Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices

Aline Talhouk[1] *, Arnaud Doucet[2], and Kevin Murphy[1]
[1]Department of Statistics, University of British Columbia
[2]Department of Statistics, University of Oxford

August 18, 2011

Abstract

We propose a Bayesian approach for inference in the multivariate probit model, taking into account the association structure between binary observations. We model the association through the correlation matrix of the latent Gaussian variables. Conditional independence is imposed by setting some off-diagonal elements of the inverse correlation matrix to zero and this sparsity structure is modeled using a decomposable graphical model. We propose an efficient Markov chain Monte Carlo algorithm relying on a parameter expansion scheme to sample from the resulting posterior distribution. This algorithm updates the correlation matrix within a simple Gibbs sampling framework and allows us to infer the correlation structure from the data, generalizing methods used for inference in decomposable Gaussian graphical models to multivariate binary observations. We demonstrate the performance of this model and of the Markov chain Monte Carlo algorithm on simulated and real data sets.

Keywords: Bayesian inference; Correlated binary data; Gibbs sampling; Graphical models; Markov chain Monte Carlo.

1 Introduction

Examples of applications where correlated binary data arise range from the study of group randomized clinical trials to consumer behavior, panel data, sample surveys and longitudinal studies. The multivariate probit model assumes that, given a set of explanatory variables, the multivariate response is an observed indicator that some underlying Gaussian latent variables, with covariance matrix Σ, fall within certain intervals. The likelihood of the observed discrete data is obtained by integrating over the multidimensional constrained space of latent variables. In many cases, researchers are not only interested in inferring the effect of the covariates on the correlated response, but would also like to explore and account for the structure of association between the

*Research Funding provided by National Sciences and Engineering Research Council of Canada (NSERC) Canada Graham Bell Graduate Scolarship (CGS-D)

multivariate binary response. In the multivariate probit model, the association is modeled through the covariance matrix of the latent Gaussian variables. Hence, it is possible to directly appeal to the well-developed theory of covariance selection in Gaussian graphical models (Dempster 1972). The structure of association among latent variables is imposed through conditional independence by fixing certain off-diagonal elements of the inverse covariance matrix to zero.

Inference in the multivariate probit model is notoriously difficult as the likelihood is intractable and, for identifiability reasons, additional constraints must be imposed on the covariance matrix Σ. Restricting the covariance to be a correlation matrix R as in Chib and Greenberg (1998) is the standard approach. This restriction, however, adds to the computational burden as there does not exist a conjugate prior for correlation matrices. In this article, we propose an efficient computational approach for Bayesian inference in the multivariate probit model.

The first contribution of this work is to present a Markov chain Monte Carlo (MCMC) algorithm which updates the correlation matrix R within a simple Gibbs sampling approach when the prior used for R has the attractive property of being marginally uniform. This algorithm relies on a parameter expansion for data augmentation (PXDA) strategy (Liu and Wu 1999; van Dyk and Meng 2001). We expand the correlation matrix into a covariance matrix, update this covariance matrix using standard simulation steps before projecting it back to a correlation matrix. This approach was originally proposed in Liu (2001) and Lawrence et al. (2008) when the prior on R is obtained from the Jeffreys prior on Σ. Similar ideas have been pursued in Liu and Daniels (2006) and Zhang et al. (2006) but their algorithms require Metropolis-Hastings (M-H) steps. Related strategies have also been developed for multinomial probit models. For these models, the covariance Σ is also not likelihood identified and a common approach is to set the first diagonal element of the covariance matrix $\Sigma_{11} = 1$; see for example (McCulloch et al. 2000; Nobile 2000; Linardakis and Dellaportas 2003). In this context Imai and van Dyk (2005) have proposed an efficient MCMC algorithm based on a PXDA strategy which expands the constrained covariance matrix where $\Sigma_{11} = 1$ into an unconstrained covariance matrix.

