Bayesian Analysis of Multivariate Probit Models

SIDDHARTHA CHIB* EDWARD GREENBERG

July 1996
First version: September 1995

Abstract

This paper provides a unified simulation-based Bayesian and non-Bayesian analysis of correlated binary data using the multivariate probit model. The posterior distribution is simulated by Markov chain Monte Carlo methods, and maximum likelihood estimates are obtained by a Monte Carlo version of the E-M algorithm. Computation of Bayes factors from the simulation output is also considered. The methods are applied to a bivariate data set, to a 534-subject, four-year longitudinal data set from the Six Cities study of the health effects of air pollution, and to a seven-year data set on the labor supply of married women from the Panel Survey of Income Dynamics.

Keywords : Bayes factors; correlated binary data; Gibbs sampling; marginal likelihood; Markov chain Monte Carlo; Metropolis-Hastings algorithm.

1 Introduction

Correlated binary data arise in settings ranging from multivariate measurements on a random cross-section of subjects to repeated measurements on a sample of subjects across time. A central issue in the analysis of such data is model formulation. One strategy, outlined by Carey, Zeger, and Diggle (1993) and Glonek and McCullagh (1995), relies on the generalization of the binary logistic model to multivariate outcomes in conjunction with a particular parameterized representation for the correlations. Another strategy, discussed by Ashford and Sowden (1970) and Amemiya (1985), generalizes the binary probit model. The resulting multivariate probit model is described in terms of a correlated Gaussian distribution for underlying latent variables that are manifested as discrete variables through a threshold specification. Despite this connection to the Gaussian distribution (which allows

Address for correspondence: John M. Olin School of Business, Washington University, One Brookings Drive, St. Louis, MO 63130; e-mail: chib@simon.wustl.edu. Siddhartha Chib is Professor, John M. Olin School of Business and Edward Greenberg is Professor, Department of Economics, Washington University.
We acknowledge the helpful comments on the paper by the editor, associate editor and the referees.

for flexible modeling of the correlation structure and straightforward interpretation of the parameters), the model is not commonly used, mainly because its likelihood function is difficult to evaluate, except under simplifying assumptions [Ochi and Prentice (1984)]. Thus, few applications of the model have appeared and much of the potential of the model has not been realized.

The purpose of this paper is to provide a unified simulation-based inference methodology for overcoming the problems in fitting multivariate probit models. We discuss various aspects of the inference problem including simulation of the posterior distribution, calculation of maximum likelihood estimates and the computation of Bayes factors from the simulation output. The approach makes extensive use of recent developments both in Markov chain Monte Carlo methods [Gelfand and Smith (1990), Roberts and Smith (1993), Tierney (1994), Chib and Greenberg (1995)] and in the Bayesian analysis of binary and polychotomous data [Albert and Chib (1993)]. Two important technical advances made in this work may be highlighted. First, the paper provides an approach for sampling the posterior distribution of the correlation matrix. The same approach can be used in other problems with a restricted covariance matrix. Second, we extend Chib's (1995) marginal likelihood estimation procedure to a problem where some of the full conditional densities in the Markov chain Monte Carlo simulation do not have known normalizing constants.

The paper proceeds as follows. In Section 2 we summarize the model and in Section 3 we consider the sampling of the posterior distribution and the computation of the marginal likelihood. The computation of maximum likelihood estimates is discussed in Section 4. These estimates are obtained by utilizing a Monte Carlo version of the EM algorithm [Wei and Tanner (1990) and Meng and Rubin (1993)]. The E-step in this approach is implemented by Monte Carlo, while the M-step is conducted in two sub-steps; latent data are re-simulated after the first conditional maximization. Section 5 presents three real data applications, and Section 6 contains conclusions.

2 The multivariate probit model

Let Y_{ij} denote a binary 1-0 response on the ith observation unit ($1 \leq i \leq n$) and jth variable, and let $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ denote the collection of responses on all J variables.

According to the multivariate probit model, the probability that $Y_i = y_i$ (conditioned on parameters γ, Ω, and a set of covariates x_{ij}) is given by

$$\Pr(Y_i = y_i|\gamma,\Omega) \equiv \Pr(y_i|\gamma,\Omega) = \int_{A_J} \cdots \int_{A_1} \phi_J(t|0,\Omega)\, dt, \tag{1}$$

where $\phi_J(t|0,\Omega)$ is the density of a J-variate normal distribution with mean vector 0 and covariance matrix $\Omega = \{\omega_{jk}\}$, A_j is the interval

$$A_j = \begin{cases} (-\infty, x_{ij}'\gamma_j) & \text{if} \quad y_{ij} = 1 \\ (x_{ij}'\gamma_j, \infty) & \text{if} \quad y_{ij} = 0, \end{cases}$$

$\gamma_j \in R^{k_j}$ is an unknown parameter vector, and $\gamma' = (\gamma_1', \ldots, \gamma_J') \in R^k$, $k = \sum k_j$.

For our purposes a more convenient formulation of the multivariate probit model is in terms of Gaussian latent variables. Let $W_i = (w_{i1}, \ldots, w_{iJ})'$ have the distribution

$$W_i \sim N_J(X_i\gamma, \Omega), \tag{2}$$

where $X_i = \text{diag}(x_{i1}', \ldots, x_{iJ}')$ is a $J \times k$ covariate matrix, and let Y_{ij} be 1 or 0 according to the sign of w_{ij}, i.e.,

$$Y_{ij} = \begin{cases} 1 & \text{if} \quad w_{ij} > 0 \\ 0 & \text{if} \quad w_{ij} \leq 0. \end{cases} \tag{3}$$

It is then easy to show that the probability that W_i lies in the subspace of R^J generated by $\cap_{j=1}^J A_j$ equals the probability in (1).

It follows from the latent variable formulation that the parameters γ and Ω are not likelihood identified. To see this, pre-multiply W_i by a diagonal matrix C with positive elements and observe that the outcomes Y_{ij} are left unaffected. A set of identified parameters is defined by letting $C = \text{diag}\{\omega_{11}^{-1/2}, \ldots, \omega_{JJ}^{-1/2}\}$ and defining

$$\beta_j = \omega_{jj}^{-1/2}\gamma_j \quad \text{and} \quad \Sigma = C\Omega C',$$

whence $\Pr(y_i|\gamma,\Omega)$ is equal to $\Pr(y_i|\beta,\Sigma)$. The parameters of the identified model thus consist of β and Σ , a correlation matrix, with the $p = J(J-1)/2$ free parameters denoted by $\sigma \equiv (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{J-1,J})$. The corresponding latent variable model is

$$Z_i \sim N_J(X_i\beta, \Sigma), \quad y_{ij} = I[z_{ij} > 0], \tag{4}$$

where $Z_i = CW_i$. This representation forms the basis of our sampling method.

