such an identification is reasonable given the nature of the question.) In addition, there is little evidence of a departure of the normal approximation from the posterior distribution. Comparing the estimated density to the normal curve with matching mean and standard error (.009), an overall concordance is observed (Fig. 7b). A similar conclusion is reached by examining the corresponding rankit plot (Fig. 8). Regarding the lower mode (Figs. 7a and 9), some evidence against the normal approximation ($\hat{\mu}$ = .039, $\hat{\sigma}$ = .006) is noted. In particular, the posterior distribution is slightly skewed to the right.

6. THEORETICAL DEVELOPMENT

In this section, we return to the study of the algorithm motivated and outlined in Section 2. In previous examples, it was seen that the algorithm converged to the true posterior. The results in this section will explain *why* the al-

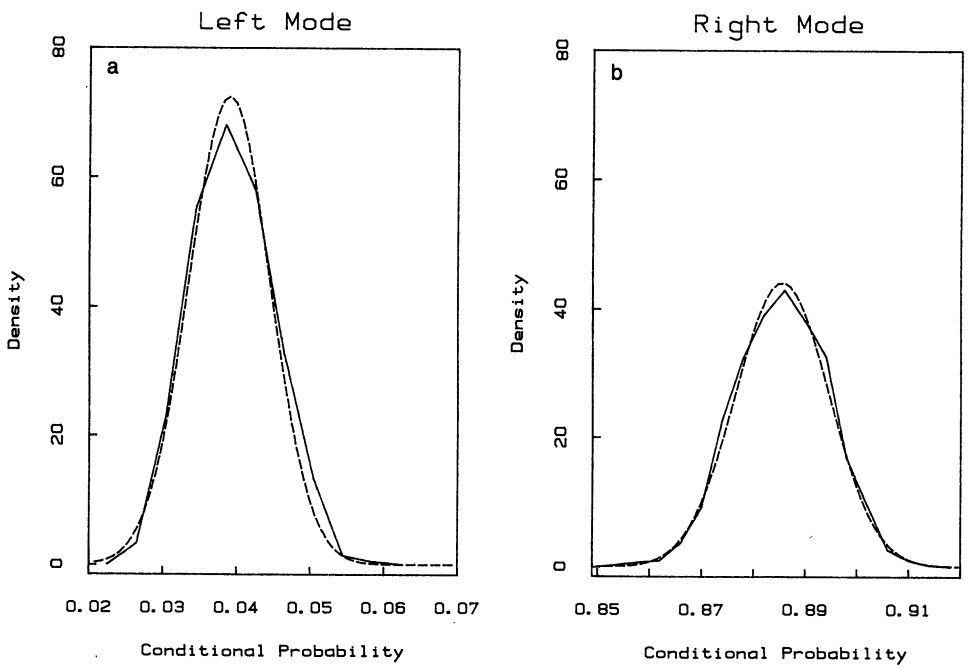

Figure 7. Posterior Density of π_{11}^{AX}. The solid and dashed lines represent the estimated and true posterior density, respectively. (a) Left mode. (b) Right mode.

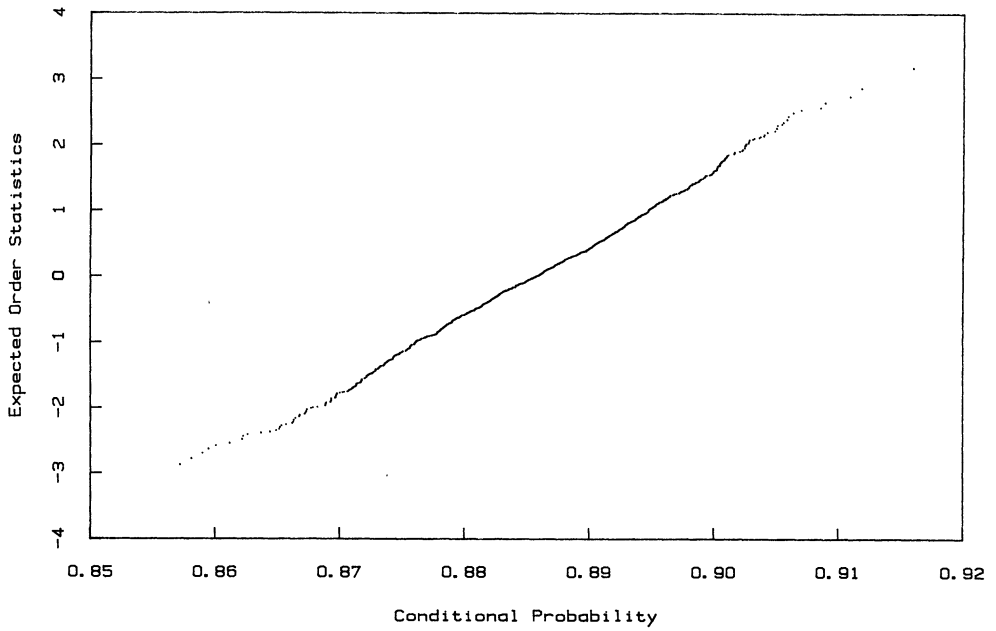

Figure 8. Rankit Plot for Right Mode.

gorithm should converge and at what rate it does so. For simplicity we will first assume that Θ is a connected subset of R^p. The theory is essentially the same for discrete Θ, as discussed briefly at the end of this section. Let L_1 be the space of (Lebesque) integrable functions of $\theta \in \Theta$, and $\|f\| = \int |f(\theta)|\, d\theta$ for $f \in L_1$. Let $g_i(\theta)$, $K(\theta, \phi)$, and T be defined as in (2.3)–(2.5). Clearly, T is a bounded linear operator on L_1. Let us denote the true posterior density by $g_*(\theta)$. Then according to (2.3), g_* is a fixed point under T; that is, $Tg_* = g_*$.

The main results of this section are, roughly, (a) g_* is the only density that satisfies the fixed point equation and (b) for essentially any starting value, the iteration (2.5) converges linearly to g_*, that is, the deviation in the L_1 norm decreases at a geometric rate. These statements hold under some regularity conditions [Condition (C), given subsequently].

The first theorem shows that the L_1 distances from the true posterior are nonincreasing in the iterations.

Theorem 1. $\|g_{i+1} - g_*\| \leq \|g_i - g_*\|$.

Proof. The proof will make use of the following elementary facts: (a) $\int K(\theta, \phi)\, d\theta = 1$; thus if $f(\theta) \geq 0$ for all θ, then $\|Tf\| = \|f\|$. (b) If $f(\theta) \geq g(\theta)$ for all θ, then $Tf(\theta) \geq Tg(\theta)$ for all θ. To prove the theorem, let $f = g_i - g_*$. Then

$$Tf = g_{i+1} - g_*,$$

$$\|Tf\| = \int |Tf(\theta)|\, d\theta \leq \int (T|f|)(\theta)\, d\theta = \|T|f|\| = \|f\|.$$

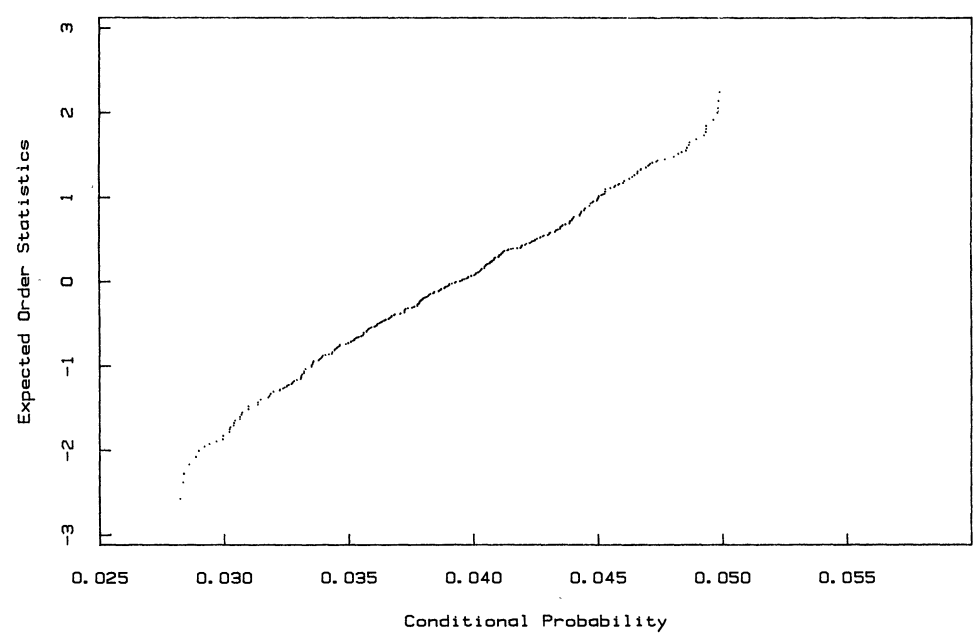

Figure 9. Rankit Plot for Left Mode.

Can the distances from the truth be strictly decreasing? Is g_* the only density that satisfies the fixed point equation? To obtain positive results, we must impose some regularity conditions.

Condition (C). $K(\theta, \phi)$ is uniformly bounded and is equicontinuous in θ. For any $\theta_0 \in \Theta$, there is an open neighborhood U of θ_0, so $K(\theta, \phi) > 0$ for all $\theta, \phi \in U$.

The second part of this condition says that if θ and ϕ are close, then it is possible to generate some latent data pattern z from $p(z \mid \phi, y)$ such that $p(\theta \mid z, y)$ is nonzero, which is a reasonable condition.

Lemma 1. Under Condition (C), any density g that is a fixed point of T must be continuous and strictly positive.