The second and main contribution of this article is to extend the prior and MCMC algorithm to accommodate a sparse structure in R^{-1} using the theory of decomposable graphical models developed in Dawid and Lauritzen (1993) and Lauritzen (1996). This allows us to develop an approach to infer the structure of the decomposable graph. This generalizes the earlier work of Giudici (1996), Giudici and Green (1999) and Wong et al. (2003) to binary data and is an alternative to the model developed recently in Webb and Forster (2008). In their work, Webb and Forster (2008) parametrize R^{-1}, the precision matrix in terms of its Cholesky decomposition $R^{-1} = \Psi\Psi'$ where Ψ is an upper triangular matrix with diagonal elements equal to 1. The elements of Ψ can be interpreted as the regression coefficients obtained by regressing the latent variable on its predecessors. The disadvantage of using this approach is that the resulting prior is not invariant w.r.t the ordering of the variables. Dobra et al. (2004) propose an algorithm to search over possible orderings, however this becomes very computationally expensive in high dimensions. Webb and Forster (2008) present a reversible jump MCMC algorithm that allows for switching in the orderings of the variables but this adds to the computational complexity.

The rest of the paper proceeds as follows: In Section 2, we detail the multivariate probit model using a Bayesian approach. In Section 3, we propose an efficient MCMC algorithm which draws the correlation matrix jointly. In Section 4, we extend the model and the MCMC algorithm to incorporate a conditional independence structure on the correlation matrix. We also show how to perform model selection on the space of correlation matrices, treating the structure as an unknown parameter. Section 5 presents some simulation results and an application to two real data sets.

2 The Multivariate Probit Model

2.1 Introduction

Let $Y_i = (Y_{i1}, \dots, Y_{iJ})$ denote the J-dimensional vector of observed binary 0/1 responses on the ith subject $(1 \leqslant i \leqslant n)$, X is a $n \times p$ design matrix, β is a $p \times J$ matrix of regression coefficients and $R = (r_{ij})$ a $J \times J$ correlation matrix. In the multivariate probit model, conditioned on β, R and X, we have

$$\mathrm{pr}(Y_i = y_i \mid X, \beta, R) = \int_{A_{iJ}} \dots \int_{A_{i1}} \mathcal{N}_J(u; 0, R) du \tag{1}$$

where A_{ij} is the interval $[(X\beta)_{ij}, \infty)$ if $y_{ij} = 1$ and $(-\infty, (X\beta)_{ij})$ otherwise and $\mathcal{N}_k(u; m, \Sigma)$ is the density of a k-variate Gaussian distribution with argument u, mean vector m and covariance matrix Σ. A parameterization in terms of a covariance matrix in (1) is not likelihood identified. To illustrate this, assume a covariance matrix $\Sigma = DRD$ where R is a correlation matrix and $D = \mathrm{diag}\,(d_1, ..., d_J)$ with $d_i > 0$, then it is easy to verify that $\mathrm{pr}(Y_i = y_i \mid X, \gamma, \Sigma) = \mathrm{pr}(Y_i = y_i \mid X, \beta, R)$ where $\gamma = \beta D$. Therefore we restrict the covariance matrix Σ to be a correlation matrix R as in Chib and Greenberg (1998).

From a computational viewpoint, it is more convenient to introduce the multivariate probit model using Gaussian latent variables. We denote by $\mathcal{N}_{k,l}\,(U;\, M,\, \Omega,\, \Pi)$ the matrix-variate normal density of argument U given by

$$(2\pi)^{-kl/2}\, |\Omega|^{-\frac{k}{2}}\, |\Pi|^{-\frac{l}{2}} \exp\left(-\frac{1}{2} \mathrm{tr}\left[\Omega^{-1} (U - M)' \,\Pi^{-1} (U - M) \right] \right)$$

where U and M are $k \times l$ matrices, Ω is $l \times l$ and Π is $k \times k$. Let $Z = (Z_1' \cdots Z_n')'$ denote the $n \times J$ matrix of latent variables such that

$$\pi(Z \mid \beta,\, R) = \mathcal{N}_{n,J}\,(Z;\, X\beta,\, R,\, I_n) \tag{2}$$

where I_n is the $n \times n$ identity matrix. The response Y_{ij} is 1 or 0 according to the sign of the corresponding Z_{ij}; that is

$$Y_{ij} = \mathbb{I}(Z_{ij} \geqslant 0), \quad j = 1, \dots, J \tag{3}$$

where $\mathbb{I}\,(A)$ is the indicator function of the event A. Using this Gaussian latent representation, it is easy to check that the probability in (1) can be rewritten as

$$\mathrm{pr}(Y_i = y_i \mid X, \beta, R) = \int_{B_{iJ}} \dots \int_{B_{i1}} \mathcal{N}_J(Z_i; (X\beta)_i\,, R) dZ_i$$

where $(X\beta)_i$ is the i^{th} row of $X\beta$, B_{ij} is the interval $[0, \infty)$ if $y_{ij} = 1$ and $(-\infty, 0)$ otherwise. We define by $B_i = B_{i1} \times \cdots \times B_{iJ}$ the set-valued inverse of the mapping in (3).