3 Posterior analysis

Suppose that we have a random sample of nJ observations $y = (y_1, \ldots, y_n)$ and a prior density $\pi(\beta, \sigma)$ on the parameters of a given multivariate probit model. Then, the posterior density is

$$\pi(\beta, \sigma|y) \propto \Pr(y|\beta, \Sigma)\pi(\beta, \sigma), \quad \beta \in \Re^k, \quad \sigma \in \mathbf{A}, \tag{5}$$

where

$$\Pr(y|\beta, \Sigma) = \prod_{i=1}^{n} \Pr(y_i|\beta, \Sigma)$$

is the likelihood function and $\mathbf{A}$ is a convex solid body in the hypercube $[-1, 1]^p$ that leads to a proper correlation matrix [see Rousseeuw and Molenberghs (1994) for more on the shape of correlation matrices]. Assume prior independence between β and σ, and let

$$\pi(\beta) = \phi_k(\beta|\beta_0, B_0^{-1}) \quad \text{and} \quad \pi(\sigma) \propto \phi_{\mathrm{p}}(\sigma|\sigma_0, \mathrm{G}_0^{-1}), \quad \sigma \in \mathbf{A}, \tag{6}$$

where ϕ_s denotes the density of a s - variate normal distribution (which in the case of σ is truncated to the region $\mathbf{A}$) and the hyperparameters $(\beta_0, B_0, \sigma_0, G_0)$ are chosen to reflect the available prior information. The location of the prior information is controlled by the vectors β_0 and σ_0 and the strength by the precision matrices B_0 and G_0.

Our first goal is to use Markov chain Monte Carlo methods to summarize the posterior distribution. It is easy to see that $\pi(\beta, \sigma|y)$ cannot be simulated, even with a Metropolis type algorithm, due to the intractability of $\Pr(y|\beta, \Sigma)$. We therefore follow Albert and Chib (1993) and augment the parameter space to include $Z = (Z_1, \ldots, Z_n)$. Let $f(Z|\beta, \Sigma)$ denote the density of Z conditioned on the parameters. Then, under the assumptions in (4), we have

$$f(Z|\beta, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Z_i - X_i\beta)'\Sigma^{-1}(Z_i - X_i\beta)\right) I[\sigma \in \mathbf{A}]\,.$$

From this, we consider a Gibbs sampling scheme based on the following full conditional distributions

$$\begin{aligned}
[Z|y, \beta, \Sigma] &\stackrel{d}{=} \prod_{i=1}^{n}[Z_i|y_i, \beta, \Sigma]\,, \\
[\beta|y, Z, \Sigma] &\stackrel{d}{=} [\beta|Z, \Sigma]\,, \text{ and} \\
[\sigma|\mathrm{y}, \mathrm{Z}, \beta] &\stackrel{d}{=} [\sigma|Z, \beta]\,,
\end{aligned}$$

where " $\stackrel{d}{=}$ " denotes equality in distribution. As we show below, only the conditional density of σ cannot be simulated by standard means. This complication is unavoidable because the alternative of sampling $\pi(\{W_i\}, \gamma, \Omega|y)$ and then transforming the output to obtain a sample of the identified parameters (as in McCulloch and Rossi (1994)) cannot be recommended for several reasons. First, this approach requires a carefully balanced weak (proper) prior distribution on the *unidentified parameters*—too weak of a prior leads to a flat posterior and convergence problems with respect to the Markov chain Monte Carlo simulations, whereas a strong prior completely determines the posterior. Second, it does not work when Σ is patterned or restricted, as in some of the examples below. Finally, the dependence on weak priors is inconsistent with the computation of Bayes factors, giving rise to problems akin to those described by the Lindley paradox, especially if the model dimensions are quite different.

3.1 Posterior simulations

We now explain how each of the full conditional distributions can be sampled, beginning with the full conditional distribution of the latent data. Conditional on the observed outcomes and the parameters, the distribution of Z_i is truncated to a subset of R^J, where the region of truncation is determined by y_i [see Albert and Chib (1993) for further details]. For example, if $J = 2$ and $y_i = (1, 1)'$, then Z_i lies in the positive orthant. More generally, the full conditional distribution of the latent data is truncated multivariate normal

$$Z_i|(y_i, \beta, \Sigma) \propto N_J(X_i\beta, \Sigma) \prod_{j=1}^{J} \left(I(0 < z_{ij})I(y_{ij} = 1) + I(z_{ij} \leq 0)I(y_{ij} = 0)\right), \qquad (7)$$

where $I[A]$ is the indicator function of the event A. Following Geweke (1991), this distribution can be simulated by composing a cycle of J Gibbs steps through the components of Z_i; i.e., z_{ij} is simulated from $z_{ij}|(y_{ij}, z_{ik}\,(k \neq j), \beta, \Sigma)$. This distribution is truncated univariate normal, truncated to $(0, \infty)$ if $y_{ij} = 1$ and to $(-\infty, 0]$ if $y_{ij} = 0$. The parameters of the untruncated normal distribution $z_{ij}|(z_{ik}\,(k \neq j), \beta, \Sigma)$ are obtained from the usual formulas and are omitted.

For the full conditional density of β given values of Z and Σ, we combine the normal density of Z with the multivariate normal prior density of β to obtain by standard Bayesian

calculations that

$$\beta|(Z,\Sigma) \sim N_k(\hat{\beta}, B^{-1}), \tag{8}$$

where $\hat{\beta} = B^{-1}(B_0\beta_0 + \sum_{i=1}^n X_i'\Sigma^{-1}Z_i)$ and $B = B_0 + \sum_{i=1}^n X_i'\Sigma^{-1}X_i$. Thus, the simulation of β is straightforward.

Next, we consider the full conditional density of the unique elements of Σ

$$\begin{aligned}\pi(\sigma|Z,\beta) &\propto \pi(\sigma)\, f(Z|\beta,\Sigma) \\ &\propto \pi(\sigma)\,|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\mathrm{Z}-\Delta)'\Sigma^{-1}(\mathrm{Z}-\Delta)\right) I[\sigma \in \mathbf{A}],\end{aligned}$$

where $Z = (Z_1, Z_2, ..., Z_n)$ and $\Delta = (X_1\beta, \ldots, X_n\beta)$ are both $J \times n$ matrices. As the analysis of this density (and the search for suitable bounds and dominating functions) is difficult, we make use of the Metropolis-Hastings algorithm. Recall that for a given target density $f(\sigma)$, the Metropolis-Hastings algorithm consists of a proposal density $q(\sigma,\sigma')$, which supplies candidate values σ' given the current value σ, and a probability of move

$$\alpha(\sigma,\sigma') = \min\left(\frac{f(\sigma')q(\sigma',\sigma)}{f(\sigma)q(\sigma,\sigma')}, 1\right),$$

to determine whether the proposal value is accepted [Hastings (1970), Chib and Greenberg (1995)]. The function $f(\sigma) = \pi(\sigma)\, f(Z|\beta,\Sigma)I[\sigma \in \mathbf{A}]$ is easily evaluated, and the main question concerns the formulation of a suitable proposal density. Note that the proposal density need not enforce the positive definiteness constraint, because that constraint is part of $f(\sigma)$. Thus, if Σ' is not positive definite, the conditional posterior is zero, and the proposal value is rejected with certainty. It may also be mentioned that when the dimension of Σ is large (as in our third example in Section 5 below) it is best to partition σ into blocks and to apply the Metropolis-Hastings algorithm in sequence, cycling through the various blocks.

The simplest proposal density is described by the *random walk chain*

$$\sigma' = \sigma + h,$$

where σ' is the candidate value, σ is the current value, and h is a zero mean increment vector. It is convenient to assume that h follows a symmetric distribution, such as the normal, which leads to a cancellation of the q functions in the computation of $\alpha(\sigma,\sigma')$. The variance of the increment may be set to a multiple of either $1/n$, which is the large-sample variance of the marginal posterior of a correlation coefficient, or λ, the smallest

characteristic root of Σ [see Marsaglia and Olkin (1984) for an explanation]. This proposal density is generally effective for small p (less than five).