Proof. By hypothesis, $g(\theta) \geq 0$, $g(\theta) = \int K(\theta, \phi)g(\phi)\, d\phi$. Hence $|g(\theta_1) - g(\theta)| \leq \int |K(\theta_1, \phi) - K(\theta, \phi)|\, g(\phi)\, d\phi$, which tends to 0 as $\theta_1 \to \theta$, by dominated convergence. This proves continuity of g. To prove positivity, consider $A = \{\theta \in \Theta : g(\theta) > 0\}$. If $A \neq \Theta$, then there must be a $\theta_0 \in \Theta$ that is also on the boundary of A. By Condition (C), there is a neighborhood U of θ_0 such that $K(\theta, \phi) > 0$ for all $\theta, \phi \in U$. Since θ_0 is on the boundary we must have $g(\phi) > 0$ for some open subset of U. Hence $0 = g(\theta_0) \geq \int_U K(\theta_0, \phi)g(\phi)\, d\phi > 0$, a contradiction. Hence $A = \Theta$.

Lemma 2. Under Condition (C), if $f \in L_1$ is a function so that neither its positive part f^+ nor its negative part f^- are identically 0, then $\|Tf\| < \|f\|$.

Proof. By connectedness of Θ and Condition (C), we must have support of $Tf^+ \supset$ support of f^+, and support of $Tf^- \supset$ support of f^-. Note that the inclusions are strict. It follows that

$$(\text{support of } Tf^+) \cap (\text{support of } Tf^-) \tag{6.1}$$

is nonempty. Now

$$|(Tf)(\theta)| = |Tf^+(\theta) - Tf^-(\theta)|,$$
$$(T|f|)(\theta) = Tf^+(\theta) + Tf^-(\theta).$$

Hence under (6.1) we must have

$$\int |(Tf)(\theta)|\, d\theta < \int (T|f|)(\theta)\, d\theta.$$

Corollary. Under (C), the distance of g_i to g_* is strictly decreasing.

Now we are ready to state and prove the main theorems. Theorem 2 guarantees the uniqueness of the solution to the fixed point equation. Theorem 3 gives the rate of convergence of the iteration (2.5) in terms of L_1 distances.

Theorem 2. Under Condition (C), the posterior density g_* is the only density that satisfies $Tg = g$.

Proof. The fact that g_* satisfies the fixed point equation was derived in Section 2. Suppose that g_{**} is a different density satisfying $Tg = g$. Let $f = g_* - g_{**}$, then f must be continuous by Lemma 1. In addition, since $\int f(\theta)\, d\theta = 0$ and $f \neq 0$, neither f^+ nor f^- can be identically 0. Hence, by Lemma 2, $\|Tf\| < \|f\|$. But on the other hand, $Tf = Tg_* - Tg_{**} = g_* - g_{**} = f$, a contradiction.

Theorem 3. Suppose that Condition (C) holds and that the starting value g_0 satisfies $\sup_\theta (g_0(\theta)/g_*(\theta)) < \infty$. Then there exists a constant α $(0 < \alpha < 1)$, such that

$$\|g_{i+1} - g_*\| \leq \alpha^i \|g_0 - g_*\|.$$

Proof. The proof proceeds in five steps:

(a) For any $M > 0$, if $(g_0(\theta)/g_*(\theta)) < M$ for all $\theta \in \Theta$, then $(g_i(\theta)/g_*(\theta)) < M$ for all i, for all $\theta \in \Theta$.

(b) For any $M > 0$, the set $\{f \in L_1 : |f(\theta)/g_*(\theta)| < M$ for all $\theta\}$ is weakly sequentially compact in L_1.

(c) Let $f_i = g_i - g_*$ and let $\alpha = \sup_{i>1} (\|Tf_i\|/\|f_i\|)$. There exists a subsequence $\{f_{i'}\}$ such that $\|Tf_{i'}\|/\|f_{i'}\| \to \alpha$, and $f_{i'}$ converges to some f_* weakly in L_1.

(d) Since the set $\{f_{i'}\}$ is bounded and equicontinuous, we must actually have $f_{i'}$ converges to f_* strongly in L_1, and f_* can be chosen to be continuous.

(e) Hence $\alpha = \lim(\|Tf_{i'}\|/\|f_{i'}\|) = \|Tf_*\|/\|f_*\|$. But $\int f_*(\theta)\, d\theta = 0$; hence by Lemma 2, $0 \leq \alpha < 1$. From this, the theorem follows directly.

It remains to establish statements (a)–(e). Statement (e) needs no proof, statement (a) follows from elementary manipulation, and statement (b) is a well-known property of L_1 spaces (see, e.g., Dunford and Schwartz 1958, p. 294). To prove (c), let $\{f_{i'}\}$ be a subsequence of $\{f_i\}$ such that $\|Tf_{i'}\|/\|f_{i'}\| \to \alpha$. Now by (a) and (b), $\{f_{i'}\}$ is weakly sequentially compact, so there must exist a further subsequence $\{f_{i''}\}$ of $\{f_{i'}\}$ convergent weakly in L_1. This establishes (c). Finally, (d) can be established by standard analytical arguments.

Remark 1. One of the conditions of Theorem 3 requires that $g_0(\theta)/g_*(\theta)$ be uniformly bounded. For a compact parameter space Θ, this condition is automatic if Condition (C) holds, since under (C), g_* is continuous and strictly positive. For an unbounded parameter space, we need to make sure that the decay of $g_0(\theta)$ when $|\theta| \to \infty$ is not slower than that of $g_*(\theta)$. This suggests using g_0 of bounded support.

Remark 2. Theorem 3 says that the convergence rate is linear. Unfortunately, the rate α is dependent on the initial value g_0. If Θ is compact, it can be shown that the supremum of α over all possible g_0 is still less than 1; that is, we get a linear rate independent of the starting values. If Θ is unbounded, however, α can be arbitrarily close to 1, depending on the starting value. This seems to be an intrinsic limit imposed by an unbounded parameter space and should not be regarded as a weakness of the method.

Remark 3. The whole theory can be developed in the same way for finite or countable Θ. The simplest replacement for Condition (C) is to require $K(\theta, \phi) > 0$ for all $\theta, \phi \in \Theta$. Weaker conditions exist but they are cumbersome to state.

Remark 4. It is clear from properties (a) and (b) in the proof of Theorem 1 that T is a Markov transition operator. However, a search through standard references, including Doob (1953), does not produce results directly suitable

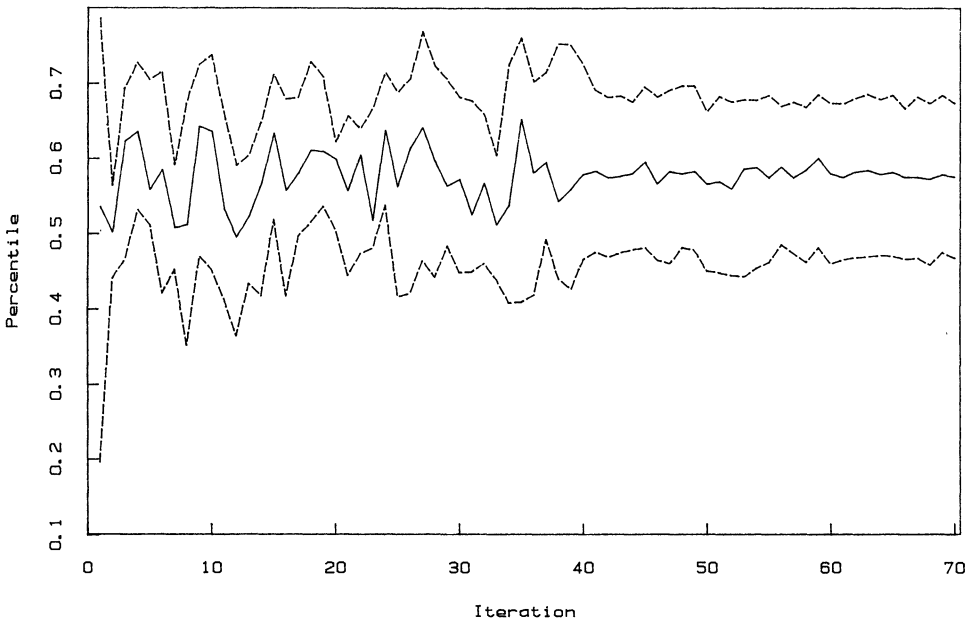

Figure 10. Median and Upper and Lower Quartiles of θ Values Across Iterations. The upper dashed line, the solid line, and the lower dashed line represent the upper quartile, the median, and the lower quartile, respectively.

for our use. Especially, the L_1 convergence rate in Theorem 3 seems to be new.

Remark 5. Similarly, there is a vast literature on fixed point operator equations and the method of successive substitution (see, e.g., Rall 1969, pp. 64–74). Again, we have not found results directly usable here.

7. PRACTICAL IMPLEMENTATION OF THE ALGORITHM

As indicated in the introduction, if the sample size m is taken to be large in each iteration, then the algorithm can be interpreted as the method of successive substitution for solving a fixed point problem. In practice, however, it is inefficient to take m large during the first few iterations when the estimated posterior distribution is far from the true distribution. Rather, it is suggested that m initially be small and then increased with successive iterations. In addition, we have found it helpful to monitor the progress of the algorithm by examining selected percentiles of the estimated posterior distribution, for example, the 25%, 50%, and 75% percentiles.