2.2 Prior specification

We follow a standard Bayesian conjugate approach for β by assuming that it follows a matrix-variate Gaussian density

$$\pi(\beta \mid R) = \mathcal{N}_{p,J}(\beta;\, 0_{p\times J},\, R,\, \Psi) \tag{4}$$

where Ψ is a $p \times p$ matrix of hyperparameters representing the prior belief on the dispersion of β.

A prior specification for R, on the other hand, is not straightforward as there is no conjugate prior for correlation matrices. Several priors on R have been proposed in the literature. In an influential paper, Chib and Greenberg (1998) use a jointly uniform prior for the correlation $\pi(R) \propto 1$. This prior poses several problems. Computationally, it results in a posterior that is not easy to sample. Chib and Greenberg (1998) propose a MH random walk algorithm to sample the correlation matrix drawing the correlation coefficients in blocks. The resulting proposal is not guaranteed to be a correlation matrix and moreover, as with random walk algorithms in general, the mixing is slow in high dimensions. Furthermore, Barnard et al. (2000) show that this prior tends to push marginal correlations to zero in high dimensions, making it very informative marginally. Similarly, the Jeffreys prior used by Liu (2001) tends to push marginal correlations to ± 1 in high dimensions.

Barnard et al. (2000) propose a marginally uniform prior for R given by

$$\pi(R) \propto |R|^{\frac{J(J-1)}{2}-1} (\prod_{i=1}^{J} |R_{ii}|)^{-(J+1)/2} \tag{5}$$

where R_{ii} is the principal submatrix of R. Despite not being conjugate, this prior will be adopted here. Combined with a PXDA strategy, it allows us to develop an efficient MCMC algorithm presented in Section 3.

3 Bayesian Computation using Parameter Expansion

Given n observations $y = (y_1, \dots, y_n)$ and the prior density $\pi(\beta, R)$, we are interested in the posterior density

$$\pi(\beta, R \mid y) \propto \pi(R)\pi(\beta \mid R) \prod_{i=1}^{n} \mathrm{pr}(y_i \mid X, \beta, R).$$

This expression is not convenient as we are not even able to easily evaluate the likelihood. The standard approach to bypass the calculation of the likelihood consists of introducing explicitly the latent variables Z and to consider instead

$$\pi(\beta, R, Z \mid y) \propto \pi(R)\pi(\beta \mid R)\pi(Z \mid \beta, R) \prod_{i=1}^{n} \mathbb{I}\left(Z_i \in B_i\right). \tag{6}$$

The full conditional density of the latent variables Z factorizes

$$\pi(Z \mid y, \beta, R) = \prod_{i=1}^{n} \pi(Z_i \mid y, \beta, R)$$

where

$$\pi(Z_i \mid y, \beta, R) \propto \mathcal{N}_J(Z_i \mid (X\beta)_i\,, R)\, \mathbb{I}\,(Z_i \in B_i)$$

is a truncated multivariate Gaussian. For $J = 1$, we can sample from it directly. Otherwise, we sample approximately from it by cycling through a series of univariate truncated Gaussians (Geweke 1991). In each step Z_{ij} is simulated from $\pi\,(Z_{ij}|y, Z_{i,-j}, \beta, R)$, which is a univariate Gaussian distribution truncated to $[0, \infty)$ if $y_{ij} = 1$ and to $(-\infty, 0)$ if $y_{ij} = 0$. To sample the univariate truncated Gaussians, we use the algorithm of Robert (1995).

