



The Calculation of Posterior Distributions by Data Augmentation

Author(s): Martin A. Tanner and Wing Hung Wong

Source: *Journal of the American Statistical Association*, Vol. 82, No. 398, (Jun., 1987), pp. 528-540

Published by: American Statistical Association

Stable URL: <http://www.jstor.org/stable/2289457>

Accessed: 19/08/2008 10:34

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

# 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  $p(\theta | y, z)$  can be calculated. But the posterior density that we want is  $p(\theta | y)$ , which may be difficult to calculate directly. If, however, one can generate multiple values of  $z$  from the predictive distribution  $p(z | y)$  (i.e., multiple imputations of  $z$ ), then  $p(\theta | y)$  can be approximately obtained as the average of  $p(\theta | y, z)$  over the imputed  $z$ 's. However,  $p(z | y)$  depends, in turn, on  $p(\theta | y)$ . Hence if  $p(\theta | y)$  was known, it could be used to calculate  $p(z | y)$ . This mutual dependency between  $p(\theta | y)$  and  $p(z | y)$  leads to an iterative algorithm to calculate  $p(\theta | y)$ . 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  $p(\theta | y, z)$  and  $p(z | \theta, y)$ . 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  $p(\theta | y, z)$  arising from parametric submodels of the multinomial, we develop a primitive but generally applicable way to approximately sample  $\theta$ . 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  $\theta$ . 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,  $p(\theta | x)$ , is of known form]. The method consists of iterating the following two steps: (a) Given the current guess of the posterior distribution of  $\theta$  given  $y$ , generate a sample of  $m > 0$  latent data patterns from the predictive distribution of  $z$  given  $y$ . (b) Update the posterior of  $\theta$ , 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

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:

$$p(\theta | y) = \int_z p(\theta | z, y)p(z | y) dz, \quad (2.1)$$

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

$$p(z | y) = \int_{\Theta} p(z | \phi, y)p(\phi | y) d\phi. \quad (2.2)$$

In the above equations, the sample space for the latent data  $z$  is denoted by  $Z$  and the parameter space for  $\theta$  is denoted by  $\Theta$ . (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  $p(\theta | y)$  must satisfy the integral equation

$$g(\theta) = \int K(\theta, \phi)g(\phi) d\phi,$$

where  $K(\theta, \phi) = \int p(\theta | z, y)p(z | \phi, y) dz.$  (2.3)

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

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

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

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

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

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  $p(\theta | y)$ ,

- (a) generate a sample  $z^{(1)}, \dots, z^{(m)}$  from the current approximation to the predictive density  $p(z | y)$
- (b) update the current approximation to  $p(\theta | y)$  to be

the mixture of conditional densities of  $\theta$  given the augmented data patterns generated in (a), that is,

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

In the above, we must either be able to calculate  $p(\theta | z, y)$  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  $p(z | y)$ . Given that the current approximation to  $p(\theta | y)$  is  $g_i(\theta)$ , (2.2) then suggests that  $z$  can be generated from the current predictive distribution in two steps:

- (a1) generate  $\theta$  from  $g_i(\theta)$ .
- (a2) generate  $z$  from  $p(z | \phi, y)$ , where  $\phi$  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  $p(\theta | x)$  over the augmented data patterns generated across iterations will converge to the  $p(\theta | y)$ . 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  $\theta$  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

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

and the augmented likelihood is of the form

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

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  $\theta$  from the current estimate of  $p(\theta | y)$ .
  - (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  $\theta$  equal to the mixture of beta distributions, mixed over the  $m$  imputed values of  $x_2$ ; that is,

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

where

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

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  $\theta$  is assumed to be uniform in  $(0, 1)$ .