To illustrate these ideas, let us return to the linkage

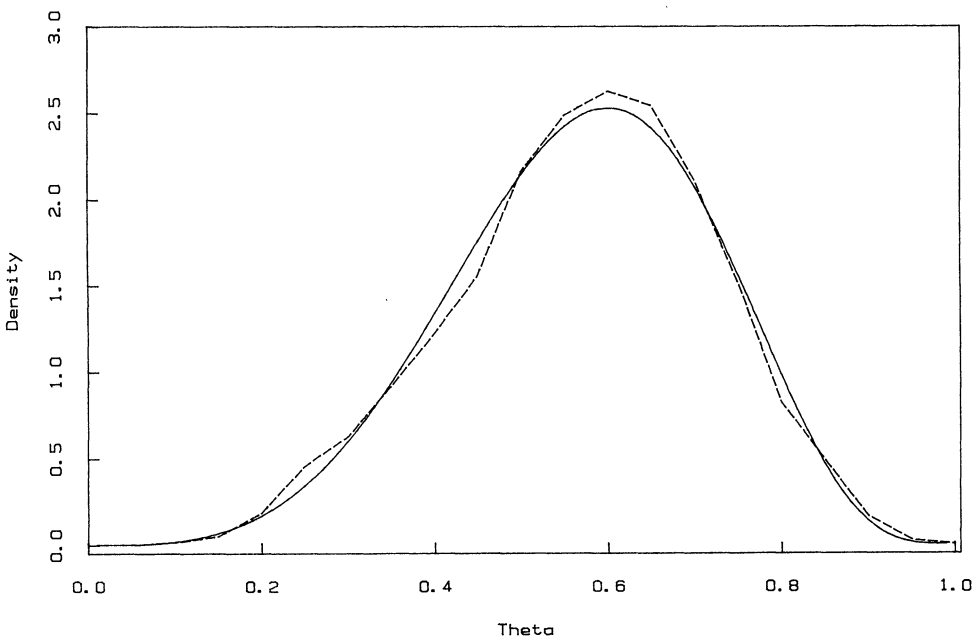

Figure 11. The Posterior Density of θ. The dashed and solid lines represent the estimated and true posterior, respectively.

example, where the observed data is taken to be (13, 2, 2, 3). At the initial iteration, m is taken to be 20. The algorithm then runs through 40 iterations, at which point it appears (see Fig. 10) that the process has become stationary. The sample size is then increased to 400 and the algorithm proceeds through 20 further iterations. From Figure 10, we see that the effect of increasing m has been to reduce substantially the system variability. The final 10 iterations are run with $m = 1{,}600$, and the estimated posterior distribution is then obtained by pooling the imputed theta values from the final iterations. Figure 11 is obtained by pooling the results of iterations 67–70.

For obvious reasons, the statistical fluctuations exhibited in iterations 20–40 cannot be reduced by further iterations without increasing the sample size m (for the sample of augmented data). Typically, graphical displays, such as Figure 10, will give a good idea of how m should be varied. A more formal procedure can be obtained by comparing the within-iteration variance to the between-iteration variance.

Another point illustrated in the linkage example is the possibility of pooling among iterations. For example, in iterations 20–40 we see that the process has stabilized. These samples are then pooled to form a combined sample of 400 to initialize the new iteration with $m = 400$. This pooled sample should not be regarded for all purposes as a random sample because the values from different iterations are dependent. If the process has reached equilibrium, however, then the histogram constructed from the sample will give the correct shape. Thus, for example, let m, $\bar{\theta}_m$, and s denote, respectively, the sample size, mean, and standard deviation of the pooled sample. It then follows that $\bar{\theta}_m$ will be a consistent estimate (as $m \to \infty$) of the posterior mean of θ, but the standard error of this estimate will typically be larger than $s/\sqrt{m}$. To see this, consider the extreme case in which $m = 1$, so that iteration i produces only one value $\theta(i)$. In this case, $\theta(i)$ $(i = 1, 2, \ldots)$ forms a Markov process with transition function equal to $K(\theta, \phi)$, as defined in (2.3). Under the regularity conditions of Section 6, this is an ergodic Markov process with an equilibrium distribution satisfying the fixed point equation given in (2.3). Hence $\bar{\theta}_m$ will converge to the mean of this equilibrium distribution, which is identical to the mean of the posterior distribution.

Finally, it is noted that the computation in Section 2.1 (10 iterations with $m = 1{,}600$) required 13 minutes on a VAX 750, whereas the computations in Section 3 (15 iterations with $m = 6{,}400$) and Section 5 (15 iterations with $m = 1{,}600$) required 23 minutes and 171 minutes, respectively, on a VAX 750.

[*Received April 1985. Revised October 1986.*]

REFERENCES

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Clogg, C. C. (1977), "Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users," Working Paper 1977-09, Pennsylvania State University, Population Issues Research Center.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Doob, J. L. (1953), *Stochastic Processes*, New York: John Wiley.

Dunford, N., and Schwartz, J. T. (1958), *Linear Operators. Part I: General Theory*, New York: Interscience Publishers.

Goodman, L. A. (1974a), "The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable. Part I—A Modified Latent Structure Approach," *American Journal of Sociology*, 79, 1179–1259.

——— (1974b), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231.

Haberman, S. J. (1979), *Analysis of Qualitative Data, Volume 2: New Developments*, New York: Academic Press.

Li, K. H. (1985a), "Hypothesis Testing in Multiple Imputation With Emphasis on Mixed-Up Frequencies in Contingency Tables," unpublished Ph.D. dissertation, University of Chicago, Dept. of Statistics.

——— (1985b), "Imputation Using Markov Chains," technical report, The Chinese University of Hong Kong, Dept. of Statistics.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Murray, G. D. (1977), Comment on "Maximum Likelihood From Incomplete Data Via the EM Algorithm," by A. P. Dempster, N. M. Laird, and D. B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, 39, 27–28.

Odell, P. L., and Feiveson, A. H. (1966), "A Numerical Procedure to Generate a Sample Covariance Matrix," *Journal of the American Statistical Association*, 61, 199–203.

Rall, L. B. (1969), *Computational Solution of Nonlinear Operator Equations*, New York: John Wiley.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Rubin, D. B. (1978), "Multiple Imputation in Sample Surveys. A Phenomenological Bayesian Approach to Nonresponse," in *Imputation and Editing of Faulty or Missing Survey Data*, Washington, DC: U.S. Department of Commerce, Social Security Administration.

——— (1980), "Handling Non-Response in Sample Surveys by Multiple Imputation," U.S. Department of Commerce Bureau of Census Monograph.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C., and Dransfield, M. (1985), "The Implementation of the Bayesian Paradigm," *Communications in Statistics, Part A—Theory and Methods*, 14, 1079–1102.

Tiao, G. C., and Fienberg, S. (1969), "Bayesian Estimation of Latent Roots and Vectors With Special Reference to the Bivariate Normal Distribution," *Biometrika*, 56, 97–104.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–393.

Comment

A. P. DEMPSTER*

The article by Tanner and Wong is a notable achievement because it combines theoretical elegance and insight with an important contribution to the emerging field of practical Bayesian statistics. I believe that Bayesian statistics is fundamentally a computational theory whereby the implications of a set of statistical data are understood by constructing a probability model and computing a set of relevant probabilities and expectations. There is little need for the traditional mathematical studies of properties of statistical techniques that dominate statistical journals. There is a great need for better understanding of the principles of model construction, criticism, and revision, and for studies of sensitivity of inferences to model change as a function of features of the data. Such studies depend critically on the availability of sophisticated Bayesian computing tools, especially for realistically complex data systems. The data augmentation idea seems destined to greatly facilitate the handling of complexity, so that situations that recently were computationally intractable now appear feasible. I look forward to much further development, both of related algorithms and of implementations for specific models, now that Tanner and Wong have shown us the way.

For example, hierarchical models with several levels of Gaussian randomness have a long history and many important applications. Standard practice in this area has been to replace the unknown hyperparameters (i.e., variance components) by point estimates, and then to report posterior distributions for interesting quantities, assuming that the point estimates are known true values. Standard methods for adjusting these posterior distributions to reflect increased variability due to inaccuracy of the point estimates of the hyperparameters do not exist. For a Bayesian such standard methods are necessarily dubious in principle, in the sense that second-level variability typically has relatively small sample size so that dependence on specific choices of hyperpriors is practically meaningful. The immediate task is to implement methods for constructing and computing with genuine hyperpriors, and then to study the results of applying the methods to data sets with specified features. Data augmentation methods (EM algorithms) have in recent years facilitated the implementation and study of methods based on maximum likelihood estimation of hyperparameters. I now expect parallel developments for logically more satisfying Bayesian methods.

I believe that Tanner and Wong might have helped readers unfamiliar with the problem of Bayesian computing by presenting an outline of the range of currently available methods and of their strengths and weaknesses in various settings. These include exact analytic expressions, analytic approximations and expansions, numerical integration procedures, and Monte Carlo methods, especially importance sampling. The basic idea of data augmentation exhibited in Tanner and Wong's formula (2.1) is to depend on numerical mixing via $p(z \mid y)$ of the analytically tractable complete data posteriors $p(\theta \mid z, y)$, where the relative contributions of each part to posterior variability depends on the fraction of missingness inherent in the adopted model. The numerical mixing operation could in principle draw on any of the approximate methods listed previously, and opportunities exist for developing algorithms along many different lines.

Tanner and Wong do not use any of the standard approximate methods, but instead adopt an ingenious iterative scheme based on successive substitution. The result is that the Tanner and Wong method has two striking parallels to the EM algorithm, namely, (a) dependence on the analytic simplicity of the likelihood function given augmented data, and (b) use of an iterative process with a monotone convergence property. Although the iterative scheme is mathematically very appealing, its use in the actual numerical problem may be somewhat unnatural, because Monte Carlo approximation is typically used for each successive iterate $g_i(\theta)$, whereas it seems to me that if Monte Carlo is needed it might more efficiently be applied directly to approximating $g(\theta)$.