To sample (β, R), we adopt a parameter expansion framework. We define a transformation on the latent variables $W = ZD$ where $D = \text{diag}(d_1, \ldots, d_J)$ is the expansion parameter with $d_i > 0$. It follows that

$$\pi(W \mid \beta, R, D) = \mathcal{N}_{n,J}\,(W;\, X\beta D,\, DRD,\, I_n)\,.$$

We then define the new posterior

$$\pi(\beta, R, D, W \mid y) \propto \pi(R)\pi(\beta \mid R)\pi(D \mid R)\pi(W \mid \beta, R, D) \prod_{i=1}^{n} \mathbb{I}\,(W_i \in B_i)\,. \tag{7}$$

Due to the lack of identifiability mentioned in Section 1, the marginals in (β, R) under (6) and (7) are similar whatever being the density $\pi(D \mid R)$; hence the somewhat abusive notation consisting of using the symbol π for both (6) and (7) is justified. We set $\pi(D \mid R) = \prod_{i=1}^{J} \pi(d_i \mid R)$ with

$$d_i^2 \sim \mathcal{IG}\left((J+1)/2, r^{ii}/2\right) \tag{8}$$

where the notation r^{ij} is used to refer to the ij^{th} element of R^{-1} and $\mathcal{IG}\,(a, b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. The reason for this choice is that it follows from Barnard et al. (2000) that

$$\Sigma = DRD \sim \mathcal{IW}\,(2,\, I_J)$$

when R follows (5) and where $\mathcal{IW}\,(\nu,\, \Omega)$ denotes the inverse Wishart distribution with degrees of freedom ν and inverse scale matrix Ω defined as in Dawid and Lauritzen (1993)

$$\frac{\left|\frac{\Omega}{2}\right|^{\left(\frac{\nu+J-1}{2}\right)}}{\Gamma_J\left(\frac{\nu+J-1}{2}\right)} |\Sigma|^{-(\nu+2J)/2} \text{exp}\left(-\frac{1}{2}\text{tr}\left[\Sigma^{-1}\Omega\right]\right) \tag{9}$$

where Γ_J denotes the multivariate Gamma function. Note that this convention for the Wishart distribution is different from the one used in Barnard et al. (2000).

Hence to sample jointly from (β, R, D) from $\pi(\beta, R, D \mid y, W) = \pi(\beta, R, D \mid W)$, we can simply perform a change of variables $\Sigma = DRD$ and $\gamma = \beta D$, sample (Σ, γ) according to their resulting full conditional density

$$\pi(\gamma, \Sigma \mid W) = \pi(\Sigma \mid W)\pi(\gamma \mid W, \Sigma)$$

then compute $R = D^{-1}\Sigma D^{-1}$, $\beta = \gamma D^{-1}$ where $d_i = (\sigma_{ii})^{1/2}$, where σ_{ii} is the i-th diagonal element of Σ. It is easy to check that

$$\pi(\Sigma \mid W) = \mathcal{IW}(\Sigma;\, 2+n,\, W'W + I_J - M'\Xi^{-1}M) \tag{10}$$

and

$$\pi(\gamma \mid W, \Sigma) = \mathcal{N}_{p,J}(\gamma;\, M,\, \Sigma,\, \Xi) \tag{11}$$

where $\Xi^{-1} = X'X + \Psi^{-1}$ and $M = \Xi X'W$.

To summarize the algorithm, assume we have (β, R, Z) then we update these parameters as follows:

- Sample $Z_{ij} \sim \pi\left(Z_{ij} | Z_{i,-j}, \beta, R\right)$ for all i, j.
- Sample D using (8) and compute $W = ZD$.
- Sample (Σ, γ) using (10)-(11).
- Compute $R = D^{-1}\Sigma D^{-1}$, $\beta = \gamma D^{-1}$ where $d_i = (\Sigma_{ii})^{-1/2}$.

4 Extension to Structured Correlation Matrix

4.1 Decomposable Gaussian Graphical Models

We model the association between the J columns of the $n \times J$ matrix of binary responses $Y = \left(Y^1, \ldots, Y^J\right)$ via the correlations between the J columns of the $n \times J$ matrix of latent Gaussian variables $Z = \left(Z^1, \ldots, Z^J\right)$ distributed according to (2). When dealing with high dimensional problems, imposing a structure of association is helpful both conceptually and computationally. From a computational viewpoint, this reduces the number of free parameters to be estimated compared with the saturated model. Hence, for a properly specified structure, it will result in more accurate and efficient estimates of parameters. Furthermore, researchers are often able to identify or reject certain parsimonious structures based on their conceptual understanding of the data and their subject matter expertise.