Figure 1 presents the posterior density estimates of  $\theta$  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)

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

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

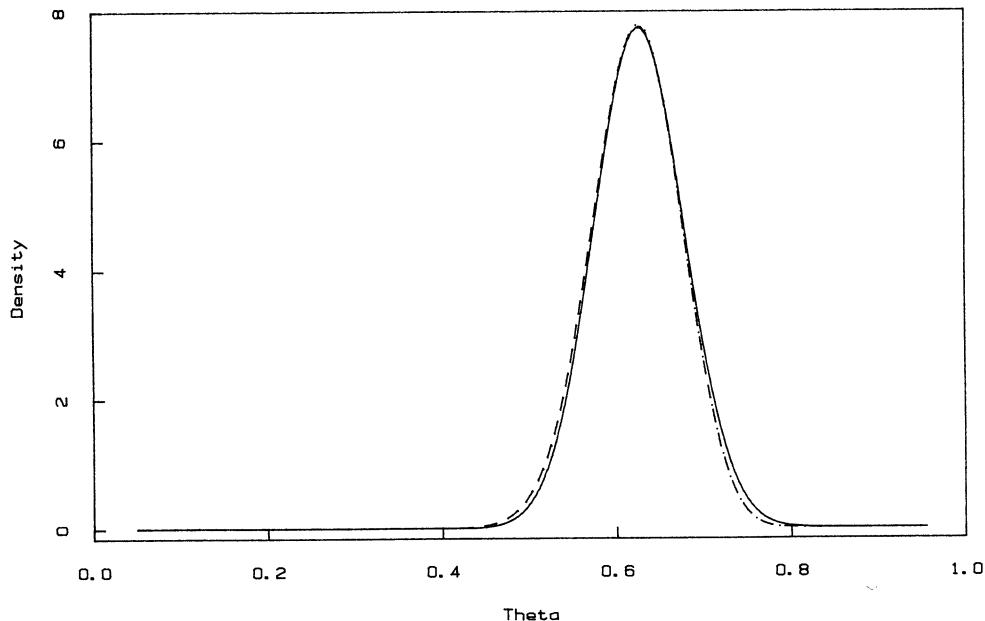


Figure 1. Posterior Density of  $\theta$  for Data (125, 18, 20, 34). The solid, dashed, and dotted lines represent the normal approximation, the estimated posterior distribution, and the true posterior, respectively. The dashed and dotted lines are superimposed.

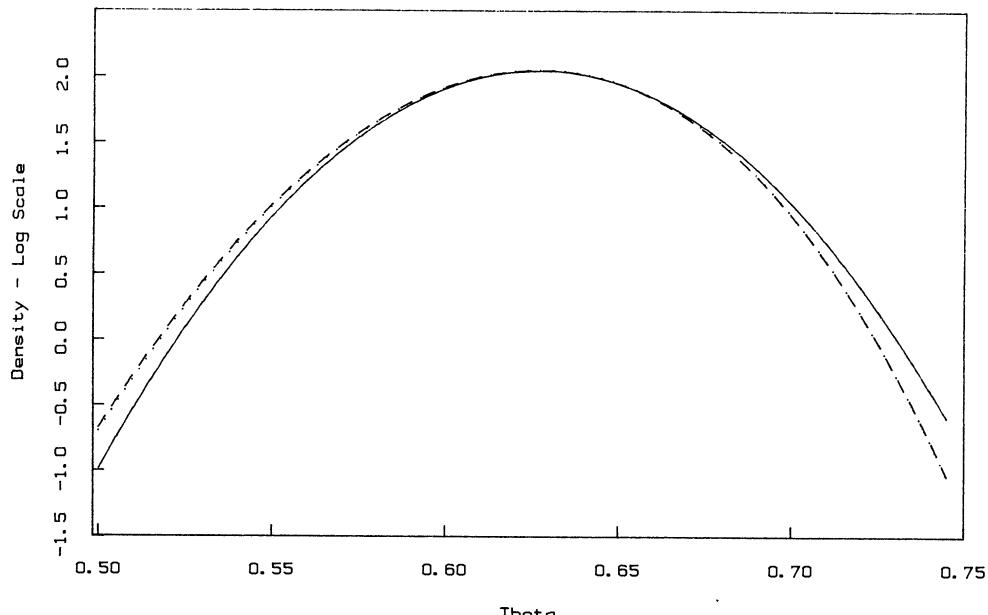


Figure 2. Log-Posterior Density of  $\theta$  (same data and legend as in Fig. 1).

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  $\rho$ , and variances  $\sigma_1^2$  and  $\sigma_2^2$ . 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  $\rho$  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  $\sigma_1^2$  and  $\sigma_2^2$  in the eight incomplete observations cannot be ignored because information regarding  $\sigma_1^2$  and  $\sigma_2^2$  is of use in making inference regarding  $\rho$ .