A problem with direct use of Monte Carlo for numerical mixing is that available algorithms generally permit only sampling from approximate posteriors, whence importance sampling weights are required for the direct use of (2.1). Two desiderata for importance sampling are (a) it should be easy to compute the weights, and (b) the weights should be as constant as possible. The feasibility of (a) depends on the model. For the hierarchical Gaussian models mentioned previously, simulation of z produces the likelihood $l(\theta)$ as a byproduct so that weights depending only on θ may be used. For other models, more variable weights based on (θ, z) pairs may be necessary. There would still be a role for successive iterates $g_i(\theta)$ for successive stages of sampling, in order to stabilize the importance sampling weights, but the iterates themselves would involve weights and so would differ from the Tanner and Wong iterates. My central point is that there are many options for using Monte Carlo in conjunction with the basic data augmentation principle (2.1), whence opportunities for studying and comparing them need to be pursued.

* A. P. Dempster is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This work was facilitated in part by National Science Foundation Grant DMS-85-04332.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

Comment

Simulation in Hierarchical Models

C. N. MORRIS*

1. SIMULATION FOR GENERAL HIERARCHICAL MODELS

Hierarchical Bayes models and related empirical Bayes models are suited ideally for analysis by Tanner and Wong's data augmentation algorithm. The following amplifies on the general approach and examines the case of normal means in particular.

In the usual three-stage hierarchical setting, assume known forms of densities: $p_1(y \mid \theta)$ for observed data y (stage I in the hierarchy); $p_2(\theta \mid \alpha)$ for unknown parameters θ (stage II), given the unknown hyperparameter α; and $p_3(\alpha)$ (stage III) for α. The joint distribution is

$$p(y, \theta, \alpha) = p_1(y \mid \theta)p_2(\theta \mid \alpha)p_3(\alpha). \tag{1.1}$$

This can be rewritten in two alternative forms:

$$p(y, \theta, \alpha) = q_1(y)q_2(\theta \mid y)q_3(\alpha \mid \theta) \tag{1.2}$$

$$= r_1(y)r_2(\theta \mid \alpha, y)r_3(\alpha \mid y), \tag{1.3}$$

with the q_i and r_j densities derived from p_1, p_2, p_3. Of course $q_1(y) = r_1(y)$. Note the useful simplifying fact that the conditional probability of α given θ and y does not depend on y and hence is $q_3(\alpha \mid \theta)$.

Tanner and Wong obtain a sample $\theta^{(1)}, \ldots, \theta^{(m)}$ of size m from $q_2(\theta \mid y)$, the posterior density, or a jointly distributed sample $(\theta^{(j)}, \alpha^{(j)})$ $(j = 1, \ldots, m)$ from

$$\bar{p}(\theta, \alpha \mid y) \equiv p(y, \theta, \alpha)/q_1(y). \tag{1.4}$$

Their method accomplishes the latter objective, and hence the former, by ignoring the $\alpha^{(j)}$. A sample of size m is drawn from (1.4) by iterating two steps, the imputation and posterior steps. The imputation step (I step) imputes missing data, which in this model are taken to be values of the unknown hyperparameter α. Given an initial sample $\theta^{(1)}, \ldots, \theta^{(m)}$, impute $\alpha^{(1)}, \ldots, \alpha^{(m)}$ independently:

$$\alpha^{(j)} \sim q_3(\alpha \mid \theta^{(j)}), \qquad j = 1, \ldots, m. \tag{1.5}$$

Then, given $\alpha^{(1)}, \ldots, \alpha^{(m)}$, the posterior step (P step) simulates a new sample $\theta^{(1)}, \ldots, \theta^{(m)}$ according to

$$\theta^{(j)} \sim r_2(\theta \mid \alpha^{(j)}, y) \tag{1.6}$$

independently for $j = 1, \ldots, m$. Iterating (1.5) and (1.6) many times, beginning with any starting point, produces a final sample $(\theta^{(j)}, \alpha^{(j)})$ $(j = 1, \ldots, m)$ from the desired distribution $\bar{p}(\theta, \alpha \mid y)$, even if the initial sample is not distributed as $\bar{p}$.

In some important hierarchical problems the P step is conducted easily whenever $p_2(\theta \mid \alpha)$ is a conjugate prior distribution relative to $p_1(y \mid \theta)$, because then $r_2(\theta \mid \alpha, y)$ takes the same form as $p_2(\theta \mid \alpha)$. For example, inferences about the usual parameters θ when p_1 is a normal, Poisson, gamma, binomial, or negative binomial distribution and p_2 is conjugate to p_1, only require normal, gamma, or beta variates to sample from r_2. The I step is simplified in all hierarchical models, because $q_3(\alpha \mid \theta)$ does not depend on y, and is simplified further when $p_3(\alpha)$ is chosen as conjugate to $p_2(\theta \mid \alpha)$. Thus simple forms for q_3 and r_2 often can be chosen for the iterations (1.5)–(1.6).

2. EXAMPLE: INFERENCES ABOUT NORMALLY DISTRIBUTED OBSERVATIONS

Suppose that k population means $\theta = (\theta_1, \ldots, \theta_k)'$ are to be estimated after observing independent normally distributed sample means $y = (y_1, \ldots, y_k)'$ with

$$y_i \mid \theta_i \overset{\text{iid}}{\sim} N(\theta_i, V_i), \qquad i = 1, \ldots, k, \tag{2.1}$$

and $V_i = \text{var}(y_i \mid \theta_i)$ known. The conjugate prior distribution $p_2(\theta \mid \alpha)$ is assumed for each θ_i independently with $\alpha = A > 0$ unknown and

$$\theta_i \mid A \overset{\text{iid}}{\sim} N(0, A), \qquad i = 1, \ldots, k. \tag{2.2}$$

Further assume that $p_3(\alpha)$ has the conjugate (relative to p_2) form

$$p_3(A) = cA^{-1-q/2} \exp(-.5\lambda/A) \tag{2.3}$$

for known $q > -k/2$ and $\lambda \geq 0$. These choices lead to proper posterior densities, but proper prior densities require $q > 0$ and $\lambda > 0$, in which case A is distributed as λ/χ^2_q.

Let initial values $A^{(1)}, \ldots, A^{(m)}$ be given. The posterior distribution of θ given (y, A) is normally distributed and the P step samples

$$\theta_i^{(j)} \overset{\text{iid}}{\sim} N((1 - B_i^{(j)})y_i, V_i(1 - B_i^{(j)})) \tag{2.4}$$

$(i = 1, \ldots, k; j = 1, \ldots, m)$ independently, with $B_i^{(j)} \equiv V_i/(V_i + A^{(j)})$ $(i = 1, \ldots, k; j = 1, \ldots, m)$. Next, given values $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_k^{(j)})$ from (2.4), (2.2), and (2.3) imply a reciprocal chi-squared distribution for

* C. N. Morris is Professor, Department of Mathematics, and Director, Center for Statistical Sciences, University of Texas, Austin, TX 78712. Support for this research was provided by National Science Foundation Research Grant DMS-8407876. This comment was written while the author was a Visiting Fellow at the University of Warwick, England. The author extends his appreciation to Don Rubin for key discussions on this topic.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

A. The I step (1.5), therefore, samples new values $A^{(1)}, \ldots, A^{(m)}$ according to

$$A^{(j)} \sim \frac{\lambda + \|\theta^{(j)}\|^2}{\chi^2_{k+q}}, \quad j = 1, \ldots, m, \tag{2.5}$$

with χ^2_{k+q} sampled independently for each j, $\|\theta\|^2$ denoting sum of squares. The final iteration yields m independent samples $(\theta_1^{(j)}, \ldots, \theta_k^{(j)}, A^{(j)})$ $(j = 1, \ldots, m)$, distributed as $\bar{p}(\theta, \alpha \mid y)$.

There can be ambiguities in the use of these data. Estimation of the posterior mean $(1 - B_i)y_i$, for example, is achieved by averaging either the $(1 - B_i^{(j)})y_i$ values or the $\theta_i^{(j)}$ values for $j = 1, \ldots, m$, but these values will not agree precisely.

The preceding can be expanded straightforwardly to encompass an unknown prior mean so that $\alpha = (\mu, A)$, $\mu = E\theta_i$, or to include regression forms $E\theta_i = \beta' x_i$ with $\alpha = (\beta_1, \ldots, \beta_r, A) = (\beta', A)$ and each x_i a known vector, assuming a normal or a flat prior distribution is used for μ or β.

The simulations described here will be costly when substantial accuracy is required, because that usually necessitates large m and many iterations. The number of iterations will be reduced dramatically, however, if a good starting approximation to the distribution $r_3(\alpha \mid y)$ is available. Assuming the model (2.1)–(2.3), a suggestion follows, based on Morris (1987).

Let $l(A)$ denote the logarithm of the modified posterior density $A \cdot r_3(A \mid y)$, and let $l'(A)$, $l''(A)$ be its first two derivatives.

(a) Find $\hat{A} > 0$ a (usually unique) value satisfying $l'(\hat{A}) = 0$, with

$$2l'(A) = -(k + q)A^{-1} + \lambda A^{-2} + \sum_1^k y_i^2/(V_i + A)^2. \tag{2.6}$$

(b) Define

$$p \equiv (-2\hat{A}^2 l''(\hat{A}))^{-1/2}. \tag{2.7}$$

(c) For $j = 1, \ldots, m$, let $u_j = (j - .5)/m$, and set

$$A^{(j)} = \hat{A}(u_j/(1 - u_j))^p. \tag{2.8}$$

The values (2.8) are approximations to the expected order statistics from the posterior density $r_3(A \mid y)$.

Tanner and Wong are to be congratulated for offering this approach to Bayesian simulation. Perhaps this note helps to expand further the usefulness of their method and to emphasize that their "missing data" concept can be used to include unknown parameters or latent data. Promising as their data augmentation method is, however, much still must be learned about affordable accuracy in large problems, about criteria for stopping the iterations (1.5)–(1.6), and whether simulation times are feasible in high-dimensional problems.

ADDITIONAL REFERENCE

Morris, C. N. (1987), "Empirical Bayes Interval Estimation," Technical Report 42, The University of Texas at Austin, Center for Statistical Sciences.

