Analysis of multivariate probit models

By SIDDHARTHA CHIB

John M. Olin School of Business, Washington University, One Brookings Drive, St. Louis, Missouri 63130, U.S.A.

chib@simon.wustl.edu

AND EDWARD GREENBERG

Department of Economics, Washington University, One Brookings Drive, St. Louis, Missouri 63130, U.S.A.

edg@wuecona.wustl.edu

SUMMARY

This paper provides a practical 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 EM algorithm. A practical approach for the computation of Bayes factors from the simulation output is also developed. The methods are applied to a dataset with a bivariate binary response, to a four-year longitudinal dataset from the Six Cities study of the health effects of air pollution and to a seven-variate binary response dataset on the labour supply of married women from the Panel Survey of Income Dynamics.

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 & Diggle (1993) and Glonek & McCullagh (1995), relies on the generalisation of the binary logistic model to multivariate outcomes in conjunction with a particular parameterised representation for the correlations. Another strategy, discussed by Ashford & Sowden (1970) and Amemiya (1972), generalises 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 for flexible modelling 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 & Prentice, 1984). Thus, few applications of the model have appeared, and much of the potential of the model has not been realised.

The purpose of this paper is to provide a unified simulation-based inference method-

ology 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 & Smith, 1990; Smith & Roberts, 1993; Tierney, 1994; Chib & Greenberg, 1995) and in the Bayesian analysis of binary and polytomous data (Albert & Chib, 1993). Two important technical advances are made. 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. Secondly, 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 normalising constants.

The paper proceeds as follows. In § 2 we summarise the model and in § 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 § 4. These estimates are obtained by utilising a Monte Carlo version of the EM algorithm (Wei & Tanner, 1990; Meng & 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 maximisation. Section 5 presents three real data applications, and § 6 contains a brief discussion.

2. THE MULTIVARIATE PROBIT MODEL

Let Y_{ij} denote a binary 0/1 response on the ith observation unit and jth variable, and let $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ $(1 \leqslant i \leqslant n)$ 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 β, Σ and a set of covariates x_{ij}, is given by

$$\text{pr}(Y_i = y_i \,|\, \beta, \Sigma) \equiv \text{pr}(y_i \,|\, \beta, \Sigma) = \int_{A_{iJ}} \cdots \int_{A_{i1}} \phi_J(t \,|\, 0, \Sigma)\, dt, \tag{1}$$

where $\phi_J(t \,|\, 0, \Sigma)$ is the density of a J-variate normal distribution with mean vector 0 and correlation matrix $\Sigma = \{\sigma_{jk}\}$, A_{ij} is the interval

$$A_{ij} = \begin{cases} (-\infty, x_{ij}'\beta_j) & \text{if } y_{ij} = 1, \\ [x_{ij}'\beta_j, \infty) & \text{if } y_{ij} = 0, \end{cases}$$

$\beta_j \in R^{k_j}$ is an unknown parameter vector and $\beta' = (\beta_1', \ldots, \beta_J') \in R^k$, $k = \sum k_j$. We denote the $p = J(J-1)/2$ free parameters of Σ by $\sigma \equiv (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{J-1,J})$.

It is important to note that Σ must be in correlation form for identifiability reasons. Suppose that (γ, Ω) is an alternative parameterisation, where γ is the regression parameter vector and Ω is the covariance matrix. Then it is easy to show that $\text{pr}(y_i \,|\, \gamma, \Omega) = \text{pr}(y_i \,|\, \beta, \Sigma)$, where $\beta_j = \omega_{jj}^{-1/2}\gamma_j$, $\Sigma = C\Omega C'$ and $C = \text{diag}\{\omega_{11}^{-1/2}, \ldots, \omega_{JJ}^{-1/2}\}$. A parameterisation in terms of covariances is therefore not likelihood identified.

For our purposes a more convenient formulation of the multivariate probit model is in terms of Gaussian latent variables. Let $Z_i = (z_{i1}, \ldots, z_{iJ})$ denote a J-variate normal vector with distribution $Z_i \sim N_J(X_i\beta, \Sigma)$, 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 z_{ij}:

$$y_{ij} = I(z_{ij} > 0) \quad (j = 1, \ldots, J), \tag{2}$$

where $I(A)$ is the indicator function of the event A. In this formulation the probability in (1) may be expressed as

$$\int_{B_{iJ}} \cdots \int_{B_{i1}} \phi_J(Z_i \mid X_i\beta, \Sigma)\, dZ_i, \tag{3}$$

where B_{ij} is the interval $(0, \infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$. We let $B_i = B_{i1} \times B_{i2} \times \ldots B_{iJ}$ and note that B_i is the set-valued inverse of the mapping in (2). It is important to bear in mind that B_i, unlike the A_{ij}, depends only on the value of y_i and not on the parameters. This latent variable representation and the inverse mapping from y_{ij} to z_{ij} form the basis of our posterior sampling method.

3. Posterior analysis

3·1. *Introduction*

Given a random sample of nJ observations $y = (y_1, \ldots, y_n)$ and a prior density $\pi(\beta, \sigma)$ on the parameters of a multivariate probit model, the posterior density is

$$\pi(\beta, \sigma \mid y) \propto \pi(\beta, \sigma)\, \mathrm{pr}(y \mid \beta, \Sigma), \quad \beta \in \Re^k, \quad \sigma \in C,$$

where $\mathrm{pr}(y \mid \beta, \Sigma) = \prod \mathrm{pr}(y_i \mid \beta, \Sigma)$ is the likelihood function, C is a convex solid body in the hypercube $[-1, 1]^p$ that leads to a proper correlation matrix, see Rousseeuw & Molenberghs (1994) for more on the shape of correlation matrices, and $\mathrm{pr}(y_i \mid \beta, \Sigma)$ is the integral of the multivariate normal density given in (3). This form of the posterior density is not particularly useful for Bayesian estimation, because the evaluation of the likelihood function is computationally intensive.