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

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

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

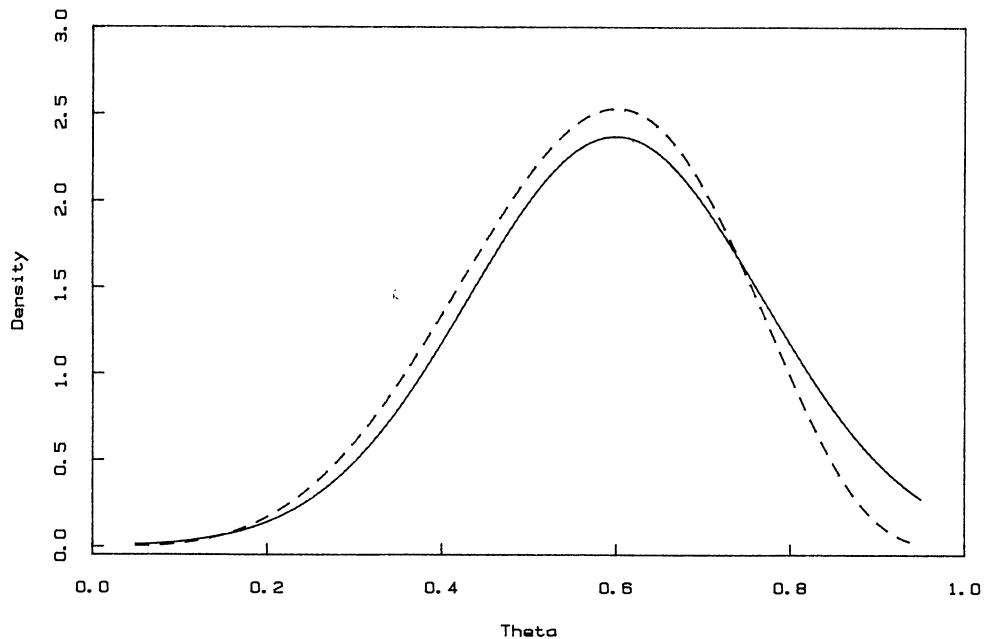


Figure 3. Posterior Density of  $\theta$  for Data (13, 2, 2, 3) (same legend as in Fig. 1). The dashed lines and dotted lines are superimposed.

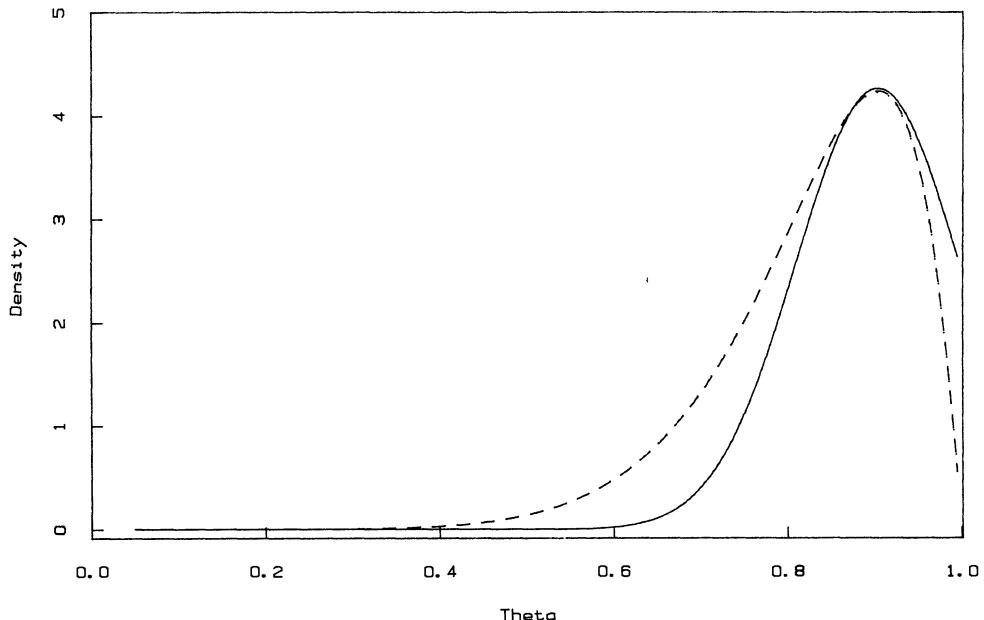


Figure 4. Posterior Density of  $\theta$  for Data (14, 0, 1, 5) (same legend as in Fig. 1). The dashed and dotted lines are superimposed.