Comment

A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm

DONALD B. RUBIN*

Tanner and Wong are to be congratulated for this fine addition (henceforth, TW) to modern statistical theory and practice, which makes heavy use of current computational capabilities to draw inferences using simulation techniques. I firmly believe (Rubin 1985, sec. 2.5) that given today's computing environments, obtaining inferences in applied problems using simulation is often highly desirable because of the resultant flexibility of matching models to data without tedious and tangential mathematical analysis.

I expect that TW will simulate a variety of interesting efforts involving inference via simulation. For instance, it is easy to imagine a major article or sequence of articles providing details of the data augmentation algorithm in the series of EM examples outlined in Dempster, Laird, and Rubin (1977), and explicated and extended in subsequent literature. Similarly, more technical articles on this algorithm concerning rates of convergence, choices of

* Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This comment was written while the author was visiting the University of Warwick, England, and was also partially supported by National Science Foundation Grants SES-8311428 and DMS-8504332.

A. The I step (1.5), therefore, samples new values $A^{(1)}, \ldots, A^{(m)}$ according to

$$A^{(j)} \sim \frac{\lambda + \|\theta^{(j)}\|^2}{\chi^2_{k+q}}, \quad j = 1, \ldots, m, \tag{2.5}$$

with χ^2_{k+q} sampled independently for each j, $\|\theta\|^2$ denoting sum of squares. The final iteration yields m independent samples $(\theta_1^{(j)}, \ldots, \theta_k^{(j)}, A^{(j)})$ $(j = 1, \ldots, m)$, distributed as $\bar{p}(\theta, \alpha \mid y)$.

There can be ambiguities in the use of these data. Estimation of the posterior mean $(1 - B_i)y_i$, for example, is achieved by averaging either the $(1 - B_i^{(j)})y_i$ values or the $\theta_i^{(j)}$ values for $j = 1, \ldots, m$, but these values will not agree precisely.

The preceding can be expanded straightforwardly to encompass an unknown prior mean so that $\alpha = (\mu, A)$, $\mu = E\theta_i$, or to include regression forms $E\theta_i = \beta' x_i$ with $\alpha = (\beta_1, \ldots, \beta_r, A) = (\beta', A)$ and each x_i a known vector, assuming a normal or a flat prior distribution is used for μ or β.

The simulations described here will be costly when substantial accuracy is required, because that usually necessitates large m and many iterations. The number of iterations will be reduced dramatically, however, if a good starting approximation to the distribution $r_3(\alpha \mid y)$ is available. Assuming the model (2.1)–(2.3), a suggestion follows, based on Morris (1987).

Let $l(A)$ denote the logarithm of the modified posterior density $A \cdot r_3(A \mid y)$, and let $l'(A)$, $l''(A)$ be its first two derivatives.

(a) Find $\hat{A} > 0$ a (usually unique) value satisfying $l'(\hat{A}) = 0$, with

$$2l'(A) = -(k + q)A^{-1} + \lambda A^{-2} + \sum_1^k y_i^2/(V_i + A)^2. \tag{2.6}$$

(b) Define

$$p \equiv (-2\hat{A}^2 l''(\hat{A}))^{-1/2}. \tag{2.7}$$

(c) For $j = 1, \ldots, m$, let $u_j = (j - .5)/m$, and set

$$A^{(j)} = \hat{A}(u_j/(1 - u_j))^p. \tag{2.8}$$

The values (2.8) are approximations to the expected order statistics from the posterior density $r_3(A \mid y)$.

Tanner and Wong are to be congratulated for offering this approach to Bayesian simulation. Perhaps this note helps to expand further the usefulness of their method and to emphasize that their "missing data" concept can be used to include unknown parameters or latent data. Promising as their data augmentation method is, however, much still must be learned about affordable accuracy in large problems, about criteria for stopping the iterations (1.5)–(1.6), and whether simulation times are feasible in high-dimensional problems.

ADDITIONAL REFERENCE

Morris, C. N. (1987), "Empirical Bayes Interval Estimation," Technical Report 42, The University of Texas at Austin, Center for Statistical Sciences.

Comment

A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm

DONALD B. RUBIN*

Tanner and Wong are to be congratulated for this fine addition (henceforth, TW) to modern statistical theory and practice, which makes heavy use of current computational capabilities to draw inferences using simulation techniques. I firmly believe (Rubin 1985, sec. 2.5) that given today's computing environments, obtaining inferences in applied problems using simulation is often highly desirable because of the resultant flexibility of matching models to data without tedious and tangential mathematical analysis.

I expect that TW will simulate a variety of interesting efforts involving inference via simulation. For instance, it is easy to imagine a major article or sequence of articles providing details of the data augmentation algorithm in the series of EM examples outlined in Dempster, Laird, and Rubin (1977), and explicated and extended in subsequent literature. Similarly, more technical articles on this algorithm concerning rates of convergence, choices of

* Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This comment was written while the author was visiting the University of Warwick, England, and was also partially supported by National Science Foundation Grants SES-8311428 and DMS-8504332.

number of replications at each iteration, methods for speeding convergence, and so forth certainly should be forthcoming.

1. RELEVANCE OF THE DATA-AUGMENTATION ALGORITHM TO THE CREATION OF MULTIPLY-IMPUTED DATA BASES

TW should also stimulate related work concerning the use of simulated complete data sets—multiply-imputed data sets—to draw inferences. I have been deeply involved in this topic in recent years [work now summarized in Rubin (1987)], although my efforts have been primarily focused on cases with only a few (e.g., $m = 2$–10) draws from the posterior predictive distribution of the missing values, $p(z \mid y)$ in TW's notation, where z is the missing data and y is the observed data. This focus has arisen from the desirability of supplementing an incomplete public-use data base with a few imputations for each missing value so that the resultant multiply-imputed public-use data base can be analyzed using standard complete-data methods applied to each of the m data sets completed by imputation. Such complete-data analyses can be easily combined to form one posterior distribution that is conditional on the output of the m complete-data analyses rather than on y (Rubin 1987, chap. 3). Inferences based on this posterior distribution, which include an adjustment for small m, have been evaluated (Li 1985; Rubin 1987, chap. 4; Rubin and Schenker 1986), and they can be perfectly adequate even when m is 2 or 3, provided that the fraction of missing information is modest, where the fraction of missing information about the parameter Θ is measured with normal posterior distributions, for example, by the eigenvalues of $\text{var}[E(\Theta \mid y, z) \mid y]$ relative to $\text{var}(\Theta \mid y)$.

Creating a multiply-imputed data set can be a major effort, however, depending on the model and pattern of missing data. Consequently, TW's work on the data augmentation algorithm initially appeared to be of great relevance to this work, even though the algorithm could be expensive to apply to large public-use data bases. After some reflection, however, I believe that in those situations where useful inferences can be drawn from a multiply-imputed data set with small m (i.e., situations where the fraction of missing formation is modest), a few draws from $p(z \mid y)$ can often be easily created *noniteratively* using a method that may be of general interest to potential users of the data augmentation algorithm, especially considering that it can be applied even when the imputation step of the data augmentation algorithm is intractable.

2. A SAMPLING/IMPORTANCE RESAMPLING ALGORITHM FOR DRAWING A FEW IMPUTATIONS FROM THE POSTERIOR PREDICTIVE DISTRIBUTION OF THE MISSING VALUES

Step 1. Obtain a decent first-pass approximation to the joint posterior density of (Θ, z) say $h(\Theta, z \mid y) > 0$ for all possible (Θ, z), usually formulated as

$$h(\Theta, z \mid y) = h(\Theta \mid y)h(z \mid \Theta, y),$$

where $h(\Theta \mid y)$ is an approximate posterior density for Θ, and $h(z \mid \Theta, y)$ is an approximate posterior predictive density for z given Θ.

Step 2. Draw M values of (Θ, z) at random from $h(\Theta, z \mid y)$, where M is large relative to m = the final number of imputations desired for each missing value. Call these (Θ_j, z_j) $(j = 1, \ldots, M)$.

Step 3. Calculate the importance ratios for each (Θ_j, z_j),

$$r(\Theta_j, z_j \mid y) \propto p(y, z_j \mid \Theta_j)p(\Theta_j)/h(\Theta_j, z_j \mid y),$$

where the actual joint sampling density for (y, z), $p(y, z \mid \Theta)$, is designed to be easy to evaluate (up to a multiplier depending only on y) by the construction of the missing data, z.

Step 4. Draw m values of z from the z_j $(j = 1, \ldots, M)$ with probability proportional to $r_j = r(\Theta_j, z_j \mid y)$, thereby creating m values of z, z_l^* $(l = 1, \ldots, m)$; the associated values of Θ, Θ_l^* $(l = 1, \ldots, m)$ are not needed for imputation. Methods for such drawing appear in the survey literature for pps sampling (e.g., Cochran 1977, chap. 9).

If $p(z \mid \Theta, y)$ is tractable in the sense that (a) the imputation step of the data augmentation algorithm is straightforward [i.e., if $h(z \mid \Theta, y) = p(z \mid \Theta, y)$] and (b) $p(z \mid \Theta, y)$ can be explicitly evaluated [i.e., if $p(y \mid \Theta) = p(y, z \mid \Theta)/p(z \mid \Theta, y)$, in addition to $p(y, z \mid \Theta)$, can be explicitly evaluated up to a multiplier depending only on y], the sampling/importance resampling (SIR) algorithm can be streamlined. In particular, in Step 2, only the Θ_j then need to be drawn, since the r_j calculated in Step 3 and used in Step 4 do not depend on the z_j:

$$\begin{aligned} &r(\Theta_j, z_j \mid y) \\ &\quad \propto p(y \mid \Theta_j)p(z_j \mid \Theta_j, y)p(\Theta_j)/[p(z_j \mid \Theta_j, y)h(\Theta_j \mid y)] \\ &\quad \propto p(y \mid \Theta_j)p(\Theta_j)/h(\Theta_j \mid y). \end{aligned}$$

Having calculated the r_j in Step 3 from the Θ_j $(j = 1, \ldots, M)$, then in Step 4, m values of Θ_j are drawn with probability proportional to r_j, and for each Θ_l^*, z_l^* is drawn from $p(z \mid \Theta = \Theta_l^*, y)$, thereby creating the m imputed values of z.