In the context of Gaussian latent variables Z, conditional independence assumptions are encoded by the inverse correlation R^{-1}, such that $r^{ij} = 0$ implies that Z^i and Z^j are conditionally independent, given all the other variables in the model. It allows us to filter out indirect dependence caused by intermediate or confounding variables. Gaussian graphical models provide a convenient framework for imposing a conditional independence structure of association between variables. As noted in Webb and Forster (2008), the interpretation of conditional independence on the latent variables Z does not translate to the observed binary variables, so that $Z^1 \perp\!\!\!\perp Z^2 \mid Z^3$ does not imply $Y^1 \perp\!\!\!\perp Y^2 \mid Y^3$, but does imply $Y^1 \perp\!\!\!\perp Y^2 \mid Z^3$.

For the time being, we will consider the scaled latent Gaussian variables $W = ZD$ which follow a matrix-variate normal distribution $\mathcal{N}_{n,J}\left(X\gamma,\, \Sigma,\, I_n\right)$. Gaussian graphical models are a class of undirected graphical models that provide a graphical representation of the inverse covariance matrix Σ^{-1}, characterizing conditional independence between Gaussian variables. Graphical models have been extensively used in Bayesian inference on Gaussian data (Giudici (1996), Giudici and Green (1999), and Wong et al. (2003)). For a complete account of graphical model theory, the reader is referred to Lauritzen (1996). In this paper, we restrict our attention to decomposable graphs. We first present key graph theory concepts and definitions necessary to understand the statistical model developed here.

An undirected graph is a pair $G = (V, E)$, where V is a set of vertices representing variables and E, the edge-set, is a subset of the set of unordered distinct pair of vertices. Visually, each vertex i is a node representing the random variable i and an edge $(i, j) \in E$ is an undirected edge connecting nodes i and j unless they are conditionally independent. A *subgraph* is a graph which has as its vertices some subset of the vertices of the original graph. A graph or a subgraph is *complete* or fully connected if there is an edge connecting any two nodes. A *clique* is a complete subgraph. A set S is said to *separate* A from B if all paths from A to B go through S. Subgraphs (A, B, S) form a decomposition of G if $V = A \cup B$, $S = A \cap B$, where S is complete and separates A from B. A sequence of subgraphs that cannot be further decomposed are the *prime components* of a graph. A graph is said to be *decomposable* if every prime component is complete. A distribution $p(.)$ is Markov with respect to a graph G, if for any decomposition (A, B) of G, $A \perp\!\!\!\perp B | A \cap B$, where $A \cap B$ is complete and separates A from B.

Let Σ be a $J \times J$ structured covariance matrix consistent with the decomposable graph G, then the multivariate Gaussian distribution is Markov with respect to any decomposable graph G; that is

$$\mathcal{N}_{n,J}\left(W;\, X\gamma,\, \Sigma,\, I_n\right) = \frac{\prod_{P \in \mathcal{P}} \mathcal{N}_{n,|P|}(W^P;\, X\gamma^P,\, \Sigma^P,\, I_{|P|})}{\prod_{S \in \mathcal{S}} \mathcal{N}_{n,|S|}(W^S;\, X\gamma^S,\, \Sigma^S,\, I_{|S|})} \tag{12}$$

where $\mathcal{P}$ is the set of all prime components and $\mathcal{S}$ is the set of all separators of G whereas, for each component U, $|U|$ denotes its cardinality, γ^U is the corresponding $p \times |U|$ matrix of regression coefficients and Σ^U is the covariance matrix of component U. Similarly the prior on γ satisfies

$$\mathcal{N}_{p,J}(\gamma;\, 0,\, \Sigma,\, \Psi) = \frac{\prod_{P \in \mathcal{P}} \mathcal{N}_{p,|P|}(\gamma^P;\, 0,\, \Sigma^P,\, \Psi)}{\prod_{S \in \mathcal{S}} \mathcal{N}_{p,|S|}(\gamma^S;\, 0,\, \Sigma^S,\, \Psi)} \tag{13}$$