In this paper we invoke an alternative procedure that is based on the framework developed by Albert & Chib (1993). The idea is to focus on the joint posterior distribution of the parameters and the latent data $\pi(\beta, \sigma, Z_1, \ldots, Z_n \mid y)$. From Bayes' theorem,

$$\pi(\beta, \sigma, Z \mid y) \propto \pi(\beta, \Sigma)f(Z \mid \beta, \Sigma)\, \mathrm{pr}(y \mid Z, \beta, \Sigma),$$

where we have let $Z = (Z_1, \ldots, Z_n)$. Use the mapping in equation (2) to note that $\mathrm{pr}(y_i \mid Z_i, \beta, \Sigma) = I(Z_i \in B_i)$. Hence,

$$\pi(\beta, \sigma, Z \mid y) \propto \pi(\beta, \sigma) \prod_{i=1}^{n} f(Z_i \mid \beta, \Sigma)I(Z_i \in B_i), \tag{4}$$

where

$$f(Z_i \mid \beta, \Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(Z_i - X_i\beta)'\Sigma^{-1}(Z_i - X_i\beta)\}I(\sigma \in C),$$

independently across observations. Thus, the effect of y_i appears only through B_i, and likelihood evaluation is not required. The problem is now much simplified, and we can take a sampling-based approach, in conjunction with Markov chain Monte Carlo methods, to summarise (4). The Markov chain sampling scheme can be constructed from the distributions $[Z_i \mid y_i, \beta, \Sigma]$ $(i \leqslant n)$, $[\beta \mid y, Z, \Sigma]$ and $[\sigma \mid y, Z, \beta]$. As we show below, each of these distributions can be sampled either directly or by Markov chain methods.

In passing we mention that no other Markov chain sampling scheme appears viable for this model. For a number of reasons, we do not think that it is a good idea to sample the non-identified parameters, γ and Ω, as in McCulloch & Rossi (1994) in a different context. First, this approach requires that a prior be formulated on unidentified param-

eters. This prior cannot be too weak, because that would lead to a flat posterior and convergence problems, or too strong, because then it would determine the posterior. Secondly, this approach does not work when Σ is patterned or restricted, as in some of the examples below. Finally, it is difficult to compute Bayes factors using this approach. With weak priors, problems similar to those described by the Lindley paradox can arise, especially if the model dimensions are quite different.

3·2. Posterior simulations

We begin with the sampling of the latent data Z_i from $[Z_i \mid y_i, \beta, \Sigma]$ $(i \leqslant n)$, given values of (β, Σ). From (4) it follows that

$$f(Z_i \mid y_i, \beta, \Sigma) \propto \phi_J(Z_i \mid X_i\beta, \Sigma) \prod_{j=1}^{J} \{I(z_{ij} > 0)I(y_{ij} = 1) + I(z_{ij} \leqslant 0)I(y_{ij} = 0)\}, \qquad (5)$$

where we have written out the set B_i in full. This is a multivariate normal density truncated to the region specified by B_i. For example, if $J = 2$ and $y_i = (1, 1)'$, then the normal distribution is truncated to the positive orthant. To sample this distribution one can use the method of Geweke (1991) to compose a cycle of J Gibbs steps through the components of Z_i. In the jth step of this cycle, z_{ij} is simulated from $z_{ij} \mid \{y_{ij}, z_{ik} \ (k \neq j), \beta, \Sigma\}$, which is a univariate normal distribution truncated to the region B_{ij}. The parameters of the untruncated normal distribution $z_{ij} \mid \{z_{ik} \ (k \neq j), \beta, \Sigma\}$ are obtained from the usual formulae, and the truncated version is simulated by the inverse distribution function method (Devroye, 1985, p. 38).

To describe the sampling of β, we assume prior independence between β and σ and let $\pi(\beta) = \phi_k(\beta \mid \beta_0, B_0^{-1})$, where the location is controlled by the vector β_0 and its strength by the precision matrix B_0. By combining terms of β in (4) we obtain the standard linear models result

$$\beta \mid (Z, \Sigma) \sim N_k(\beta \mid \hat{\beta}, B^{-1}), \qquad (6)$$

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.

Finally, we consider the sampling of σ from $\pi(\sigma \mid Z, \beta)$ given the prior

$$\pi(\sigma) \propto \phi_p(\sigma \mid \sigma_0, G_0^{-1}), \quad \sigma \in C,$$

a normal distribution truncated to the region C. From (4), $\pi(\sigma \mid Z, \beta) \propto \pi(\sigma)f(Z \mid \beta, \Sigma)$, where

$$f(Z \mid \beta, \Sigma) = |\Sigma|^{-n/2} \exp\{-\tfrac{1}{2} \operatorname{tr}(Z^* - \Delta)'\Sigma^{-1}(Z^* - \Delta)\}I(\sigma \in C),$$

and $Z^* = (Z_1, Z_2, \ldots, Z_n)$ and $\Delta = (X_1\beta, \ldots, X_n\beta)$ are $J \times n$ matrices. This distribution is nonstandard. In sampling this density we must pay attention to the facts that (Z, β) are refreshed across iterations, that suitable bounds and dominating functions are difficult to obtain, and that σ must lie in C. We deal with these problems by adopting the Metropolis–Hastings algorithm (Hastings, 1970; Chib & Greenberg, 1995). Let $q(\sigma' \mid \sigma, Z, \beta)$ denote a density that supplies candidate values σ' given the current value σ. The choice of this density is explained below. Our algorithm is then formally described by the following two steps.