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  $\Sigma$  is then generated from the current guess of the posterior distribution  $p(\Sigma | y)$ . At the first iteration,  $\rho$  can be generated from  $U[-1, 1]$  and  $\sigma_1^2$  and  $\sigma_2^2$  can be generated from weighted  $\chi^2_7$  distributions. At succeeding iterations, the updated posterior  $p(\Sigma | y)$  is a mixture of inverted Wishart distributions. This last point follows from the fact that  $p(\Sigma | x)$  is an inverted Wishart distribution (Box and Tiao 1973, p. 428) when the prior of  $\Sigma$  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 pos-

terior 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.

#### 4. THE DIRICHLET SAMPLING PROCESS

In the linkage example of Section 2, the augmented posterior distribution  $p(\theta | x)$  is a beta distribution. Thus it is a trivial matter to carry out the P step. In more complicated models, the sampling of  $\theta$  from  $p(\theta | x)$  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

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

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

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).

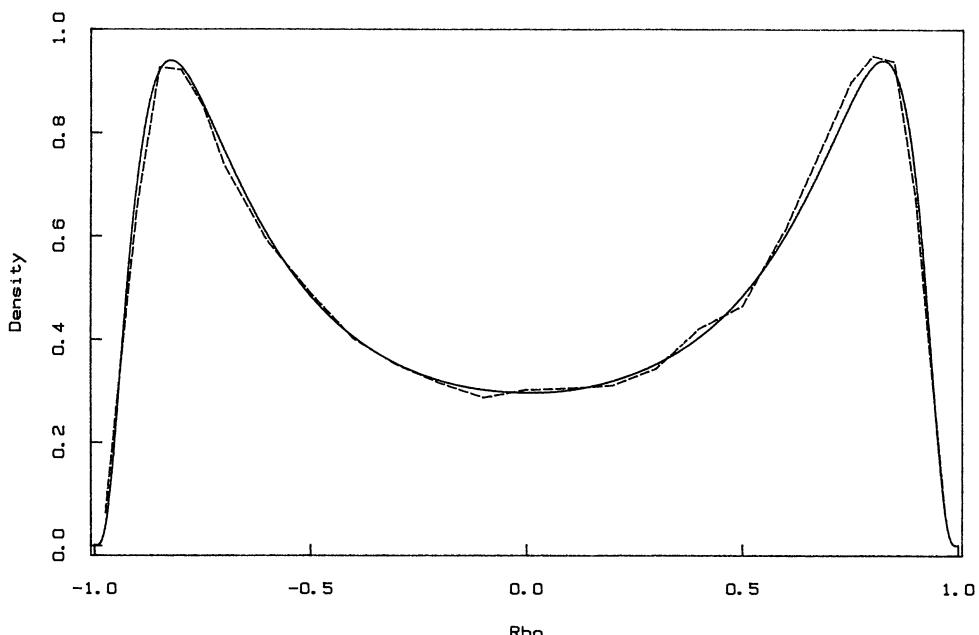


Figure 5. Posterior Density of the Correlation Coefficient. The solid and dashed lines represent the true and estimated posterior, respectively.

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

*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.