A more general procedure is to tailor the proposal density to $f(\sigma)$. Then, the form of the proposal generating process is

$$\begin{aligned} \sigma' &= \mu^* + h \\ \mu^* &= \mu + B(\sigma - \mu), \end{aligned}$$

where μ is a vector and $B : p \times p$ a diagonal matrix. Tailoring is achieved by letting μ be the approximate mode of the function $f(\sigma)$ and specifying the variance of the increment h to be similar to the curvature of $f(\sigma)$ around μ. The matrix B is taken either as equal to zero or equal to minus the identity matrix of order p. We refer to the former case as a *tailored independence chain* and the latter as a *tailored reflection chain* because the proposal values are reflected around μ. The reflection chain provides a way for the chain to make large moves, in this case to the region on the other side of the point μ where the density is likely to be about as high as at the current point. The simplest way to find μ is by a few, perhaps two, iterative Newton-Raphson steps, initialized at the mode from the previous round. We estimate the curvature of the target at μ by the inverse of the negative Hessian matrix $V = -\left(\frac{\partial^2 f}{\partial \sigma \partial \sigma'}\right)^{-1}$. With these ingredients, the proposal density is

$$q(\sigma, \sigma') = f_{mvt}(\mu^*, \tau^2 V, \nu), \tag{9}$$

a multivariate-t density with mean vector μ^*, dispersion matrix $\tau^2 V$, and ν degrees of freedom. The tuning parameter τ^2 is adjusted by experimentation, and ν is specified arbitrarily at 10 (or some similar value). From our experience, this version of the Metropolis-Hastings algorithm, although computationally more demanding, leads to lower serial correlation than the random walk proposal density.

The simulation of these full conditional distributions in fixed or random order completes one cycle of the Markov chain Monte Carlo algorithm. To generate a sample of draws from the posterior distribution, this cycle is run a large number of times. All values beyond those in an initial transient stage are collected and used to summarize the posterior density. For example, the posterior mean is estimated as the average of the simulated values, and posterior credibility intervals are estimated from the sample percentiles.

3.2 Computation of marginal likelihood

We next consider the calculation of marginal likelihoods and Bayes factors (ratios of marginal likelihoods) for competing multivariate probit models that may be obtained by restricting the covariate or correlation structure. For example, one might be interested in the restriction that $\beta_j = \beta$ for all j. One might also be interested in imposing restrictions on Σ, for example, by specifying that Σ is in the equi-correlated form $(1-\rho)I_J + \rho 1_J 1_J'$, where $|\rho| < 1$ [Ochi and Prentice (1984)]. Other models arise if the index j represents time (as in a panel data setting), when Σ may be specified to reflect the assumption of serially correlated errors or the assumption of 1-dependence. Finally, it may be of interest to determine whether the responses are independent [see Kiefer (1982)].

By definition the marginal likelihood of M_k is given by

$$m(y|M_k) = \int \Pr(y|M_k, \beta, \Sigma)\, \pi(\beta, \sigma|M_k)\, d\beta\, d\sigma. \tag{10}$$

This integral cannot be estimated by the Laplace method or by importance sampling [Kass and Raftery (1995)], because these require the repeated evaluation of the likelihood function $\Pr(y|M_k, \beta, \Sigma)$. We therefore discuss an approach based on Chib (1995) that relies on an alternative expression for the marginal likelihood:

$$m(y|M_k) = \frac{\Pr(y|M_k, \beta, \Sigma)\pi(\beta, \sigma|M_k)}{\pi(\beta, \sigma|M_k, y)}, \tag{11}$$

where the numerator is the product of the sampling density and the prior, with all integrating constants included, and the denominator is the posterior density. This expression arises from the formula for the posterior density and holds for any value of (β, σ). We refer to it as the basic marginal likelihood identity. The key point is that an estimate of the posterior density at a point (β^*, σ^*) delivers an estimate of the marginal likelihood. On the computationally convenient log scale

$$\begin{aligned} \ln \hat{m}(y|M_k) &= \ln \Pr(y|M_k, \beta^*, \Sigma^*) + \ln \pi(\beta^*|M_k) + \ln \pi(\sigma^*|M_k) \\ &\quad - \ln \hat{\pi}(\beta^*|M_k, y, \sigma^*) - \ln \hat{\pi}(\sigma^*|M_k, y). \end{aligned} \tag{12}$$

A considerable virtue of this estimate is that it requires only *one* evaluation of the sampling density. Although this expression may be evaluated at any point (β^*, Σ^*) in the parameter

space, it is important that it be evaluated at a high density point, such as the posterior mean. Such a choice leads to a more accurate estimate.

We now discuss the estimation of the terms in (12) that are not available through direct computation, starting with $\pi(\sigma^*) = \phi_p(\sigma|\sigma_0, G_0^{-1})/\Pr(\sigma \in \mathrm{A})$, where we have suppressed the dependence on M_k. The issue is how to estimate the normalizing constant, $\Pr(\sigma \in \mathrm{A})$. Since analytical evaluation is not feasible, we generate a large number of observations from $\phi_p(\sigma|\sigma_0, G_0^{-1})$ and find the proportion that satisfy the positive definiteness constraint. This proportion is the Monte Carlo estimate of $\Pr(\sigma \in \mathrm{A})$.

The next goal is to specify appropriate estimators for the conditional posterior ordinate

$$\pi(\beta^*|y, \Sigma^*) = \int \pi(\beta^*|y, \Sigma^*, Z)\, p(Z|y, \Sigma^*)\, dZ$$

and the marginal ordinate

$$\pi(\sigma^*|y) = \int \pi(\sigma^*|Z, \beta)\, p(Z, \beta|y)\, d\beta\, dZ.$$

For the former, note that if $Z^{(g)} \sim (Z|y, \Sigma^*)$, then

$$\hat{\pi}(\beta^*|y, \Sigma^*) = G^{-1} \sum_{g=1}^{G} \pi(\beta^*|y, Z^{(g)}, \Sigma^*) \tag{13}$$

is a simulation-consistent estimate of $\pi(\beta^*|y, \Sigma^*)$, where $\pi(\beta^*|y, Z, \Sigma^*) = \pi(\beta^*|Z, \Sigma^*)$ is the multivariate normal density in (8) with β set equal to β^* and Σ to Σ^*. Following Chib (1995), draws $Z^{(g)} \sim (Z|y, \Sigma^*)$ may be obtained by fixing Σ at the value Σ^* and applying the Markov chain Monte Carlo sampling algorithm for G iterations to the (reduced) full conditional distributions

$$(Z_1|y_1, \beta, \Sigma^*) \times \ldots \times (Z_n|y_n, \beta, \Sigma^*) \quad \text{and} \quad \beta|(\mathrm{Z_1}, \ldots, \mathrm{Z_n}, \Sigma^*).$$

For the marginal ordinate a similar averaging estimator would have been possible if the normalizing constant of $\pi(\sigma^*|Z, \beta)$ were known. As an alternative technique we rely on kernel smoothing. The resulting estimator takes the form

$$\hat{\pi}(\sigma^*|y) = G^{-1} \sum_{g=1}^{G} \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{\sigma_j^* - \sigma_j^{(g)}}{h_j}\right),$$

where $K(x)$ is a univariate kernel density and h_j is the bandwidth parameter. In the examples below we use a Gaussian kernel and let $h_j = s_j G^{-1/(p+4)}$, where s_j is the standard deviation of $\{\sigma_j^{(g)}\}$ [see Silverman (1986) for further details]. From numerical evaluation we have found that the accuracy of this estimate can be improved by thinning the sample, say by retaining every fifth iterate, before application of the formula above.