Algorithm
1. *Sample a proposal value σ' given σ from the density $q(\sigma'|\sigma, Z, \beta)$.*
2. *Move to σ' with probability*

$$\alpha(\sigma, \sigma') = \min\left\{\frac{\pi(\sigma')f(Z|\beta, \Sigma')I(\sigma' \in C)}{\pi(\sigma)f(Z|\beta, \Sigma)I(\sigma \in C)}\frac{q(\sigma|\sigma', Z, \beta)}{q(\sigma'|\sigma, Z, \beta)}, 1\right\},$$

and stay at σ with probability $1 - \alpha(\sigma, \sigma')$.

We make two general remarks. First, the proposal density q is not truncated to C, because that constraint is part of the target density. Thus, if Σ' is not positive definite, the conditional posterior is zero, and the proposal value is rejected with certainty. By specifying q in this way we are able to simplify both the choice of q and the determination of the proposal density of the reverse move $q(\sigma'|\sigma, Z, \beta)$. Secondly, if the dimension of Σ is large, as in our third example in § 5 below, it is best to partition σ into blocks and to apply the Metropolis–Hastings algorithm in sequence through the various blocks.

We now discuss the choice of q. As in all Metropolis samplers, the general objectives are to traverse the parameter space and produce output that mixes well; see Chib & Greenberg (1995) for further details.

One simple way to generate proposal values is through 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. If h is assumed to follow a symmetric distribution such as the normal, the ratio of the $q(.)$'s is one, and the probability of move is determined by the ratio of the density ordinates. To generate suitable moves 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 the correlation coefficient, or λ, the smallest characteristic root of Σ; see Marsaglia & Olkin (1984) for an explanation. This proposal density is usually effective for small p.

A more general procedure is based on a proposal density that is tailored to the un-normalised target density $g(\sigma|Z, \beta) = \pi(\sigma)f(Z|\beta, \Sigma)I(\sigma \in C)$. The form of this proposal generating process is given by

$$\sigma' = \mu^* + h, \quad \mu^* = \mu + P(\sigma - \mu), \tag{7}$$

where μ is a vector, P is a $p \times p$ diagonal matrix and h is a multivariate-t vector with mean zero, dispersion $\tau^2 V$ and v degrees of freedom. Of the five tuning parameters (μ, V, P, τ, v) in this proposal density, two (μ, V) depend on the current values of β and Z and automatically adjust as the iterations proceed. The resulting proposal density is extremely flexible and leads to competitive proposal values. The parameter μ is taken to be the approximate mode of the function $\log\{g(\sigma|Z, \beta)\}$, and V is taken to be inverse of the negative Hessian matrix of $\log\{g(\sigma|Z, \beta)\}$ at the mode. These quantities are obtained from two or three Newton–Raphson steps, initialised at the mode from the previous round. Next, P is either equal to zero, leading to what may be called the tailored independence chain, or equal to minus the identity matrix of order p, giving the tailored reflection chain. Reflection is a useful device for generating large moves, from the current point to the other side of the mode where the ordinate is roughly similar, that have a high probability of being accepted. Finally, τ^2 is adjusted by experimentation, and v is specified arbitrarily at 10, or some similar value.

One cycle of the resulting Markov chain Monte Carlo algorithm is completed by simulating all the distributions in fixed or random order. A full sample from the posterior

distribution is generated by repeating this process a large number of times. All posterior inferences are based on the sample furnished by the Markov chain procedure.

3·3. *Computation of marginal likelihood*

We now consider the question of comparing alternative multivariate probit models. Typically, competing models arise from restrictions on the covariate or correlation structure. One example of this is the restriction that $\beta_j = \beta$ across the J responses. Another is that Σ is in the equi-correlated form $(1 - \rho)I_J + \rho 1_J 1_J'$, where $|\rho| < 1$ (Ochi & Prentice, 1984). In a panel data context, when the index j represents time, Σ may be specified to reflect the assumption of serially correlated errors or the assumption of 1-dependence. Finally, a benchmark restriction is that the responses are independent and Σ is diagonal (Kiefer, 1982).

These alternative models may be compared in a Bayesian context by Bayes factors, or ratios of model marginal likelihoods (Kass & Raftery, 1995). The marginal likelihood of model $\mathcal{M}_k$ ($k = 1, 2, \ldots, K$) is defined as

$$m(y \mid \mathcal{M}_k) = \int \mathrm{pr}(y \mid \mathcal{M}_k, \beta, \Sigma)\pi(\beta, \sigma \mid \mathcal{M}_k)\, d\beta\, d\sigma,$$

and is a function of $\mathcal{M}_k$. This function is sometimes estimated by the Laplace method or importance sampling. In the present case, however, these methods are not attractive, because they involve the repeated evaluation of the likelihood function $\mathrm{pr}(y \mid \mathcal{M}_k, \beta, \Sigma)$. For this reason, we focus on an alternative method recently developed in Chib (1995). The Chib method relies on an identity in (β, Σ). This identity, which is obtained by rewriting the expression of the posterior density, states that

$$m(y \mid \mathcal{M}_k) = \frac{\mathrm{pr}(y \mid \mathcal{M}_k, \beta, \Sigma)\pi(\beta, \sigma \mid \mathcal{M}_k)}{\pi(\beta, \sigma \mid \mathcal{M}_k, y)},$$

where all integrating constants on the right-hand side are included. Since the left-hand side is free of the parameters, the marginal likelihood may be calculated by evaluating the terms on the right-hand side at any single point (β^*, σ^*). In practice, a high density point such as the posterior mean is used. Thus, on the log scale

$$\log m(y \mid \mathcal{M}_k) = \log \mathrm{pr}(y \mid \mathcal{M}_k, \beta^*, \Sigma^*) + \log \pi(\beta^* \mid \mathcal{M}_k) + \log \pi(\sigma^* \mid \mathcal{M}_k)$$

$$- \log \pi(\beta^* \mid \mathcal{M}_k, y, \sigma^*) - \log \pi(\sigma^* \mid \mathcal{M}_k, y), \tag{8}$$

where we have expressed the posterior ordinate by a conditional/marginal decomposition.