An important feature of the SIR algorithm is that only one set of M (or m in the streamlined version) imputations of z is created. When a data base is large, the number of missing values can also be large, even with a small fraction of missing information, and then repeatedly passing through the database and creating sets of imputations for z can be quite expensive. Furthermore, the process of drawing M values of (Θ, z) (or Θ in the streamlined version) in Step 2 is designed to be relatively inexpensive by the choice of $h(\Theta, z \mid y)$ [or $h(\Theta \mid y)$ in the streamlined version].

The rationale for the SIR algorithm is based on the fact that as $M/m \to \infty$, the m values (z_l^*, Θ_l^*) are drawn with

probabilities given by

$$h(\Theta, z \mid y) \frac{r(\Theta, z \mid y)}{\int\int h(\Theta, z \mid y) r(\Theta, z \mid y)\, d\Theta\, dz}$$

$$= \frac{p(y, z \mid \Theta) p(\Theta)}{\int\int p(y, z \mid \Theta) p(\Theta)\, d\Theta\, dz}$$

$$= p(z, \Theta \mid y),$$

which implies that the m imputations, z_l^* ($l = 1, \ldots, m$), are independent draws from $p(z \mid y)$, as desired. This sampling/importance resampling technique for simulating a posterior distribution has been previously applied in small sample logistic regression problems (Rubin 1983).

The choice of an adequate ratio M/m depends on the fraction of missing information, γ: smaller γ implies satisfactory inferences from smaller M/m for two reasons. First, $h(\Theta, z \mid y)$ should be a better approximation to $p(\Theta, z \mid y)$ with smaller γ, since $p(\Theta \mid y, z)$ is easy to find—a simple measure of the adequacy of the approximation is the variance of the log r_j. Second, with smaller γ, final inferences from a multiply-imputed data set will be proportionately more determined by y than the imputed data, so there is less sensitivity to the imputed values. If γ is large, very large ratios M/m may be required for adequate performance of SIR, but in such cases a multiply-imputed data base with modest m is of limited utility anyway. In common practical cases, I expect that $M/m = 20$ will often be more than adequate, especially considering that with modest m, accurately approximating the tails of $p(z \mid y)$ is of limited importance.

3. EXAMPLE: MISSING VALUES IN A NORMAL/CONDITIONALLY NORMAL BIVARIATE SAMPLE

The case of a bivariate normal sample with missing values on both variables is a classic example (e.g., Wilks 1932) of a missing-data problem without a general closed-form solution. This case, however, is easily handled by the EM, data-augmentation, and streamlined SIR algorithms. A slightly modified situation, which has no closed-form solution and cannot be directly handled by either the EM or data-augmentation algorithms, will be used to illustrate the general SIR algorithm.

Specifically, let (w_1, w_2) be an iid sample from

$$w_1 \mid \Theta \sim N(\mu, \sigma^2), \tag{1}$$

$$w_2 \mid \Theta, w_1 \sim N(\alpha + \beta w_1 + \gamma w_1^2, \tau^2), \tag{2}$$

where

$$\Theta = (\mu, \log \sigma, \alpha, \beta, \gamma, \log \tau) \tag{3}$$

and

$$p(\Theta) \propto \text{constant}. \tag{4}$$

Suppose that a sample of n units is taken, where n_1 units have only w_1 observed, n_2 units have only w_2 observed, and n_{12} units have both w_1 and w_2 observed, where $N = n_1 + n_2 + n_{12}$, $n_{12} > n_1 + n_2 > n_2 > n_1$, and we assume that the missing data are missing at random (Rubin 1976). In TW's notation, y consists of the n_1 observations of w_1, the n_2 observations of w_2, and the n_{12} observations of (w_1, w_2), and z consists of the n_1 missing values of w_2 and the n_2 missing values of w_1. Notice that the fraction of missing information for each component of Θ is constructed to be modest, especially if τ^2 is small relative to $\text{var}(w_2 \mid \Theta)$ or if $n_{12} \gg n_1 + n_2$.

The only specific issue in implementing the SIR algorithm in this example is Step 1: choosing a decent approximation $h(\Theta, z \mid y) = h(\Theta \mid y) h(z \mid \Theta, y)$. Note that the density $p(y, z \mid \Theta)$ is easily evaluated, since it is the product over the n units of the two normal densities implied by (1) and (2), and from (3) and (4), this is the numerator of the importance ratio, $r(\Theta, z \mid y)$, used in Step 3.

Regarding $h(\Theta \mid y)$, it can be easily approximated by independent normal densities:

$$\log \sigma \mid y \sim N(\log s_1, [2(n_1 + n_{12} - 1)]^{-1}), \tag{5}$$

$$\mu \mid y \sim N(\bar{w}_1, s_1^2/(n_1 + n_{12})), \tag{6}$$

$$\log \tau \mid y \sim N(\log \hat{\tau}, [2(n_{12} - 3)]^{-1}, \tag{7}$$

and

$$(\alpha, \beta, \gamma) \mid y \sim N((\hat{\alpha}, \hat{\beta}, \hat{\gamma}), \hat{\tau}^2 C), \tag{8}$$

where $\bar{w}_1$ and s_1^2 are the mean and variance of the $(n_1 + n_{12})$ observations of w_1, and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\tau}^2, C)$ are the standard least squares summaries obtained by regressing w_2 on $(1, w_1, w_1^2)$ using the n_{12} observations of (w_1, w_2). Similarly, $h(z \mid \Theta, y)$ can be easily approximated by $n_1 + n_2$ independent normal densities:

$$n_1 \text{ missing } w_2 \mid \Theta, y \overset{\text{ind}}{\sim} N(\hat{\alpha} + \hat{\beta} w_1 + \hat{\gamma} w_1^2, \hat{\tau}^2), \tag{9}$$

and

$$n_2 \text{ missing } w_1 \mid \Theta, y \overset{\text{ind}}{\sim} N(a + b w_2 + c w_2^2, s_e^2), \tag{10}$$

where (a, b, c, s_e^2) are the standard least squares summaries obtained by regressing w_1 on $(1, w_2, w_2^2)$ using the n_{12} observations of (w_1, w_2). Thus $h(\Theta, z \mid y)$ is the product of the $(n_1 + n_2 + 3)$ univariate normal densities specified by (5), (6), (7), (9), and (10), and the trivariate normal density specified by (8), and is therefore easy to draw from at Step 2, and easy to evaluate as the denominator of the importance ratios, $r(\Theta, z \mid y)$, in Step 3. Better approximations are available especially for small n_{12}, but it is not clear whether they are worth the effort in the context of SIR relative to increasing the ratio M/m.

ADDITIONAL REFERENCES

Cochran, W. G. (1977), *Survey Techniques*, New York: John Wiley.
Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

———(1983), "Progress Report on Project for Multiple Imputation of 1980 Codes," University of Chicago.
———(1985), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
———(1987) *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.
Wilks, S. S. (1932), "Moments and Distributions of Estimates of Population Parameters From Fragmentary Samples," *Annals of Mathematical Statistics*, 2, 163–195.

Comment

SHELBY J. HABERMAN*

Tanner and Wong have provided a fascinating illustration of how serious problems of Bayesian inference can be addressed. Their approach is of interest not only to committed Bayesians but also to other statisticians who need to evaluate posterior distributions without unrealistically simple models or unrealistically simple prior distributions. Their iterative use of sampling is of particular interest.

The techniques used in this article raise a question addressed briefly by the authors. Computational requirements are quite considerable even for small problems, and the techniques used do involve judgment in their use. These problems may decrease in significance as computers improve and as investigators have more practice with the authors' approach; nonetheless, the really important question is the frequency with which the approach of Tanner and Wong will be used effectively by regular users of statistical procedures.

This computational issue is particularly troublesome given that statistical analyses of the examples are available that involve little calculation. To begin, consider the linkage example. In this case, the maximum likelihood estimate of θ is obtained by solution of a simple quadratic equation, and there is a straightforward normal approximation to the distribution of the estimate. In samples of adequate size for serious study of linkage, the normal approximation is quite adequate. A disciple of Bayes can obtain a polynomial expression for the posterior density of θ given the data under the uniform prior of the authors or even under more realistic mixtures of beta priors. Thus this example serves to illustrate feasibility of the method rather than an efficient way to treat the problem at hand.

Normally, procedures such as the EM algorithm are readily used to obtain maximum likelihood estimates in the covariance example. To be sure, the sample is too small for reliance on the normal approximation for the distribution of maximum likelihood estimates, but it is necessary to ask why anyone would try to apply a five-parameter model to so little data.

The last example can be analyzed by use of published algorithms that perform maximum likelihood estimation and obtain asymptotic standard errors of parameter estimates with little expenditure of computer time. In contrast, the analysis in this article required almost 3 hours on a VAX 750.

Thus one issue that must be considered is under what circumstances will a typical consumer of statistics be willing to make the additional investment in the authors' techniques given their current cost and given the possibility that alternative methods of analysis may be satisfactory enough.

A second issue to consider involves Bayesian inference. The use of prior distributions derived from the Dirichlet family by conditioning seems both inefficient and unrealistic. For example, it seems more realistic in the latent-class example to use a joint multivariate normal prior on the log-odds ratios $\log(P_{abcdx}/P_{11111})$, for a, b, c, d, and x equal to 1 or 2. Such a prior is easily constructed given that these log-odds satisfy a linear model under the traditional latent-class model.