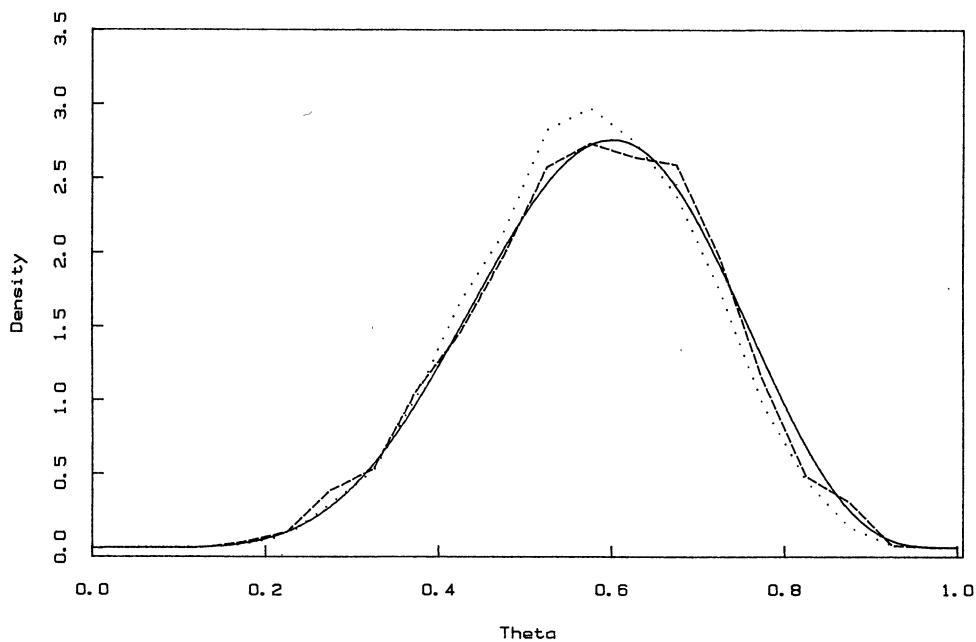


Figure 6. Posterior Density of  $\theta$  for Data (3, 2, 2, 3). The dotted, dashed, and solid lines represent the estimate based on 10,000 values, the estimate based on 3,000 values, and the true posterior distribution, respectively.

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  $\theta$ :  $P_2 = \theta/4$ ,  $P_3 = 1/4 - \theta/4$ ,  $P_4 = 1/4 - \theta/4$ , and  $P_5 = \theta/4$ , the least squares solution yields  $\hat{\theta} = 2(p_2 + p_5)$ . The approximate posterior distribution for  $\theta$  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  $\varepsilon$ -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  $\varepsilon$  is sufficiently small, then the  $\hat{\theta}$  values obtained in this way will have a distribution approximately equal to  $p(\theta | x)$ .

In practice, the value of  $\varepsilon$  is selected by plotting a sequence of estimated posterior distributions of  $\theta$  corresponding to a sequence of decreasing  $\varepsilon$  values. The curves tend to converge as the value of  $\varepsilon$  is decreased. The aforementioned procedure is generally applicable to parametric models for multinomial data if the cell probabilities are linear in  $\theta$  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  $\theta$ , 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  $\theta$  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  $\theta$  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

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).

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.

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  $\pi_{a1}^{AX}$ ). In conjunction with  $\pi_{a2}^{AX}$ , 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  $\pi_{11}^{AX}$  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  $\pi_{11}^{AX}$ , 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  $\pi_{11}^{AX}$  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  $D(n_{11111}, \dots, n_{22231}; n_{22232})$  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  $\pi_{11}^{AX}$  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  $\pi_{11}^{AX}$  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-