We now suppress the dependence on $\mathcal{M}_k$ and discuss the estimation of the terms in (8) that are not available through direct computation, starting with

$$\pi(\sigma^*) = \phi_p(\sigma^* \mid \sigma_0, G_0^{-1})\, \mathrm{pr}(\sigma \in C).$$

The normalising constant of this density can be found by simulation: simply generate a large number of observations from $\phi_p(\sigma \mid \sigma_0, G_0^{-1})$ and find the proportion that satisfy the positive definiteness constraint. This proportion is the Monte Carlo estimate of $\mathrm{pr}(\sigma \in C)$.

Next consider the estimation of the conditional posterior ordinate

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

where $\pi(\beta^* \mid y, Z, \Sigma^*)$ is the multivariate normal density in (6) with $\beta = \beta^*$ and $\Sigma = \Sigma^*$. A

key point is that this integral can be estimated very accurately by drawing a large sample of Z values from the density $p(Z | y, \Sigma^*)$. As in Chib (1995), these Z variates are produced from a reduced Markov chain Monte Carlo run consisting of the distributions

$$[Z_1 | y_1, \beta, \Sigma^*] \times \ldots \times [Z_n | y_n, \beta, \Sigma^*], \quad [\beta | Z_1, \ldots, Z_n, \Sigma^*],$$

which are the distributions discussed above but with Σ fixed at Σ^*. Given a sample of G values on Z from this run, an estimate of $\pi(\beta^* | y, \Sigma^*)$ is available as

$$\hat{\pi}(\beta^* | y, \Sigma^*) = G^{-1} \sum_{g=1}^{G} \phi_J(\beta^* | \hat{\beta}^{(g)}, B^{-1}), \tag{9}$$

where $\hat{\beta}^{(g)} = B^{-1}(B_0 \beta_0 + \sum_{i=1}^{n} X_i' \Sigma^{*-1} Z_i^{(g)})$ and $B = B_0 + \sum_{i=1}^{n} X_i' \Sigma^{-1} X_i$.

For the marginal ordinate $\pi(\sigma^* | y) = \int \pi(\sigma^* | Z, \beta) p(Z, \beta | y) \, d\beta \, dZ$ a different approach is necessary because the normalising constant of $\pi(\sigma^* | Z, \beta)$ is not known. Let $K(x)$ denote a univariate kernel density, taken to be $\phi(x | 0, 1)$ in the examples, and s_j the standard deviation of $\{\sigma_j^{(g)}\}$. Then an estimate of $\pi(\sigma^* | y)$ is

$$\hat{\pi}(\sigma^* | y) = G^{-1} \sum_{g=1}^{G} \prod_{j=1}^{p} b_j^{-1} K\{b_j^{-1}(\sigma_j^* - \sigma_j^{(g)})\}, \tag{10}$$

where the bandwidth b_j is $s_j G^{-1/(p+4)}$ (Scott, 1992, p. 150). The sample $\sigma^{(g)}$ can be thinned before smoothing is done.

The kernel approach may have to be modified somewhat if σ is high-dimensional. Then, one should partition σ into several low-dimensional blocks and apply the reduced Markov chain Monte Carlo procedure to each of the blocks. This overcomes the well-known difficulties of kernel smoothing in high dimensions at the cost of additional computations. Note that in standard applications of kernel smoothing the sample is fixed and this refinement is not possible. We illustrate the idea with two blocks, $\sigma = (\sigma_1, \sigma_2)$ where σ_i contains p_i components. 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 samples are generated from the appropriate distributions. For the first term, the values of $\sigma_1^{(g)}$ $(g = 1, \ldots, G)$ are available from the full Markov chain Monte Carlo sampler. These values are smoothed by the kernel method as described above. To estimate the next term we run a reduced Markov chain Monte Carlo sampler consisting of

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

The values $\{\sigma_2^{(g)}\}$ generated from this sampler are from $[\sigma_2 | y, \sigma_1^*]$ and can be smoothed to estimate the ordinate $\pi_2(\sigma_2^* | y, \sigma_1^*)$.

We mention that the numerical standard error of the resulting posterior density estimate $\log \hat{\pi}(\beta^*, \sigma^* | y)$ can be estimated precisely as explained in Chib (1995), despite the fact that kernel smoothing is used to estimate $\hat{\pi}(\sigma^* | y)$. The numerical standard error is based on the variance of $G^{-1} \sum_{g=1}^{G} h^{(g)}$, where $h^{(g)}$ is a vector whose components are the density ordinates $\phi_J(\beta^* | \hat{\beta}^{(g)}, B^{-1})$ from (9) and $\prod_{j=1}^{p} b_j^{-1} K\{b_j^{-1}(\sigma_j^* - \sigma_j^{(g)})\}$ from (10). The vector $h^{(g)}$ is available as a by-product of the marginal likelihood calculation. Further details may be found in Chib (1995).