It should be noted that the kernel-based estimate of the marginal ordinate is likely to inaccurate when σ is high dimensional. This problem can be overcome (at the cost of additional computations that require no new programming) by partitioning σ into several low-dimensional blocks and applying the reduced Markov chain Monte Carlo run procedure to each of the blocks. We illustrate with two blocks: $\sigma = (\sigma_1, \sigma_2)$, where σ_i contains p_i components and $\sum p_i = p$. The identity

$$\pi(\sigma^*|y) = \pi_1(\sigma_1^*|y)\, \pi_2(\sigma_2^*|y, \sigma_1^*)$$

permits us to estimate each of the two terms by kernel smoothing after generating samples from the appropriate distributions. For the first term, the values of $\sigma_1^{(g)}$ $(g = 1, \ldots, G)$ are available from the Markov chain Monte Carlo sampler already run leading to an estimate analogous to that above. The next term may be approximated by generating a sample from the full conditional distribution of σ_2 with σ_1 fixed at σ_1^*; i.e., one generates a sample of G observations from the reduced Markov chain Monte Carlo sampler based on

$$[Z|y, \beta, \sigma_1^*, \sigma_2], \quad [\beta|y, Z, \sigma_1^*, \sigma_2], \quad [\sigma_2|y, Z, \beta, \sigma_1^*].$$

The $\sigma_2^{(g)}$ generated from this sampler are a sample from $[\sigma_2|y, \sigma_1^*]$. The value $\pi_2(\sigma_2^*|y, \sigma_1^*)$ can be estimated by kernel smoothing.

Finally, consider the computation of $\ln \Pr(y|\beta^*, \Sigma^*) = \sum_{i=1}^n \ln \Pr(y_i|\beta^*, \Sigma^*)$. A simple idea is to use a Monte Carlo accept-reject procedure by iterating on the following steps for $g = 1, 2, \ldots, M$:

Step 1: Simulate $Z_i^{(g)}$ from $N_J(X_i\beta^*, \Sigma^*)$;

Step 2: Calculate $\Pr(y_i|Z_i^{(g)}, \beta^*, \Sigma^*)$.

The probability in Step 2 is 1 or 0 depending on whether $Z_i^{(g)}$ respects the constraints imposed by y_i. Then, from the law of large numbers,

$$M^{-1} \sum_{g=1}^{M} \Pr(y_i | Z_i^{(g)}, \beta^*, \Sigma^*) \to \Pr(y_i | \beta^*, \Sigma^*),$$

so that a simulation-consistent estimate of the sampling density is

$$\sum_{i=1}^{n} \ln \left[M^{-1} \sum_{g=1}^{M} \Pr(y_i | Z_i^{(g)}, \beta^*, \Sigma^*) \right].$$

For this method to be effective, M must be large, but ensuring this is relatively painless because the computation is done at only one point (β^*, Σ^*) and Step 1 requires only the generation of Gaussian samples.

4 Modal estimation

A by-product of the simulation of the latent data is an approach that yields the maximum likelihood estimates for the multivariate probit model (or the maximizer of the posterior) without computation of the likelihood function. This can be done by an adaptation of the Monte Carlo E-M algorithm (MCEM) that was proposed by Wei and Tanner (1990), which in turn is a stochastic modification of the original Dempster, Laird, and Rubin (1977) E-M algorithm.

Let $\theta = (\beta, \sigma)$ and suppose that, given the current value of the maximizer $\theta^{(t)}$, it is desired to evaluate the expectation or E-step of the E-M algorithm. This amounts to an evaluation of the integral

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \int_Z \log\left(f(y, Z|\theta)\right) d[Z|y, \theta^{(t)}] \\ &= \int_Z \log\left(f(Z|\theta)\right) d[Z|y, \theta^{(t)}], \end{aligned} \tag{14}$$

where the integral is w.r.t. the distribution of Z induced by (7), and the second line is a consequence of the fact that $f(y|Z, \theta)$ is degenerate, since knowing Z implies knowledge of y. The Q function cannot be evaluated analytically, but it can be estimated consistently by an ergodic Monte Carlo average. Given the current parameter value $\theta^{(t)}$, take a large number of draws of Z with the approach described above; the draws are denoted by $Z^{(j)}$,

$j = 1, \ldots, N$. Then the Q function is approximated by the ergodic average, rather than by the i.i.d. average in Wei and Tanner (1990),

$$\begin{aligned} \hat{Q}(\theta, \theta^{(t)}) &= N^{-1} \sum_j \log \left(f(Z^{(j)}|\theta) \right) \\ &= -\frac{1}{2} \log |\Sigma| - N^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} (Z_i^{(j)} - X_i\beta)' \Sigma^{-1} (Z_i^{(j)} - X_i\beta). \end{aligned} \tag{15}$$

In the M-step, the $\hat{Q}$ function is maximized over θ to obtain the new parameter $\theta^{(t+1)}$. The algorithm is terminated once the difference $||\theta^{(t+1)} - \theta^{(t)}||$ is negligible.

We now follow Meng and Rubin (1993) and complete the M-step through a sequence of two conditional maximizations—the maximization over β given Σ and the maximization over Σ given β. This simplifies the update of θ and parallels the blocking adopted in the Bayesian simulation presented above [a similar conditional maximization step is adopted by Natarajan, McCulloch, and Kiefer (1995) in a different context]. Specifically, on setting the derivative w.r.t. β equal to zero, we find that the update of β is given by

$$\beta^{(t+1)} = (\sum_{i=1}^{n} X_i' \Sigma^{-1} X_i)^{-1} (\sum_{i=1}^{n} X_i' \Sigma^{-1} \bar{Z}_i),$$

where $\bar{Z}_i = N^{-1} \sum_{j=1}^{N} Z_i^{(j)}$ is the average of Z_i over the N draws. The update of σ is obtained by replacing β by $\beta^{(t+1)}$ in the $\hat{Q}$ function and maximizing over σ using a Newton-Raphson type routine. Although not necessary, it is preferable for efficiency considerations to re-draw the Z values from the distribution $Z|y, \beta^{(t+1)}, \Sigma$ and re-compute the $\hat{Q}$ function before the second maximization is attempted. The update for σ is thus obtained by maximizing the function

$$-\frac{1}{2} \log |\Sigma| - N^{-1} \sum_{g=1}^{N} \sum_{i=1}^{n} (Z_i^{(g)} - X_i\beta^{(t+1)})' \Sigma^{-1} (Z_i^{(g)} - X_i\beta^{(t+1)}),$$

where $Z^{(g)}$ are the newly drawn latent values.