Despite these reservations, Tanner and Wong are to be congratulated for their powerful new approach to Bayesian inference. I await with interest their views on the likely future of their method.

* Shelby J. Haberman is Professor, Department of Statistics, Northwestern University, Evanston, IL 60201.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

———(1983), "Progress Report on Project for Multiple Imputation of 1980 Codes," University of Chicago.
———(1985), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
———(1987) *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.
Wilks, S. S. (1932), "Moments and Distributions of Estimates of Population Parameters From Fragmentary Samples," *Annals of Mathematical Statistics*, 2, 163–195.

Comment

SHELBY J. HABERMAN*

Tanner and Wong have provided a fascinating illustration of how serious problems of Bayesian inference can be addressed. Their approach is of interest not only to committed Bayesians but also to other statisticians who need to evaluate posterior distributions without unrealistically simple models or unrealistically simple prior distributions. Their iterative use of sampling is of particular interest.

The techniques used in this article raise a question addressed briefly by the authors. Computational requirements are quite considerable even for small problems, and the techniques used do involve judgment in their use. These problems may decrease in significance as computers improve and as investigators have more practice with the authors' approach; nonetheless, the really important question is the frequency with which the approach of Tanner and Wong will be used effectively by regular users of statistical procedures.

This computational issue is particularly troublesome given that statistical analyses of the examples are available that involve little calculation. To begin, consider the linkage example. In this case, the maximum likelihood estimate of θ is obtained by solution of a simple quadratic equation, and there is a straightforward normal approximation to the distribution of the estimate. In samples of adequate size for serious study of linkage, the normal approximation is quite adequate. A disciple of Bayes can obtain a polynomial expression for the posterior density of θ given the data under the uniform prior of the authors or even under more realistic mixtures of beta priors. Thus this example serves to illustrate feasibility of the method rather than an efficient way to treat the problem at hand.

Normally, procedures such as the EM algorithm are readily used to obtain maximum likelihood estimates in the covariance example. To be sure, the sample is too small for reliance on the normal approximation for the distribution of maximum likelihood estimates, but it is necessary to ask why anyone would try to apply a five-parameter model to so little data.

The last example can be analyzed by use of published algorithms that perform maximum likelihood estimation and obtain asymptotic standard errors of parameter estimates with little expenditure of computer time. In contrast, the analysis in this article required almost 3 hours on a VAX 750.

Thus one issue that must be considered is under what circumstances will a typical consumer of statistics be willing to make the additional investment in the authors' techniques given their current cost and given the possibility that alternative methods of analysis may be satisfactory enough.

A second issue to consider involves Bayesian inference. The use of prior distributions derived from the Dirichlet family by conditioning seems both inefficient and unrealistic. For example, it seems more realistic in the latent-class example to use a joint multivariate normal prior on the log-odds ratios $\log(P_{abcdx}/P_{11111})$, for a, b, c, d, and x equal to 1 or 2. Such a prior is easily constructed given that these log-odds satisfy a linear model under the traditional latent-class model.

Despite these reservations, Tanner and Wong are to be congratulated for their powerful new approach to Bayesian inference. I await with interest their views on the likely future of their method.

* Shelby J. Haberman is Professor, Department of Statistics, Northwestern University, Evanston, IL 60201.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

Comment

A. O'HAGAN*

The idea of data augmentation is certainly intriguing. I have tried to consider in what applications it would be valuable. My problem in this regard is that the authors have not provided such applications themselves. When one sees a new numerical technique proposed, one expects to be shown examples where the new method requires less computation than the old ones. The opposite is true of the present article. In all of the examples there is a single parameter whose unnormalized posterior density may be written down directly. Simple numerical integration suffices to normalize it and obtain any summaries of interest. The authors have, therefore, not demonstrated any improvement over existing, well-tried techniques.

It is true that in the genetic linkage model a normal approximation to the posterior density is poor, but who would use a normal approximation for a parameter in [0, 1]? A beta approximation, having the same mode and curvative at the mode, is easily fitted. Trivial computation shows it to be excellent in all of the cases shown in Figures 1, 3, and 4. Indeed, on the scale used in those diagrams it would also "superimpose" the true posterior. For the record, the beta approximation in the case of Figure 4 is Beta(8.163, .872).

So when would data augmentation be valuable? Presumably only when existing methods require very substantial computation. Furthermore, there must be a data augmentation available that dramatically improves the problem. I suggest, therefore, that the authors might look at problems with many parameters, for here we can obtain high-dimensional posterior densities that require laborious numerical integration to obtain moments or marginal densities. Then we require a data augmentation to simplify the posterior density so that it becomes very much more tractable.

It is hard to think of applications having both of these features. However, the authors could examine some problems of the following kind. A traditional designed experiment, such as a factorial experiment, is performed but has missing data. With the full data set the likelihood factorizes and, supposing that the corresponding parameters are independent a priori, they will be independent a posteriori. Therefore, data augmentation makes marginal densities immediately available, and other summaries are easily computed. With missing data, orthogonality is lost and the likelihood no longer factorizes in the same way. If we had normal prior distributions, we could still invoke general linear model theory and obtain marginal densities, et cetera. Nonnormal (but independent) priors, however, would induce a massive numerical integration problem. This is still not a good example, because it requires rather special kinds of prior belief. On the other hand, one often has prior knowledge about the sign of an effect, which would produce truncated priors. Alternatively, it is my opinion that prior beliefs often demand thick-tailed prior distributions. So, if one is happy to accept normality in the likelihood, data augmentation may offer a tractable way of using such priors.

I hope that the authors will justify their approach better in the future by presenting convincing examples. Until then I will remain doubtful of its practical value. Opportunities for data augmentation are not obviously common; *good* opportunities, where data augmentation produces computational advantages, seem likely to be rare.

* A. O'Hagan is Lecturer, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

Rejoinder

MARTIN A. TANNER and WING HUNG WONG

1. REPLY TO DEMPSTER, MORRIS, AND RUBIN

1.1 Hierarchical Models

Dempster and Morris point out the usefulness of data augmentation in the analysis of hierarchical models, and we agree with them completely. We are especially grateful to Morris for his derivation of the explicit calculations needed for the I and P steps in the multivariate normal means example. A type of educational data, recently brought to our attention by Michael Meltzer of the University of Chicago, can serve to illustrate the technique, as well as some possible extensions. In this application, one has student performance data y, which is modeled given school indicators X_1 and other student-level covariates X_2; that is, the observational model is

$$y = X_1 z_1 + X_2 z_2 + \varepsilon_1.$$

In addition, there is a latent model for the school effects given school characteristics (W) such as whether it is a Catholic or a public school; that is,

$$z_1 = W\gamma + \varepsilon_2.$$

Here y is the observed data, z is the latent data and $\theta = (\gamma, \sigma_1, \sigma_2)$ if the errors in the two models are assumed to be Gaussian with standard deviations σ_1 and σ_2, respectively. Typically, an EM-type algorithm is used to obtain the maximum likelihood estimate of θ (e.g., Raudenbush and Bryk 1986). As remarked by Dempster, however, it would be desirable to perform full Bayesian analyses with genuine priors, and data augmentation offers a natural method to accomplish this. Under standard conjugate priors, the computations needed for the I and P steps can be derived in the same way as in the example in Morris's discussion. To see the effect of changing priors, one can attach weights (ratio between the true prior and the conjugate prior) to the sampled values of θ.

Very often it is desirable and prudent to perform the analysis using heavy-tailed errors, especially for the latent model, for which diagnostics are hard to come by and a Gaussian error may lead to unwarranted pooling of truly outlying schools. Thus a natural elaboration of the above model is to use a t distribution with a small degree of freedom for ε_2. The data augmentation scheme can also be extended to handle this situation. This is done by using the normal/gamma mixture representation of the t distribution; that is, ε_2 is distributed as normal/$\sqrt{q}$, where q is a mean square variable. In such a problem, there are three sets of quantities whose posterior distributions may be of interest, namely, z, q, and θ. The posterior distribution of either one given the other two is easy to obtain. The data augmentation algorithm will now be extended to sample iteratively from each of these conditional posterior distributions in turn (compare Li 1985). We hope that this example can illustrate the freedom and flexibility offered by the data augmentation scheme (and to a much greater extent, the Bayesian outlook) for the application to, and study of, more complex and realistic models—the need for which being succinctly summarized by Dempster in his opening paragraph.

1.2 Importance Sampling

Both Dempster and Rubin comment on the use of importance sampling, and Rubin proposes a highly innovative method (SIR algorithm) for creating a few imputations when the fraction of missing information is modest, which also exploits the simplicity of inference based on the augmented data. We agree with their opinion that importance sampling is a powerful tool that should be employed whenever appropriate. We regard data augmentation iteration and importance sampling as complementary, however, rather than rival concepts. To clarify this issue, we compare the efficiency of the two procedures in the canonical problem of estimating a posterior moment of the parameters. We do, however, wish to point out that in so doing we have distorted Rubin's intention, since the SIR algorithm was initially designed for a more limited task under more limited conditions.