Finally, we consider the computation of $\log \mathrm{pr}(y | \beta^*, \Sigma^*) = \sum_{i=1}^{n} \log \mathrm{pr}(y_i | \beta^*, \Sigma^*)$, where $\mathrm{pr}(y_i | \beta^*, \Sigma^*)$ is given in (3). One way to compute this integral is by the approach called the Geweke–Hajivassiliou–Keane method in the econometrics literature. See Greene

(1997, p. 196) for further details. This method is based on writing $\Sigma = LL'$, where L is the lower triangular Choleski factorisation, and making a change of variable from Z_i to ε_i, where $Z_i = X_i \beta + L\varepsilon_i$. Then

$$\text{pr}(y_i | \beta^*, \Sigma^*) = \int_{c_{iJ}^*}^{d_{iJ}^*} \cdots \int_{c_{i1}^*}^{d_{i1}^*} \phi_J(t | 0, I) \, dt, \tag{11}$$

where

$$c_{ij}^* = \frac{c_{ij} - x_{ij}'\beta_j - \sum_{k=1}^{j-1} l_{jk}\varepsilon_{ik}}{l_{jj}}, \quad d_{ij}^* = \frac{d_{ij} - x_{ij}'\beta_j - \sum_{k=1}^{j-1} l_{jk}\varepsilon_k}{l_{jj}},$$

and (c_{ij}, d_{ij}) denotes the lower and upper endpoints of B_{ij} $(j \leqslant J)$. This integral can now be calculated by recursive Monte Carlo simulations that are repeated a large number of times. In the examples we use 10 000 iterations.

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 maximiser of the posterior without computation of the likelihood function. This can be done by an adaptation of the Monte Carlo EM algorithm (MCEM) that was proposed by Wei & Tanner (1990), which in turn is a stochastic modification of the original Dempster, Laird & Rubin (1977) EM algorithm.

Let $\theta = (\beta, \sigma)$ and consider the expectation or E-step of the EM algorithm, given the current value of the maximiser $\theta^{(t)}$:

$$Q(\theta, \theta^{(t)}) = \int_Z \log\{f(y, Z | \theta)\} \, d[Z | y, \theta^{(t)}] = \int_Z \log\{f(Z | \theta)\} \, d[Z | y, \theta^{(t)}],$$

where the integral is with respect to the truncated normal distribution of Z and the second equality is a consequence of (4). This calculation is analytically intractable. In the MCEM algorithm, the Q function is estimated by

$$\hat{Q}(\theta, \theta^{(t)}) = N^{-1} \sum_j \log\{f(Z^{(j)} | \theta)\}$$

$$= -\frac{1}{2} \log|\Sigma| - N^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} (Z_i^{(j)} - X_i\beta)'\Sigma^{-1}(Z_i^{(j)} - X_i\beta),$$

where $Z^{(j)}$ $(j = 1, \ldots, N)$ are draws from (5). In the M-step of the algorithm, the $\hat{Q}$ function is maximised over θ to obtain the new parameter $\theta^{(t+1)}$. The MCEM algorithm is terminated once the difference $\|\theta^{(t+1)} - \theta^{(t)}\|$ is negligible.

We now follow Meng & Rubin (1993) and complete the M-step through a sequence of two conditional maximisations, the maximisation over β given Σ and the maximisation over Σ given β. This simplifies the update of θ and parallels the blocking adopted in the Bayesian simulation presented above; a similar conditional maximisation step is adopted in a different context in unpublished work by R. Natarajan, C. E. McCulloch and N. M. Kiefer 'Maximum likelihood for the multinomial probit model'. Specifically, if we set the derivative with respect to β equal to zero, the update of β is given by

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

where $\bar{Z}_i = N^{-1} \sum Z_i^{(j)}$ 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 maximising over σ with a Newton–Raphson type routine. Although not necessary, it is preferable for efficiency considerations to re-draw the Z values from the distribution $[Z \mid y, \beta^{(t+1)}, \Sigma]$ and re-compute the $\hat{Q}$ function before the second maximisation is attempted. The update for σ is thus obtained by maximising 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 & Tanner (1990), these iterations are started with a small value of N that is increased as the maximiser is approached. Given the modal value $\hat{\theta}$, the standard errors of the estimate are obtained by the formula of Louis (1982). Specifically, the observed information matrix is given by

$$-E \left\{ \frac{\partial^2 \log f(Z \mid \theta)}{\partial \theta \, \partial \theta'} \right\} - \text{var} \left\{ \frac{\partial \log f(Z \mid \theta)}{\partial \theta} \right\},$$

where the expectation and variance are with respect to the distribution $Z \mid y, \hat{\theta}$. Each of these terms is estimated by taking an additional M draws $\{Z^{(1)}, \ldots, Z^{(M)}\}$ from $Z \mid y, \hat{\theta}$. Standard errors are equal to the square roots of the diagonal elements of the inverse of the estimated information matrix.

5. APPLICATIONS

5·1. *Bivariate probit for voter behaviour*

Our first application is to survey data of the voting behaviour of 95 residents of Troy, Michigan, in which the first decision, Y_{i1}, is whether or not to send at least one child to public school and the second, Y_{i2}, is whether or not to vote in favour of a school budget. The objective of the study is to model the two binary responses as a function of covariates, allowing for correlation in the responses. As in Greene (1997, p. 906), the covariates in x_{i1} are 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; those in x_{i2} are a constant, INC, TAX and the number of years YRS the resident has been living in Troy.

We fit two models to these data. The first, denoted by $\mathcal{M}_1$, is the bivariate probit model in which the marginal probabilities of the ith subject are given by

$$\text{pr}(y_{ij} = 1 \mid \beta, \sigma) = \Phi(x'_{ij} \beta_j),$$

and the joint probabilities are given by the distribution function of the bivariate normal with correlation matrix

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

Model $\mathcal{M}_1$ thus contains seven unknown regression parameters and one unknown correlation parameter. The second model, denoted by $\mathcal{M}_2$, is an independent probit model in which $\sigma_{12} = 0$.

The modal maximum likelihood estimates in this instance can be computed by directly maximising the likelihood function and the MCEM algorithm is not necessary. These esti-

mates are reported in Table 1. For model $\mathcal{M}_1$, the posterior distributions of the parameters are obtained by applying the Markov chain Monte Carlo algorithm described in § 3 for 6000 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 used above, these settings imply that $\beta_0 = 0$, $B_0 = 10^{-2}I_7$, $\sigma_0 = 0$ and $G_0^{-1} = 0\cdot5$.