As suggested by Wei and Tanner (1990) and Chib (1993), in producing the iterate sequence $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)} \ldots\}$ with the above strategy it is best to begin with a small value of N and increase the number of replications of Z as the maximizer is approached. Given the modal value $\hat{\theta}$, the standard errors of the estimate can be obtained by the formula

of Louis (1982). Specifically, the observed information matrix (the negative of the second derivative matrix of the likelihood function) is given by

$$-E\left[\frac{\partial^2 \log f(Z|\theta)}{\partial\theta\partial\theta'}\right] - \mathrm{Var}\left[\frac{\partial \log f(Z|\theta)}{\partial\theta}\right],$$

where the expectation and variance are w.r.t. the distribution $Z|y,\hat{\theta}$. Each of these terms can be estimated by taking an additional M draws $\{Z^{(1)},\ldots,Z^{(M)}\}$ from $Z|y,\hat{\theta}$. Standard errors are then obtained as the square root of the diagonal elements of the inverse of the estimated information matrix.

5 Applications

5.1 Bivariate probit for voter behavior

Our first application of the methods is to survey data of the voting behavior of 95 residents of Troy, Michigan, in which the first decision (Y_{i1}) is whether to send at least one child to public school and the second (Y_{i2}) is whether to vote in favor of a school budget. The objective of the study is to model the two quantal responses as a function of covariates, allowing for correlation in the responses. As in Greene (1993), let the covariates in x_{i1} be a constant, the natural logarithm of annual household income in dollars (INC), and the natural logarithm of property taxes paid per year in dollars (TAX); and those in x_{i2} be a constant, INC, TAX, and the number of years (YRS) the resident has been living in Troy. The data collected in this survey are reproduced in Table 1.

We fit two models to this data set. The first, denoted M_1, is the bivariate probit model in which the marginal probabilities for the ith subject are given by

$$\Pr(y_{ij}=1|\beta,\sigma) = \Phi(x_{ij}'\beta_j),$$

and the joint probabilities are given through the cdf of the bivariate normal with correlation matrix equal to

$$\Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}.$$

Model M_1 thus contains 7 unknown regression parameters and 1 unknown correlation parameter. The second model, denoted M_2, is an independent probit model in which $\sigma_{12}=0$.

The modal (maximum likelihood) estimates in this instance can be computed by directly maximizing the likelihood function as discussed in Greene (1993); the MCEM algorithm is not necessary. These estimates are reported in Table 2. For model M_1, the posterior distributions of the parameters are obtained by applying the Markov chain Monte Carlo algorithm described in Section 3 for 6,000 cycles beyond 500 burn-in iterations. The prior distribution of β is multivariate normal with a mean vector of 0 and a variance matrix of 100 times the identity matrix and that of σ_{12} is proportional to a univariate normal with a mean of 0 and variance of 0.5. In terms of the notation defined in equation (6) we have set $\beta_0 = 0$, $B_0 = 10^{-2} I_7$, $\sigma_0 = 0$, and $G_0^{-1} = .5$.

We use the random walk proposal density in the Metropolis-Hastings step and let the increment random be univariate normal with standard deviation equal to $4\sqrt{1/n}$. This results in an acceptance rate of about 0.5. Our results on the posterior distribution are summarized in Table 2 (the results for the independent probit obtained via the algorithm of Albert and Chib (1993) are similar). The table reports the ML estimates, the prior moments, the posterior means and standard deviations, the numerical standard errors computed by the method of batch means, and the 2.5th and 97.5th percentiles of the posterior distribution.

The posterior distribution of σ_{12} is spread out, and its 95% credibility interval includes 0, which is evidence for the independent probit model. To formally assess the evidence for M_1 and M_2, we calculate the marginal likelihoods by running the sampler for an additional 6,000 iterations. For M_1, β^* and σ_{12}^* are specified as their respective posterior means, and for M_2, β^* is the posterior mean of β. The likelihood function is available analytically, and its value based on 10,000 simulated draws agrees with the exact expression up to the second decimal place on the log-scale. Similarly, the normalizing constant of the prior of σ_{12} is available analytically, but it is estimated as described in Section 3 above with 5,000 draws from the untruncated normal. The log marginal likelihood for M_1 is -126.31 and -126.30 for M_2. The data thus do not provide support for the correlated probit over the independent probit as evidenced by the posterior probabilities (assuming prior probabilities of 1/2 for the respective models) and the Bayes factors reported in Table 3.

5.2 Six Cities study

The second example is based on a subset of data from the Six Cities study, a longitudinal study on the health effects of air pollution, which has been analyzed by Fitzmaurice and Laird (1993) and Glonek and McCullagh (1995) using a multivariate logit model. The data, which are reproduced in Table 4, contain repeated binary measure on the wheezing status (1 = yes; 0 = no) for each of 537 children from Stuebenville, Ohio at ages 7, 8, 9, and 10 years. The objective of the study is to model the probability of wheeze status over time as a function of a binary indicator variable representing the mother's smoking habit during the first year of the study and the age of the child.

Interpreting age as category j, we fit three models to these data: the full multivariate probit model (M_1), the equi-correlated model (M_2), and the independent probit model (M_2). In each model the marginal probability of response is specified as

$$\Pr(y_{ij} = 1|\beta, \Sigma) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where x_{i1} is the age of the child (centered at 9 years), x_{i2} is a binary indicator representing the mother's smoking habit (1 = yes, 0 = no), and x_{i3} is an interaction between smoking habit and age. Note that, unlike the previous example, the regression parameter is constrained to be constant across j. In addition to the four regression parameters in each model, there are six unknown correlation parameters in M_1, one unknown correlation parameter in M_2, and zero unknown correlation parameters in M_2.

For all three models the prior on β in represented by the hyperparameters

$$\beta_0 = 0, B_0 = 10^{-1} I_4,$$

and that of σ in M_1 and M_2 by

$$\sigma_0 = 0, G_0^{-1} = .5 I_6 \quad \text{and} \quad \sigma_0 = 0,\ \mathrm{G}_0^{-1} = .5,$$

respectively, where (here and below) with abuse of notation we use the same symbols for the hyperparameters across the different sized models. Posterior sampling of the correlation parameters of model M_1 is by the tailored independence method applied in one block to all six unknown parameters of Σ. The parameters μ and V of the target function $f(\sigma)$

are obtained by the Newton-Raphson method, and $\tau = 1.5$. The same approach is used to sample the single parameter of Σ in model M_2 with $\tau = 4.0$. The sampler was run for $G = 10,000$ cycles beyond a transient stage of 500 iterations. The Metropolis-Hastings acceptance rate was about 35% for the full model and about 40% for the equi-correlated model.

To compute the modal estimates for M_1 and M_2, the MCEM algorithm described above is tuned as follows: For the first ten updates of θ, the Q function is estimated from $N = 10$ samples of the latent data; for the final ten iterations $N = 200$. The algorithm was stopped at iteration 40 when convergence was achieved for each parameter up to at least the first two decimal places.