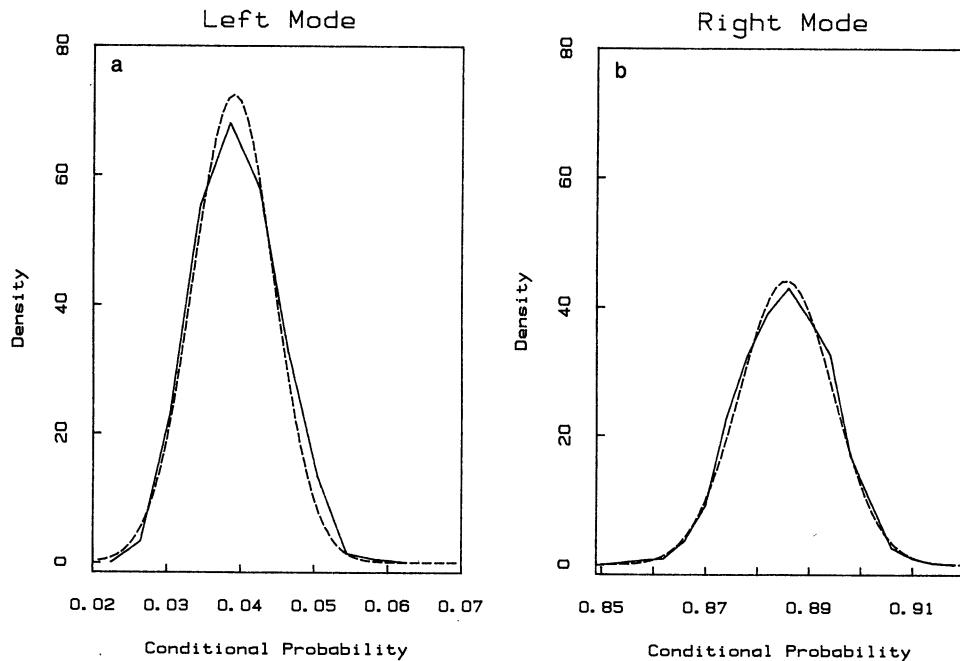


Figure 7. Posterior Density of  $\pi_{11}^{AX}$ . The solid and dashed lines represent the estimated and true posterior density, respectively. (a) Left mode. (b) Right mode.

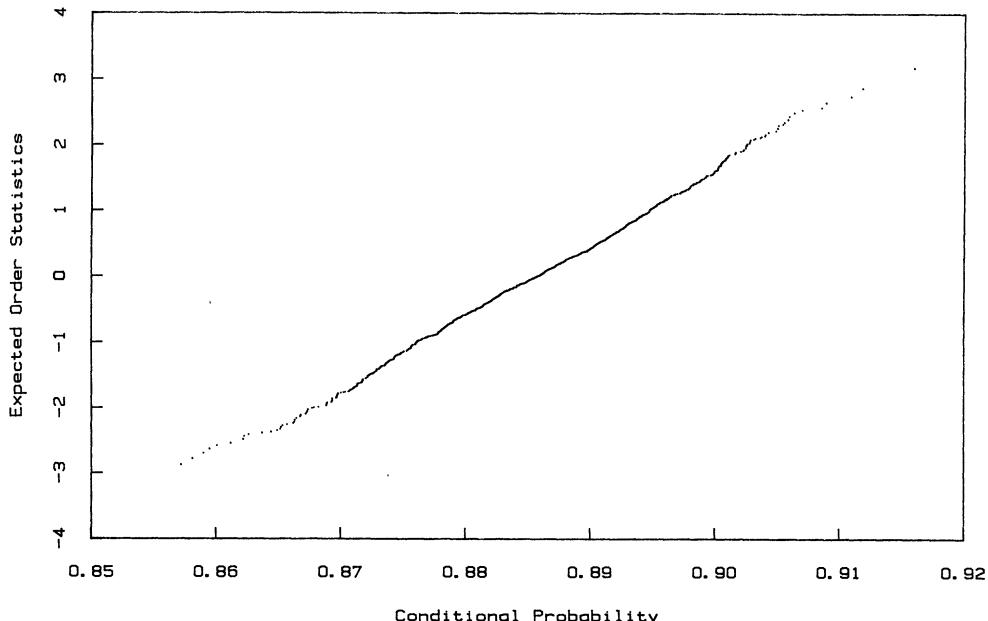


Figure 8. Rankit Plot for Right Mode.

gorithm should converge and at what rate it does so. For simplicity we will first assume that  $\Theta$  is a connected subset of  $R^p$ . The theory is essentially the same for discrete  $\Theta$ , as discussed briefly at the end of this section. Let  $L_1$  be the space of (Lebesgue) 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  $\theta$ , then  $\|Tf\| = \|f\|$ . (b) If  $f(\theta) \geq g(\theta)$  for all  $\theta$ , then  $Tf(\theta) \geq Tg(\theta)$  for all  $\theta$ . To prove the theorem, let  $f = g_i - g_*$ . Then