Table 1. *Voting data: maximum likelihood and Bayes estimates. The Bayes estimates are reported along with the mean, the numerical standard error* (NSE), *the standard deviation* (SD), *the median* (Med.) *and the lower 2·5th and upper 97·5th percentiles*

		PRIOR		Posterior					
Parameter	MLE	MEAN	SD	Mean	NSE	SD	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	0·707	0·258	0·009	0·178	0·264	−0·103	0·589

We use the random walk proposal density in the Metropolis–Hastings step and let the random increment be univariate normal with standard deviation equal to $4\sqrt{(1/n)}$. This results in an acceptance rate of about 0·5. Our results for the posterior distribution are summarised in Table 1; the results for the independent probit obtained via the algorithm of Albert & Chib (1993) are similar. Table 1 reports the maximum likelihood 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 assess the evidence more formally, we calculate the marginal likelihoods of $\mathcal{M}_1$ and $\mathcal{M}_2$, evaluating all the quantities in (8) at the posterior mean. Note that both the likelihood function and the normalising constant of the truncated normal prior of σ_{12} are available without simulation. To obtain the conditional ordinate of β at β^* we use a reduced run of 6000 iterations. The results, along with the posterior probabilities of the respective models under the assumption that the prior odds equal one, are reported in Table 2. The data thus do not provide support for the correlated probit model over the independent probit model.

Table 2. *Voting data: Marginal likelihood, posterior probabilities and Bayes factors for alternative models*

Model $\mathcal{M}_i$	$\log m(y\mid\mathcal{M}_i)$	$\mathrm{pr}(\mathcal{M}_i\mid y)$	Bayes factor
$\mathcal{M}_1$: correlated probit	−126·31	0·495	0·980
$\mathcal{M}_2$: independent probit	−126·30	0·505	1·020

5·2. *Six Cities study*

The second example is based on a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution, which has been analysed by Fitzmaurice & Laird (1993) and Glonek & McCullagh (1995) with a multivariate logit model. The data, reproduced in Table 3, contain repeated binary measures of 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.

Table 3. *Six Cities dataset: Child's wheeze frequency*

| No maternal smoking | | | | | Maternal smoking | | | | |
| Age of child | | | | | Age of child | | | | |
7	8	9	10	Frequency	7	8	9	10	Frequency
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

Interpreting age as category j, we fit three models to these data: the full multivariate probit model, $\mathcal{M}_1$, the equi-correlated model, $\mathcal{M}_2$, and the independent probit model, $\mathcal{M}_3$. In each model the marginal probability of response is specified as

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

where x_{i1} is the age of the child, centred 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, in contrast to 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 $\mathcal{M}_1$, one unknown correlation parameter in $\mathcal{M}_2$ and no unknown correlation parameter in $\mathcal{M}_2$.

For all three models the prior for β is represented by the hyperparameters $\beta_0 = 0$ and $B_0 = 10^{-1} I_4$, and that for σ in $\mathcal{M}_1$ and $\mathcal{M}_2$ by $\sigma_0 = 0$, $G_0^{-1} = 0\cdot5 I_6$ and $\sigma_0 = 0$, $G_0^{-1} = 0\cdot5$, respectively, where, here and below, we use the same symbols for the hyperparameters across the different models. Posterior sampling of the correlation parameters of model $\mathcal{M}_1$ is by the tailored independence method applied in one block to all six unknown parameters of Σ. The parameters μ and V are obtained by the Newton–Raphson method. The parameter τ is set equal to 1·5. The same approach is used to sample the single parameter of Σ

in model $\mathcal{M}_2$ with $\tau = 4{\cdot}0$. 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 $\mathcal{M}_1$ and $\mathcal{M}_2$, 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 summarised in Table 4. First, note that the maximum likelihood and Bayes estimates of β are very insensitive to the specification of the covariance structure. Secondly, the maximum likelihood values differ slightly from the posterior means, an indication of some asymmetry in the posterior distributions. Thirdly, the estimated standard errors, SE, of the maximum likelihood estimates are generally smaller than the corresponding posterior standard deviations. Fourthly, there is little support for $\mathcal{M}_3$ because the posterior distribution of Σ in both $\mathcal{M}_1$ and $\mathcal{M}_2$ is concentrated away from zero. The marginal likelihoods of the respective models support this conclusion. Some details of the implementation as they relate to $\mathcal{M}_1$ and $\mathcal{M}_2$ are as follows. A simulation of $10\,000$ draws from the untruncated normal prior on σ is used to estimate the normalising constant of the prior, as discussed in § 3. The likelihood contribution is computed using (11). The conditional posterior ordinate of β at β^*, the posterior mean, is estimated from a reduced Markov chain Monte Carlo run of $10\,000$ iterations, while the marginal posterior density ordinate of σ at σ^*, the posterior mean, is estimated in one block by kernel smoothing. Interestingly, the marginal posterior ordinate at σ^*, in the case of $\mathcal{M}_1$, changed only slightly when it was estimated as $\pi(\sigma_1^*, \sigma_2^* \mid y) = \pi(\sigma_1^* \mid y)\pi(\sigma_2^* \mid y, \sigma_1^*)$ with two blocks of size three. Thus, in this case, kernel smoothing is accurate for a six-dimensional density. We conjecture that this is because a large sample on σ is being used to estimate a single high density ordinate.