Results of the simulation are summarized in Table 5. First, note that the MLE and Bayes estimates of β are very insensitive to the specification of the covariance structure. Second, the MLE values differ slightly from the posterior mean, an indication of some asymmetry in the posterior distributions. Third, the standard errors (s.e.) of the MLE are generally smaller than the corresponding posterior standard deviations. Fourth, there is little support for M_2 because the posterior distribution of Σ in both M_1 and M_2 is concentrated away from zero. More formal measures of support for the models are provided by the estimates of the marginal likelihoods. For both M_1 and M_2 the reduced Markov chain Monte Carlo sampler for β is run for 10,000 iterations, the marginal posterior density of σ is estimated in one block by kernel smoothing, and the prior ordinate of the truncated normal prior of σ at the point σ^* (the posterior mean) is estimated from 10,000 simulations as discussed in Section 3. Interestingly, the marginal posterior ordinate of σ at σ^* for model M_1 changed only slightly when it was estimated as $\pi(\sigma_1^*, \sigma_2^*|y) = \pi(\sigma_1^*|y)\pi(\sigma_2^*|y, \sigma_1^*)$ with each reduced block of size three. This leads us to the conclusion that kernel smoothing can give accurate estimates of the posterior ordinate at a high density point for five or six dimensions if the Markov chain Monte Carlo sample size is reasonably large (as here).

The results show that the marginal likelihood of the independent model M_2 is the smallest of the three models and that of M_2 the largest. The log marginal likelihood increased by about .30 when the prior on σ was specified with $G_0^{-1} = 1$. This is evidence that the results are not unduly sensitive to the prior specification. In terms of the Bayes factors,

we find $B_{1,2} = 1.77 \times 10^{-4}$, $B_{1,3} \approx 10^{41}$, and $B_{2,3} \approx 10^{47}$. Unless the prior probabilities for the various models put virtually zero weight on M_1 and M_2, this is decisive evidence against independence in favor of either alternative and decisive evidence in favor of the equi-correlated model.

5.3 Labor Force Participation

The final illustration is a model of the labor force participation decision of married women in the age bracket 25–62. The data from the Panel Survey of Income Dynamics of the University of Michigan consist of a sample of 520 households over the seven-year span 1976–1982. Following Avery, Hansen, and Hotz (1983), where similar data are analyzed by the method of moments, the covariates are (1) a constant, (2) wife's education in number of grades completed, and (3) total family income excluding wife's earnings (in thousands of dollars). Means and standard deviations for the second and third covariates are 12.64 (2.34) and 22.41 (18.21), respectively (the standard deviations are in parentheses).

We consider two multivariate probit models for this data set. In M_1 the correlation matrix is fully unrestricted with 21 unknown parameters. In M_2 the correlation matrix is in equi-correlated form. In both models we let β_j be constant across j and represent our prior distribution through the hyperparameters $\beta_0 = 0$, $B_0 = 10^{-1} I_3$. In the unrestricted model, the prior on σ is represented by $\sigma_0 = 0$ (a 21 vector of zeros) and $G_0^{-1} = 0.50 I_{21}$; for the restricted model it is represented by $\sigma_0 = 0$ (a scalar) and $G_0^{-1} = 0.50$. The Markov chain Monte Carlo simulation algorithm is run for 10,000 cycles beyond a transient phase of 500 iterations. Because of the large dimension of σ in model M_1, the Metropolis-Hastings step is applied to $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ in four blocks, where σ_1, σ_2, and σ_3 each consist of six elements and σ_4 of three elements in a row-wise expansion of Σ. Thus, for example, $\sigma_1 = (\sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{15}, \sigma_{16}, \sigma_{17})$ and $\sigma_4 = (\sigma_{56}, \sigma_{57}, \sigma_{67})$. Proposal values for the correlation parameters in the four Metropolis-Hastings steps (within each cycle) are generated by the tailored independence chain, where μ is the approximate conditional mode of $f(\sigma_i)$ and V the negative of the inverse Hessian at convergence. The value of τ for the first three blocks of σ is 1.5 and that of the fourth block is 2.0. In the case of M_2, proposal values are also generated by the tailored independence chain, but with $\tau = 8$.

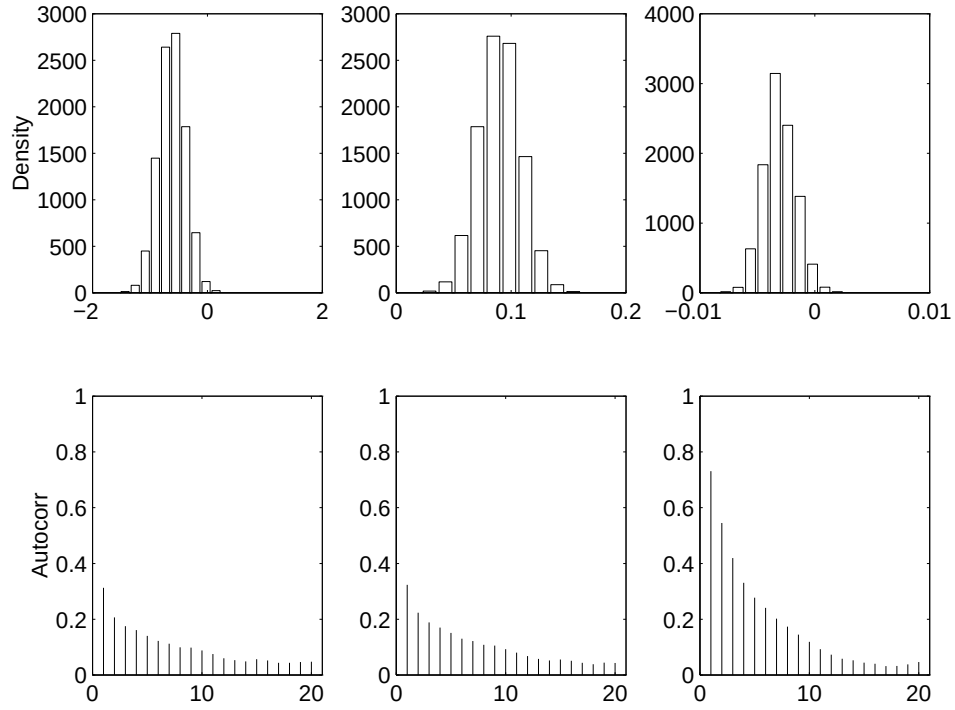

Figure 1: Posterior distributions of β and autocorrelation functions of sampled draws with PSID data and model M_1.

For the marginal likelihood calculation in M_1, the posterior density $\pi(\sigma^*|y)$ at σ^* (the posterior mean) is estimated from

$$\pi(\sigma_1^*|y)\pi(\sigma_2^*|y,\sigma_1^*)\pi(\sigma_3^*|y,\sigma_1^*,\sigma_2^*)\pi(\sigma_4^*|y,\sigma_1^*,\sigma_2^*,\sigma_3^*),$$

where each of the conditional ordinates is estimated by kernel smoothing the simulations from 10,000 values of σ_i generated from $\pi(\sigma_i|y,\sigma_1^*,\ldots,\sigma_{i-1}^*)$ in a reduced Markov chain Monte Carlo run. By breaking up σ in this manner we ensure that kernel smoothing remains accurate. Finally, we estimate the normalizing constant of the prior density of σ at σ^* from 10,000 draws and the likelihood function by iterating on Steps 1 and 2 for 50,000 cycles. The calculation of $m(y|M_2)$ is similar, but since only one parameter in σ is involved, $\pi(\sigma^*|y)$ is estimated directly in one pass by kernel smoothing.