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

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

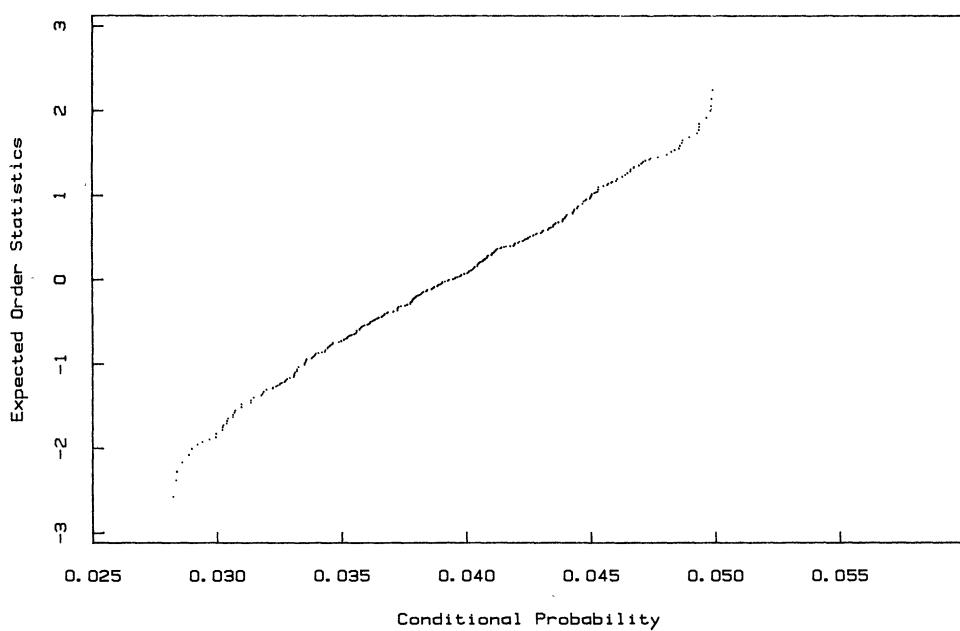


Figure 9. Rankit Plot for Left Mode.

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  $\theta$ . For any  $\theta_0 \in \Theta$ , there is an open neighborhood  $U$  of  $\theta_0$ , so  $K(\theta, \phi) > 0$  for all  $\theta, \phi \in U$ .

The second part of this condition says that if  $\theta$  and  $\phi$  are close, then it is possible to generate some latent data pattern  $z$  from  $p(z | \phi, y)$  such that  $p(\theta | z, y)$  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  $\theta_1 \rightarrow \theta$ , 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  $\theta_0$  such that  $K(\theta, \phi) > 0$  for all  $\theta, \phi \in U$ . Since  $\theta_0$  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  $\Theta$  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

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

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  $\alpha$  ( $0 < \alpha < 1$ ), 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  $\{f_{i_r}\}$  such that  $\|Tf_{i_r}\|/\|f_{i_r}\| \rightarrow \alpha$ , and  $f_{i_r}$  converges to some  $f_*$  weakly in  $L_1$ .

(d) Since the set  $\{f_{i_r}\}$  is bounded and equicontinuous, we must actually have  $f_{i_r}$  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  $\{f_{i_r}\}$  be a subsequence of  $\{f_i\}$  such that  $\|Tf_{i_r}\|/\|f_{i_r}\| \rightarrow \alpha$ . Now by (a) and (b),  $\{f_{i_r}\}$  is weakly sequentially compact, so there must exist a further subsequence  $\{f_{i_{r_s}}\}$  of  $\{f_{i_r}\}$  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  $\Theta$ , 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  $|\theta| \rightarrow \infty$  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  $\alpha$  is dependent on the initial value  $g_0$ . If  $\Theta$  is compact, it can be shown that the supremum of  $\alpha$  over all possible  $g_0$  is still less than 1; that is, we get a linear rate independent of the starting values. If  $\Theta$  is unbounded, however,  $\alpha$  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  $\Theta$ . 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