Recall that the basic principle of importance sampling is to generate values from a trial density, and then weight the values thus generated according to

$$\text{weight} \propto \frac{\text{true density}}{\text{trial density}}.$$

Let $g_*(\theta)$ be the true posterior density, $g_i(\theta)$ be the estimate for $g_*(\theta)$ at the ith iteration of the data augmentation algorithm, $h_*(\theta, z)$ be the true joint posterior density for θ and z, and $h(\theta, z)$ be the trial density in Rubin's SIR algorithm. To develop a quantitative comparison between the data augmentation algorithm and the SIR algorithm, let us consider the estimation of a specific posterior moment

$$\rho = \int a(\theta) g_*(\theta)\, d\theta.$$

For the data augmentation algorithm, we can use the estimator

$$\hat{\rho}_i = \frac{1}{m} \sum_{j=1}^{m} a(\theta_j),$$

where $\{\theta_1, \ldots, \theta_m\}$ is a sample from $g_i(\theta)$ at the ith iteration. For the SIR algorithm, we can use the estimator

$$\hat{\rho}_h = \frac{\sum r_j a(\theta_j)}{\sum r_j},$$

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

where $\{(\theta_1, z_1), \ldots, (\theta_m, z_m)\}$ is a sample from $h(\theta, z)$ and $\{r_1, \ldots, r_m\}$ are the weights calculated as in Step 3 of Rubin's discussion, his key observation being that the true joint posterior

$$h_*(\theta, z) \propto p(\theta)p(y, z \mid \theta)$$

is easy to compute (up to proportionality constant) because of the simplicity of the augmented data likelihood $p(y, z \mid \theta)$.

It is easy to obtain expressions for the means and variances:

$$E(\hat{\rho}_i) = \int a(\theta)g_i(\theta)\, d\theta,$$

$$\text{var}(\hat{\rho}_i) = \frac{1}{m}\left[\int a^2 g_i\, d\theta - (E\hat{\rho}_i)^2\right],$$

$$E(\hat{\rho}_h) = \int a(\theta)g_*(\theta)\, d\theta,$$

$$\text{var}(\hat{\rho}_h) = \frac{1}{m}\left[\int\int a^2(\theta)\frac{h_*(\theta, z)^2}{h(\theta, z)}\, dz\, d\theta - \rho^2\right].$$

We can now make the following comparisons: (a) $\hat{\rho}_h$ is unbiased and $\hat{\rho}_i$ is not—the bias of the latter, however, decreases with increasing i because g_i converges in L_1 to g_*; (b) although both variances decrease with increasing m, for any fixed m, the variance of $\hat{\rho}_h$ can be arbitrarily large because of its dependence on the ratio h_*/h.

Thus, if the trial density is not chosen well, the SIR estimate $\hat{\rho}_h$ can be very inaccurate. In practice, this difficulty is indicated if the weights come out to be highly skewed. Diagnostics for this phenomena are hence possible in the manner suggested by Rubin. But what should one do when the weights are highly skewed? It is here that we see a principal strength of the data augmentation algorithm, namely that each iteration offers an improvement over the previous estimate.

If the streamlined version of SIR is applicable, that is, when the likelihood $l(\theta \mid y)$ is easy to compute, then it is possible and may be desirable to use a procedure that combines the strength of both data augmentation iteration and importance sampling. Such a procedure would use data augmentation iteration to improve the initial approximation until the weights for importance sampling are satisfactorily distributed. Further effort will be necessary to formulate good strategies for such a combination and to explore its full potential.

2. REPLY TO HABERMAN

Haberman raises two related concerns regarding the potential acceptance of the data augmentation algorithm as a tool for routine data analysis. His first point relates to the computational intensiveness of the algorithm, and the second point questions whether the algorithm will be adopted "given the possibility that alternative methods of analysis may be satisfactory enough" (p. 546).

Regarding computational issues, it is noted that the algorithm is conceptually a very simple algorithm. One generates m augmented data sets, and then based on these data sets one repeatedly evaluates the augmented posterior or likelihood. Thus the computational burden is not due to the implementation of a complex algorithm, but rather to the repeated evaluation of the same set of operations. As we have noted in our article, the amount of work performed at each iteration by the IP algorithm is determined by the number of imputations. Thus it is a *highly parallel algorithm* in the sense that the implementation of the algorithm can be tailored to make full use of a multiprocessor machine by the careful selection of m. In contrast, regarding both the EM algorithm and Newton–Raphson algorithm, the user is constrained by the problem of how much of the resources of the machine can be used per iteration. In this way, on a multiprocessor machine, the IP algorithm may actually require a shorter execution time than either the EM or Newton–Raphson algorithms.

Haberman questions whether the data augmentation algorithm will be adopted for a particular problem given that "alternative methods of analysis may be satisfactory enough." Our opinion is that one does not know whether the normal approximation, for example, is *good enough* unless one is able to examine the likelihood (or posterior). Thus the data augmentation algorithm may be useful as a check on the validity of the normal approximation, as in the covariance example. Moreover, we do not believe that the plausibility of such departures from normality, as noted in the covariance example, can be ignored. Although fitting a five-parameter bivariate model to 12 observations may be unwise, what can one say regarding the analysis of 40 seven-dimensional observations? It is also important to note that an examination of the likelihood (or posterior) can sometimes allow one to check the *adequacy* of a given model. This is again exemplified in the covariance example. Certainly, under the bivariate normal model, we

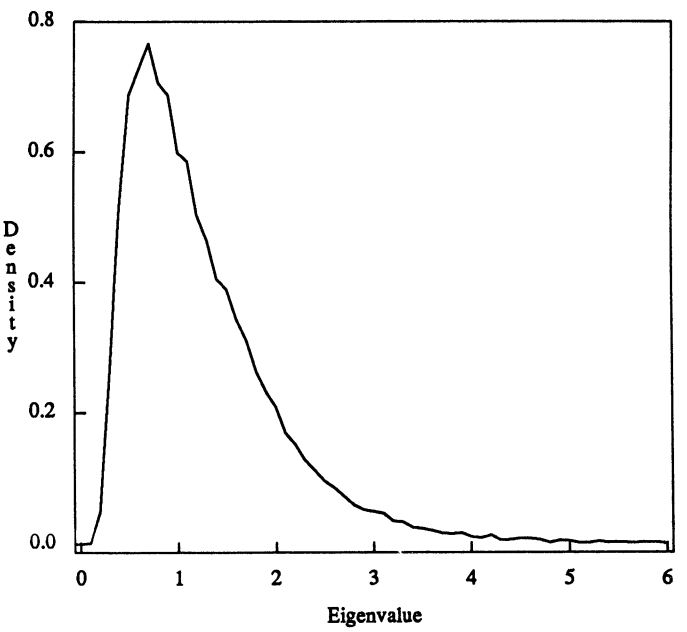

Figure 1. Posterior Distribution of the Smallest Eigenvalue—Bivariate Case.

would not expect 100 observations to yield the posterior encountered in our 12-observation data set. If such a posterior were to result from a large sample, however, we would be forced to question the validity of the bivariate normal assumption. In other words, even if we have a "large sample," we may find evidence against a model by noting some peculiarity in the likelihood. As such, by ignoring the likelihood (or posterior) and focusing on a point estimate, one may miss some feature of the data that relates to the adequacy of the model.

Regarding our analysis of the traditional latent-class model, Haberman seems to interpret our use of the Dirichlet distribution as a specification of the prior distribution for the conditional probabilities. We use the Dirichlet distribution as a vehicle for generating the probabilities, not to specify a prior. We can, however, incorporate a prior distribution by assigning weights to the probabilities.

Haberman queries our views on the future of the *method.* It is our opinion that the data augmentation algorithm is not only a useful method for calculating likelihoods (or posteriors) but it also provides a *paradigm* for handling missing data problems. For example in the context of grouped and censored data, we have used this paradigm to develop an algorithm for the nonparametric estimation of the hazard function (Tanner and Wong 1987).

3. REPLY TO O'HAGAN

We do not agree with the basic premise of O'Hagan that *the* measure of the value of a new methodological contribution is whether it minimizes numerical computation. We feel that data augmentation is valuable, not only as a numerical tool for computing posteriors, but as a way of thinking about certain types of problems (e.g., Morris's hierarchical model), as well as providing a diagnostic for both the normal approximation and the model. To focus, as does O'Hagan, on data augmentation as *only* a computational technique misses several major themes of the article.

It is unfortunate that O'Hagan misses several important features in our examples and as a consequence of this reaches the conclusion that the examples are unconvincing. He states: "In all of the examples there is a single parameter whose unnormalized posterior density may be written down directly. Simple numerical integration suffices to normalize it and obtain any summaries of interest" (p. 547). This is incorrect. In both the covariance and latent-class examples there are more than one parameter. In fact, the conditional independence model for the latent-class data has 12 parameters. We will be very surprised if in that example anyone can "write down" the unnormalized posterior density for an arbitrary parameter, for example, the quantity π_{11}^{AX}, upon which we focus. The point that we wish to make is that the data augmentation algorithm offers an approach of treating the nuisance parameters. This is because the algorithm enables one to readily obtain the marginal posterior distribution of any parameter or the joint posterior distribution of any two parameters. To illustrate, suppose that in the covariance example we are interested in the posterior distribution of the smallest eigenvalue. We can obtain this posterior distribution by finding the smallest eigenvalue for each matrix in the sample of covariance matrices drawn from the final estimate of $p(\Sigma \mid y)$. The histogram of these values yields the desired posterior (see Fig. 1 here).

O'Hagan suggests that we "look at problems with many parameters" (p. 547), as well as factorial experiments with missing data. In this regard, he seems to have missed the latent-class example. The associated log-linear model has 12 parameters and the associated 3×2^4 table, clearly, has a factorial design. O'Hagan is critical of the use of the normal approximation in the linkage example. We feel, however, that based on our analysis of the original data, the normal approximation is quite acceptable. Clearly, the adequacy of the normal approximation depends on the *nature of the data* and it is the use of data augmentation that can assist the statistician in assessing the adequacy of the approximation *for the particular data set.*

ADDITIONAL REFERENCES

Raudenbush, S., and Bryk, A. S. (1986), "A Hierarchical Model for Studying School Effects," *Sociology of Education,* 59, 1–17.

Tanner, M. A., and Wong, W. H. (1987), "An Application of Imputation to an Estimation Problem in Grouped Lifetime Analysis," *Technometrics,* 29, 23–32.