The results from the simulation show that the posterior distributions of β from the two models are virtually identical. The posterior means and standard deviations of β in model M_1 are found to be $-.620$ (0.234), 0.090 (.018), and $-.003$ (.001); the modal estimates

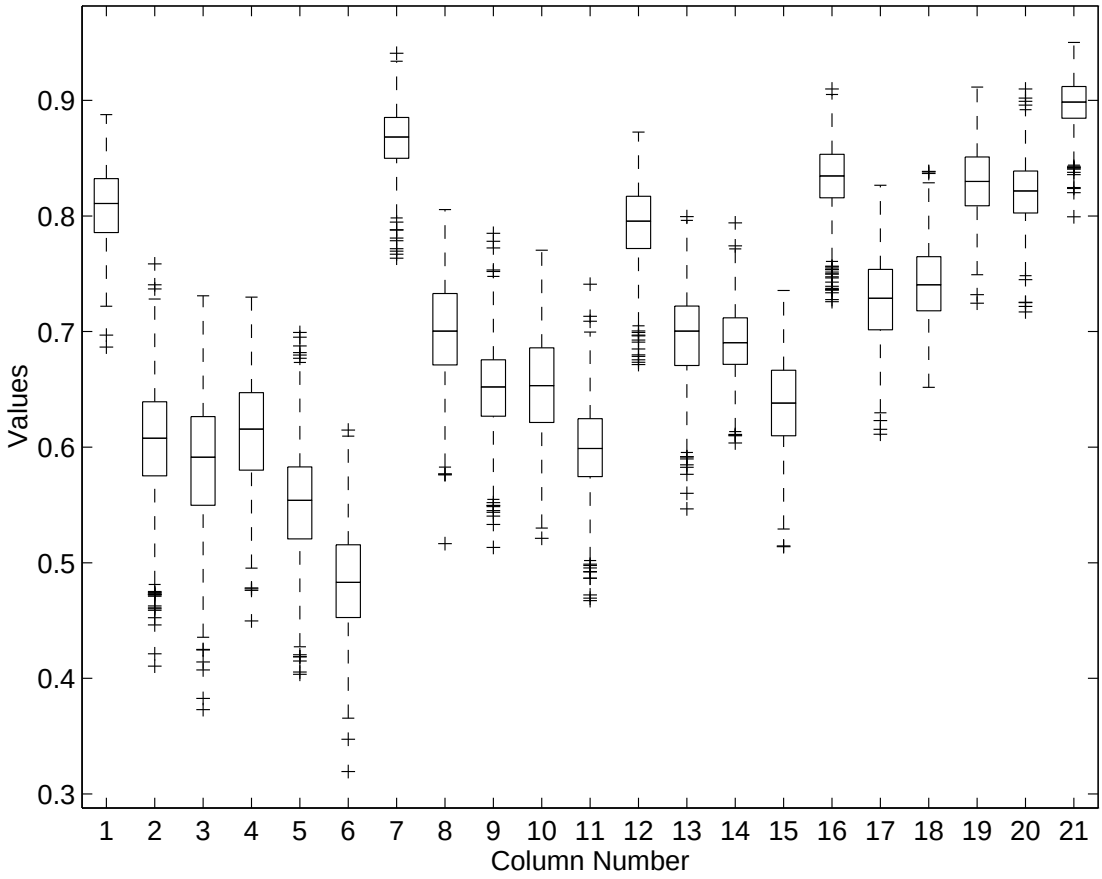

Figure 2: Marginal posterior box plots for elements of Σ in model M_1 (PSID data). The columns correspond to a row-wise expansion of Σ, ie., column 1 refers to σ_{12}, column 2 to σ_{13}, etc.

are similar and are not reported. The marginal posterior distributions are summarized in Figure 1. A box-plot summary of the marginal posterior distributions of the elements of Σ is presented in Figure 2 using every tenth draw from the simulation. The correlations are all quite large and precisely estimated, and most decline with an increase in the time lag. For M_2 the correlation parameter is estimated to be .739 with a posterior standard deviation of .057. From this evidence it would appear that the equi-correlated correlation structure is not appropriate for these data. This is confirmed from the marginal likelihood calculation, which yields $\ln \hat{m}(y|M_1) = -1561.91$ and $\ln \hat{m}(y|M_2) = -1598.36$. The evidence in favor of the unrestricted model is thus overwhelming.

6 Concluding remarks

The primary objective of this paper is to illustrate the value of Markov chain Monte Carlo methods for analyzing the multivariate probit model. We have presented techniques to simulate the posterior distributions of the unknown parameters with an unrestricted or restricted correlation structure and for finding the maximum likelihood estimates. The method we have developed for simulating elements of the correlation matrix through the Metropolis-Hastings algorithm is applicable to any problem in which the correlation or covariance matrix is subject to restrictions. In addition, the paper has established a framework for computing the marginal likelihood and Bayes factors from the output of the simulation. This same framework can be used for the computation of marginal likelihoods in other similar models, as we will report elsewhere. Our applications showed that the techniques can be applied to data sets of varying complexity and to high dimensional models that were hitherto intractable.

7 References

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669–679.

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press, Boston.

Ashford, J. R. and R. R. Sowden (1970). Multivariate probit analysis. *Biometrics*, 26, 535–546.

Avery, R. B., L. P. Hansen, and V. J. Hotz (1983). Multiperiod probit models and orthogonality condition estimation. *International Economic Review*, 24, 21–35.

Carey, V., S. L. Zeger, and P. Diggle (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80, 517–526.

Chib, S. (1993). Posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, in press.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313-1321.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327–335.

Dempster, A .P., N. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

FITZMAURICE, G. F. V. and N. M. LAIRD (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80, 141–151.

GELFAND, A. E. and A. F. M. SMITH (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

GEWEKE, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In E. Keramidas and S. Kaufman (eds.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.* Pp. 571–578. Interface Foundation of North America, Fairfax Station VA.

GLONEK, G. F. V. and P. MCCULLAGH (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533–546.

GREENE, W. (1993). *Econometric Analysis.* 2nd. ed. Macmillan, New York.

KASS, R. E. and A. E. RAFTERY (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

KIEFER, N. M. (1982). Testing for dependence in multivariate probit models. *Biometrika*, 69, 161–166.

LOUIS, T. A. (1982). Finding the observed information matrix using the EM algorithm. *Journal of the Royal Statistical Society* B, 44, 98–130.

MARSAGLIA, G. and I. OLKIN (1984). Generating correlation matrices. *SIAM Journal on Scientific and Statistical Computations*, 5, 470–475.

MCCULLOCH, R. E. and P. E. ROSSI (1994). Exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64, 207-240.

MENG, X. and D. B. RUBIN (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267–278.

NATARAJAN, R., C. E. MCCULLOCH, and N. M. KIEFER (1995). Maximum likelihood for the multinomial probit model. Manuscript.

OCHI, Y. and R. L. PRENTICE (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71, 531–543.

RIPLEY, B. D. (1987). *Stochastic Simulation.* John Wiley and Sons, New York.

ROUSSEEUW, P. and G. MOLENBERGHS (1994). The shape of correlation matrices. *American Statistician*, 48, 276–279.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

SMITH, A. F. M. and G. O. ROBERTS (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society,* B, 55, 3–24.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701–1762.

WEI, G. C. G. and M. A. TANNER (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85, 699–704.