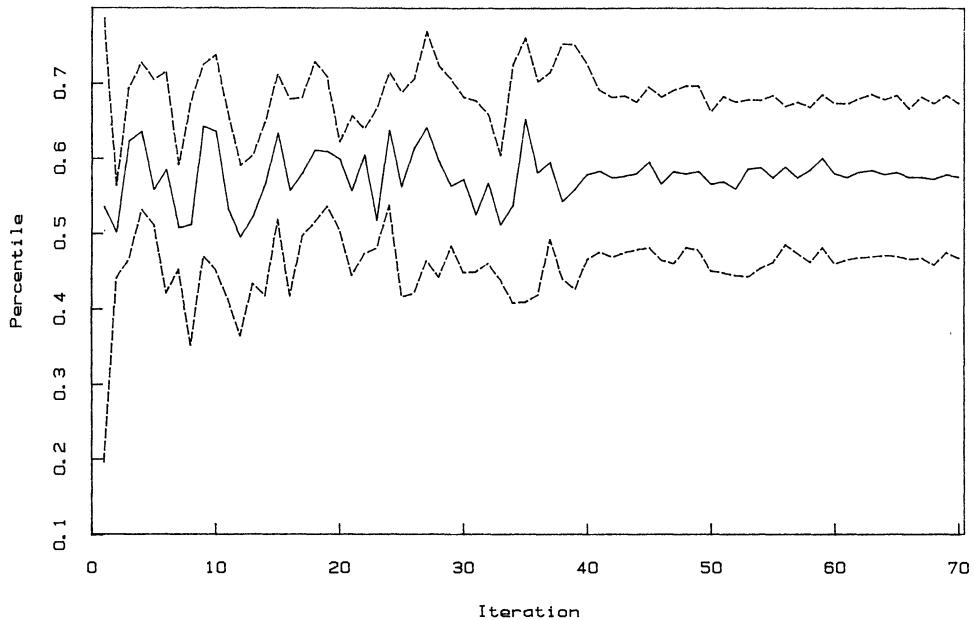


Figure 10. Median and Upper and Lower Quartiles of  $\theta$  Values Across Iterations. The upper dashed line, the solid line, and the lower dashed line represent the upper quartile, the median, and the lower quartile, respectively.

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

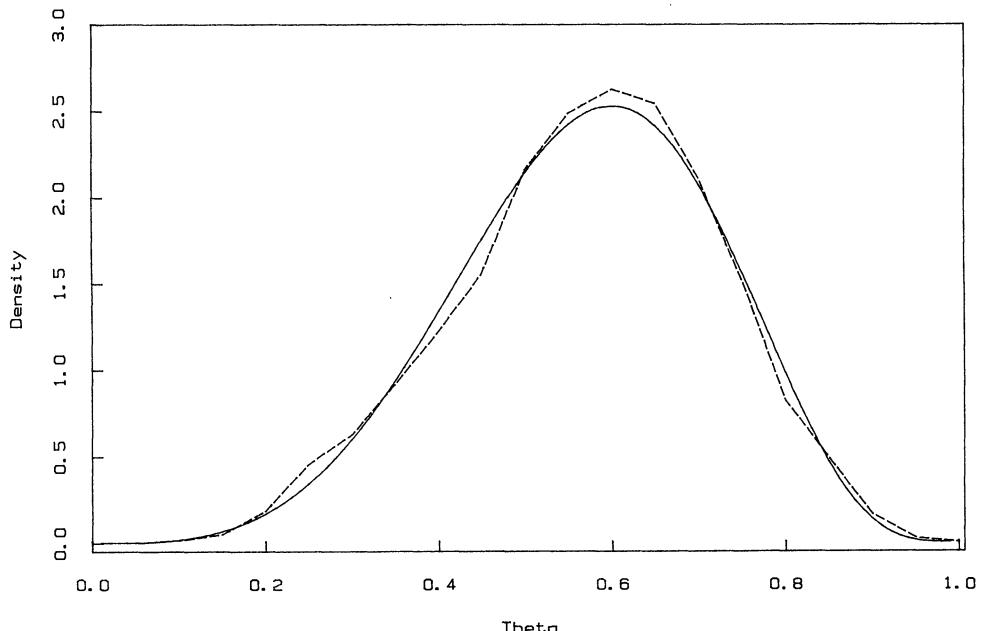


Figure 11. The Posterior Density of  $\theta$ . The dashed and solid lines represent the estimated and true posterior, respectively.

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  $m \rightarrow \infty$ ) of the posterior mean of  $\theta$ , 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,  $\theta(i)$  ( $i = 1, 2, \dots$ ) 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.