The results show that the marginal likelihood of the independence model $\mathcal{M}_3$ is the smallest of the three models and that of $\mathcal{M}_2$ the largest. The log marginal likelihood increased by about $0{\cdot}30$ 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. We computed standard errors for the marginal likelihood computation; these are $0{\cdot}13$ for $\mathcal{M}_1$ and $0{\cdot}04$ for $\mathcal{M}_2$

Table 4. *Six Cities data: Posterior results for $\mathcal{M}_1$, the unrestricted multivariate probit model, $\mathcal{M}_2$, the model with an equicorrelated correlation structure, and $\mathcal{M}_3$, the independence model*

	$\mathcal{M}_1$				$\mathcal{M}_2$				$\mathcal{M}_3$	
	MLE	(SE)	Mean	SD	MLE	(SE)	Mean	SD	Mean	SD
β	$-1{\cdot}118$	$(0{\cdot}065)$	$-1{\cdot}127$	$0{\cdot}061$	$-1{\cdot}120$	$(0{\cdot}043)$	$-1{\cdot}121$	$0{\cdot}062$	$-1{\cdot}126$	$0{\cdot}047$
	$-0{\cdot}079$	$(0{\cdot}033)$	$-0{\cdot}079$	$0{\cdot}032$	$-0{\cdot}079$	$(0{\cdot}021)$	$-0{\cdot}078$	$0{\cdot}031$	$-0{\cdot}076$	$0{\cdot}037$
	$0{\cdot}152$	$(0{\cdot}102)$	$0{\cdot}160$	$0{\cdot}099$	$0{\cdot}172$	$(0{\cdot}072)$	$0{\cdot}160$	$0{\cdot}099$	$0{\cdot}168$	$0{\cdot}076$
	$0{\cdot}039$	$(0{\cdot}052)$	$0{\cdot}040$	$0{\cdot}053$	$0{\cdot}041$	$(0{\cdot}034)$	$0{\cdot}038$	$0{\cdot}049$	$0{\cdot}035$	$0{\cdot}060$
σ	$0{\cdot}584$	$(0{\cdot}068)$	$0{\cdot}557$	$0{\cdot}068$	$0{\cdot}602$	$(0{\cdot}025)$	$0{\cdot}584$	$0{\cdot}054$	—	—
	$0{\cdot}521$	$(0{\cdot}076)$	$0{\cdot}497$	$0{\cdot}073$	—	—	—	—	—	—
	$0{\cdot}586$	$(0{\cdot}095)$	$0{\cdot}541$	$0{\cdot}075$	—	—	—	—	—	—
	$0{\cdot}688$	$(0{\cdot}051)$	$0{\cdot}656$	$0{\cdot}058$	—	—	—	—	—	—
	$0{\cdot}562$	$(0{\cdot}077)$	$0{\cdot}513$	$0{\cdot}073$	—	—	—	—	—	—
	$0{\cdot}631$	$(0{\cdot}077)$	$0{\cdot}601$	$0{\cdot}065$	—	—	—	—	—	—

$$\log \hat{m}(y \mid \mathcal{M}_i) = -825{\cdot}03 \qquad \log \hat{m}(y \mid \mathcal{M}_i) = -816{\cdot}80 \qquad \log \hat{m}(y \mid \mathcal{M}_i) = -931{\cdot}16$$

on the log scale and are negligible. The Bayes factors are as follows: $B_{1,2} = 2.76 \times 10^{-4}$, $B_{1,3} = 1.24 \times 10^{46}$ and $B_{2,3} = 4.63 \times 10^{49}$. Unless the prior probabilities for the various models put virtually zero weight on $\mathcal{M}_1$ and $\mathcal{M}_2$, this is decisive evidence against independence in favour of either alternative and decisive evidence in favour of the equi-correlated model.

5·3. *Labour force participation*

The final illustration is a model of the labour force participation decision of married women in the age range 35–62. 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. As in Avery, Hansen & Hotz (1983), who analysed similar data by the method of moments, the covariates are (i) a constant, (ii) wife's education in number of grades completed and (iii) total family income excluding wife's earnings, in thousands of dollars.

We consider two multivariate probit models for this dataset. In $\mathcal{M}_1$ the correlation matrix is fully unrestricted with 21 unknown parameters. In $\mathcal{M}_2$ 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 $\beta_0 = 0$ and $B_0 = 10^{-1} I_3$. In the unrestricted model the prior on σ is represented by $\sigma_0 = 0$, a 21-vector of zeros, and $G_0^{-1} = 0.5 I_{21}$; for the restricted model it is represented by $\sigma_0 = 0$, a scalar, and $G_0^{-1} = 0.5$. The Markov chain Monte Carlo simulation algorithm is run for 10 000 cycles beyond a transient phase of 500 iterations. As a result of the large dimension of σ in model $\mathcal{M}_1$, 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. The value of τ for the first three blocks of σ is 1·5 and that for the fourth block is 2·0. In the case of $\mathcal{M}_2$, proposal values are also generated by the tailored independence chain, but with $\tau = 8$.

For the marginal likelihood calculation in $\mathcal{M}_1$, the posterior density $\pi(\sigma^* \mid y)$ at σ^*, the posterior mean, is estimated from

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

where each of the conditional ordinates is estimated by kernel smoothing of the simulations from 10 000 values of σ_i generated from $\pi(\sigma_i \mid y, \sigma_1^*, \ldots, \sigma_{i-1}^*)$ 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 normalising constant of the prior density of σ at σ^* from 10 000 draws, and the likelihood contribution using (11). The calculation of $m(y \mid \mathcal{M}_2)$ is similar, except that $\pi(\sigma^* \mid y)$ is estimated directly in one pass by kernel smoothing since only one parameter in σ is involved.