(i)	y_1	y_2	C	INC	TAX	YRS	(i)	y_1	y_2	C	INC	TAX	YRS
1	1	1	1	9.77	7.0475	10	2	1	0	1	10.021	7.0475	8
3	1	0	1	10.021	7.0475	4	4	1	0	1	9.4335	6.3969	13
5	1	1	1	10.021	7.2792	3	6	1	0	1	10.463	7.0475	5
7	0	0	1	10.021	7.0475	4	8	1	1	1	10.021	7.2793	5
9	1	0	1	10.222	7.0475	10	10	1	1	1	9.4335	7.0475	5
11	1	1	1	10.021	7.0475	3	12	1	0	1	9.77	6.3969	30
13	1	1	1	9.77	6.7452	1	14	1	1	1	10.021	7.0475	3
15	1	1	1	10.82	6.7452	3	16	1	1	1	9.77	6.7452	42
17	1	1	1	10.222	7.0475	5	18	1	0	1	10.021	7.0475	10
19	1	1	1	10.222	7.0475	4	20	1	1	1	10.222	6.7452	4
21	1	1	1	10.463	7.0475	11	22	0	1	1	10.222	7.0475	5
23	1	1	1	9.77	6.7452	35	24	1	1	1	10.463	7.2793	3
25	1	1	1	10.021	6.7452	16	26	0	0	1	10.463	7.0475	7
27	1	1	1	9.77	6.7452	5	28	1	0	1	9.77	7.0475	11
29	1	0	1	9.77	6.7452	3	30	1	1	1	10.222	7.0475	2
31	1	1	1	10.021	6.7452	2	32	1	0	1	9.4335	6.7452	2
33	1	0	1	8.294	7.0475	2	34	0	1	1	10.463	7.0475	4
35	1	1	1	10.021	7.0475	2	36	1	0	1	10.222	7.2793	3
37	1	1	1	10.222	7.0475	3	38	1	1	1	10.222	7.4955	2
39	1	0	1	10.021	7.0475	10	40	1	1	1	10.222	7.0475	2
41	1	0	1	10.021	7.0475	2	42	1	0	1	10.82	7.4955	3
43	1	1	1	10.021	7.0475	3	44	1	1	1	10.021	7.0475	3
45	1	1	1	10.021	6.7452	6	46	1	1	1	10.021	7.0475	2
47	1	0	1	9.77	6.7452	26	48	0	0	1	10.222	7.4955	18
49	0	0	1	9.77	6.7452	4	50	0	0	1	10.021	7.0475	6
51	0	1	1	10.021	6.7452	12	52	1	1	1	9.4335	6.7452	49
53	1	1	1	10.463	7.2793	6	54	0	0	1	9.77	7.0475	18
55	1	1	1	10.021	7.0475	5	56	1	1	1	9.77	5.9915	6
57	1	0	1	9.4335	7.0475	20	58	1	1	1	9.77	6.3969	1
59	1	1	1	10.021	6.7452	3	60	1	0	1	10.463	7.0475	5
61	1	1	1	10.021	7.0475	2	62	1	0	1	10.82	7.2793	5
63	1	0	1	9.4335	6.7452	18	64	1	1	1	9.77	5.9915	20
65	0	0	1	8.9227	6.3969	14	66	1	0	1	9.4335	7.4955	3
67	1	0	1	9.4335	6.7452	17	68	1	0	1	10.021	7.0475	20
69	1	1	1	10.021	7.0475	3	70	1	1	1	10.021	7.0475	2
71	0	1	1	10.222	7.0475	5	72	1	1	1	9.77	7.0475	35
73	1	0	1	10.021	7.2793	10	74	1	1	1	9.77	7.0475	8
75	1	0	1	9.77	7.0475	12	76	1	1	1	10.222	6.7452	7
77	1	1	1	10.463	6.7452	3	78	1	0	1	10.222	6.7452	25
79	1	1	1	9.77	6.7452	5	80	1	1	1	10.222	7.0475	4
81	1	1	1	10.021	7.2793	2	82	1	1	1	10.463	6.7452	5
83	1	0	1	9.77	7.0475	3	84	1	1	1	10.82	7.4955	2
85	0	0	1	8.9227	5.9915	6	86	1	1	1	9.77	7.0475	3
87	1	1	1	9.4355	6.3969	12	88	0	1	1	9.77	6.7452	3
89	1	1	1	10.021	7.0475	3	90	0	1	1	10.021	6.7452	3
91	1	1	1	10.222	7.2793	3	92	1	1	1	10.021	7.0475	3
93	1	1	1	10.021	7.0475	5	94	0	1	1	8.9227	5.9915	35
95	1	0	1	10.463	7.4955	3							

Table 1: Voting data

		Prior		Posterior					
Param	MLE	Mean	Std dev	Mean	NSE	Std dev	Med	Lower	Upper
β_1	-4.764	0	10	-4.189	0.076	3.670	-4.193	-11.395	2.918
	0.1149	0	10	0.069	0.011	0.444	0.081	-0.820	0.911
	0.6699	0	10	0.654	0.014	0.563	0.658	-0.472	1.775
β_2	-0.3066	0	10	-0.474	0.081	3.787	-0.426	-7.878	6.923
	0.9895	0	10	1.057	0.011	0.438	1.042	0.244	1.953
	-1.3080	0	10	-1.380	0.014	0.584	-1.349	-2.599	-0.313
	-0.0176	0	10	-0.017	0.000	0.014	-0.017	-0.045	0.011
σ_{12}	0.317	0	.707	0.258	0.009	0.178	0.264	-0.103	0.589

Table 2: Voting data: ML and Bayes estimates. The Bayes estimates are reported along with the mean, the numerical standard error (NSE), the standard deviation (Std dev), the median (Med), and the 2.5th percentile and 97.5th percentiles (lower and upper).

Model M_i	$\ln m(y\|M_i)$	$\Pr(M_i\|y)$	Bayes factor
M_1: correlated probit	-126.31	0.495	0.522
M_2: independent probit	-126.30	0.505	1.01

Table 3: Voting data: Marginal likelihood, posterior probabilities and Bayes factors for alternative models.

No maternal smoking					Maternal smoking				
Age of child				Frequency	Age of child				Frequency
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

Table 4: Six Cities data set: child's wheeze status.

	M_1			M_2			M_3	
	MLE (s.e.)	Mean	Std dev	MLE (s.e.)	Mean	Std dev	Mean	Std dev
β	-1.118 (.065)	-1.127	.061	-1.120 (.043)	-1.121	.062	-1.126	.047
	-0.079 (.033)	-0.079	.031	-0.079 (.021)	-0.078	.031	-0.076	037
	0.152 (.102)	0.159	.098	0.172 (.072)	0.160	.099	0.168	.076
	0.039 (.052)	0.040	.051	0.041 (.034)	0.039	.049	0.035	.060
σ	0.584 (.068)	0.557	.069	0.602 (.025)	0.584	.054	–	–
	0.521 (.076)	0.496	.072	–	–	–	–	–
	0.586 (.095)	0.541	.075	–	–	–	–	–
	0.688 (.051)	0.656	.058	–	–	–	–	–
	0.562 (.077)	0.514	.073	–	–	–	–	–
	0.631 (.077)	0.601	.065	–	–	–	–	–
$\ln \hat{m}(y\|M_i)$	-825.58			-816.94			-931.16	

Table 5: Posterior Results: Six Cities data. M_1 is the unrestricted MVP model; M_2 is the model with an equicorrelated correlation structure, and M_3 is the independence model.