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_0, β_1 and β_2 in model $\mathcal{M}_1$ are found to be -0.620 (0·234), 0·090 (0·018) and -0.003 (0·001); the modal estimates are similar and are not reported. The marginal posterior distributions of β are fairly symmetric and concentrated about their means, and autocorrelations of the sampled draws drop off very quickly, being less than 0·2 by the tenth lag for all three β's. A boxplot summary of the marginal posterior distributions of the elements of Σ is presented in Fig. 1 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 $\mathcal{M}_2$ the correlation parameter is estimated to be 0·739 with a posterior standard deviation of 0·057. 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 $\log \hat{m}(y|\mathcal{M}_1) = -1561\cdot66$ and $\log \hat{m}(y|\mathcal{M}_2) = -1600\cdot45$. The evidence in favour of the unrestricted model is thus overwhelming, and, as above, the standard errors are negligible.
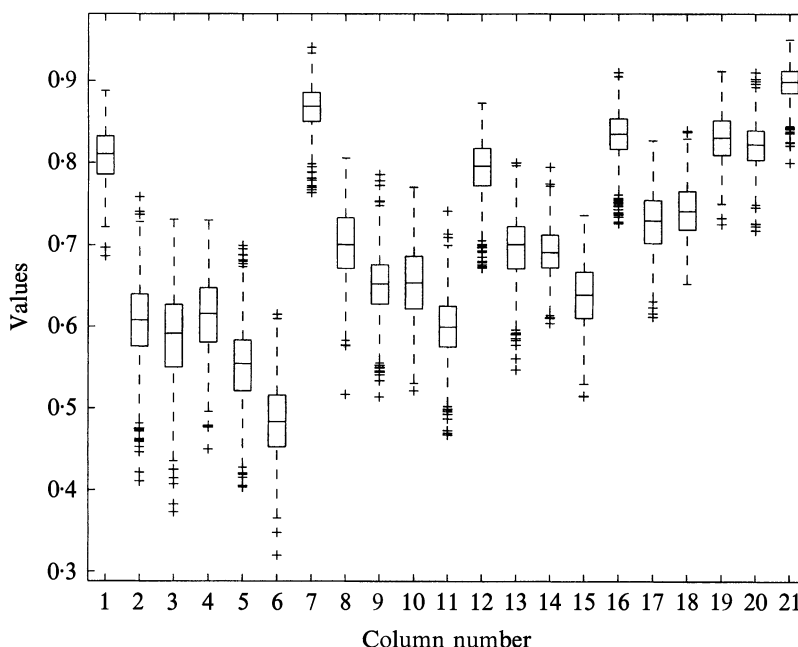

Fig. 1. Panel Survey of Income Dynamics data: Marginal posterior boxplots for elements of Σ in model $\mathcal{M}_1$. The columns correspond to a row-wise expansion of Σ, that is column 1 refers to σ_{12}, column 2 to σ_{13}, etc.

6. Discussion

Several interesting conclusions emerge from the empirical examples. First, the posterior means of β seem to be extremely robust to the specification of Σ. Secondly, although not reported above, we computed the loglikelihood contribution $\log \text{pr}(y_i|\beta^*, \Sigma^*)$ by a simple frequency approach, counting the number of times that draws from $N(X_i\beta^*, \Sigma^*)$ respect the constraints imposed by y_i, and obtained results almost identical to those in the paper. Thirdly, we achieved extremely high accuracy in computing the marginal likelihood by Chib's method, as evidenced by the very low numerical standard errors. Fourthly, in both the second and third examples, the Bayes factor decisively favoured one of the correlation specifications: an extremely large prior odds ratio would have been necessary to offset the empirical results.

REFERENCES

ALBERT, J. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.

AMEMIYA, T. (1972). Bivariate probit analysis: minimum chi-square methods. *J. Am. Statist. Assoc.* **69**, 940–4.

ASHFORD, J. R. & SOWDEN, R. R. (1970). Multivariate probit analysis. *Biometrics* **26**, 535–46.

AVERY, R. B., HANSEN, L. P. & HOTZ, V. J. (1983). Multiperiod probit models and orthogonality condition estimation. *Int. Econ. Rev.* **24**, 21–35.

CAREY, V., ZEGER, S. L. & DIGGLE, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–26.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Am. Statist. Assoc.* **90**, 1313–21.

CHIB, S. & GREENBERG, E. (1995). Understanding the Metropolis–Hastings algorithm. *Am. Statistician* **49**, 327–35.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

DEVROYE, L. (1985). *Non-Uniform Random Variate Generation*. New York: Springer Verlag.

FITZMAURICE, G. F. V. & LAIRD, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–51.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

GEWEKE, J. (1991). Efficient simulation from the multivariate normal and student-*t* distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Ed. E. Keramidas and S. Kaufman, pp. 571–8. Fairfax Station, VA: Interface Foundation of North America.

GLONEK, G. F. V. & McCULLAGH, P. (1995). Multivariate logistic models. *J. R. Statist. Soc.* B **57**, 533–46.

GREENE, W. (1997). *Econometric Analysis*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

KIEFER, N. M. (1982). Testing for dependence in multivariate probit models. *Biometrika* **69**, 161–6.

LOUIS, T. A. (1982). Finding the observed information matrix using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226–33.

MARSAGLIA, G. & OLKIN, I. (1984). Generating correlation matrices. *SIAM J. Sci. Statist. Comp.* **5**, 470–5.

McCULLOCH, R. E. & ROSSI, P. E. (1994). Exact likelihood analysis of the multinomial probit model. *J. Economet.* **64**, 207–40.

MENG, X. & RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–78.

OCHI, Y. & PRENTICE, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531–43.

ROUSSEEUW, P. & MOLENBERGHS, G. (1994). The shape of correlation matrices. *Am. Statistician* **48**, 276–9.

SCOTT, D. W. (1992). *Multivariate Density Estimation*. New York: John Wiley.

SMITH, A. F. M. & ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B **55**, 3–21.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–62.

WEI, G. C. G. & TANNER, M. A. (1990). A Monte Carlo implementation of EM algorithm and the poor man's data augmentation algorithm. *J. Am. Statist. Assoc.* **85**, 699–704.