Finally the hyper-inverse Wishart distribution, defined by Dawid and Lauritzen (1993), is a locally conjugate family of Markov probability distributions for structured covariance matrices on decomposable graphs. We say that Σ follows a hyper-inverse Wishart distribution denoted by $\mathcal{HIW}(b,\, K)$ with degrees of freedom $b > 0$ and location parameter K when

$$\mathcal{HIW}(\Sigma;\, b,\, K) = \frac{\prod_{P \in \mathcal{P}} \mathcal{IW}(\Sigma^P;\, b,\, K^P)}{\prod_{S \in \mathcal{S}} \mathcal{IW}(\Sigma^S;\, b,\, K^S)} \tag{14}$$

It is easy to check that the posterior distribution of Σ, γ given W satisfies

$$\Sigma \mid W \sim \mathcal{HIW}(b_n, K_n) \tag{15}$$

and γ is distributed according to (11) with $b_n = b + n$, $K_n = W'W + K - M'\Xi^{-1}M$.

4.2 Adaptation to the Probit Model

In the multivariate probit model, the latent variables $Z = \left(Z^1, \ldots, Z^J\right)$ are distributed according to (2). Marginally, we would like the correlation r_{ij} to be uniform on $(-1, 1)$ if variable Z^i is conditionally dependent on Z^j and we would like r^{ij}, the ij^{th} element of R^{-1}, to be null if they are conditionally independent; that is

$$\begin{cases} r^{ij} = 0 & \text{if } Z^i \perp\!\!\!\perp Z^j \mid Z^{-i,-j}; \\ r_{ij} \sim U(-1, 1) & \text{otherwise.} \end{cases} \tag{16}$$

To achieve this, we can simply select the following prior

$$\pi(R) = \frac{\prod_{P \in \mathcal{P}} \pi(R^P)}{\prod_{S \in \mathcal{S}} \pi(R^S)}$$

where the prior $\pi(R^U)$ on the prime component or separator R^U satisfies

$$\pi(R^U) \propto |R^U|^{\frac{|U|(|U|-1)}{2}-1} \left(\prod_{i \in U} |R^U_{ii}| \right)^{-\frac{(|U|+1)}{2}} \tag{17}$$

where R^U_{ii} is the principal submatrix of R^U. Similarly to the saturated case discussed in section 3, we can show that if we sample D according to (8) then

$$\Sigma = DRD \sim \mathcal{HIW}\,(2,\, I_J)$$

The PXDA algorithm can now be straightforwardly extended to accommodate a structured inverse correlation matrix. It proceeds exactly as in Section 3 with the only difference being that Σ is now sampled according to an hyper-inverse Wishart distribution (15) where $b_n = 2 + n$ and $K_n = W'W + I_J - M'\Xi^{-1}M$. We sample from the hyper-inverse Wishart distribution by using the junction tree decomposition of the graph and working directly on the prime component level; see Carvalho et al. (2007) for details.

4.3 Model Selection

For the time being, we have assumed that the graph structure G of the inverse correlation matrix is known. If it is unknown, it is possible to estimate it directly from the data by sampling it along with the other unknown parameters. While it is more desireable to use the observed data to learn the graph structure, it is computationally intractable in this case. The Gaussian latent data can be conveniently used instead. In decomposable Gaussian graphical models, the hyper-inverse Wishart distribution is conjugate to the Gaussian likelihood. This facilitates the computation of the marginal likelihood of the graph because the covariance matrix and the regression coefficients can be integrated out in the expanded parameter space. We have

$$\pi(W \mid G) = \int_{\Sigma^{-1} \in M(G)} \int_{\gamma} \pi(W \mid \Sigma, \gamma, G)\pi(\Sigma, \gamma \mid G) d\gamma d\Sigma \tag{18}$$

where $M(G)$ is the set of all possible symmetric positive definite matrices consistent with G. It is easy to check that

$$\pi(W \mid G) = (2\pi)^{-nJ/2} \frac{h(G, b, K)}{h(G, b_n, K_n)} \tag{19}$$

where h denotes the normalizing constant of the hyper-inverse Wishart given by

$$h(G, b, K) = \frac{\prod_{P \in \mathcal{P}} \left|\frac{K^P}{2}\right|^{\left(\frac{b+|P|-1}{2}\right)} \Gamma_{|P|}\left(\frac{b+|P|-1}{2}\right)^{-1}}{\prod_{S \in S} \left|\frac{K^S}{2}\right|^{\left(\frac{b+|S|-1}{2}\right)} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)^{-1}} \tag{20}$$

The posterior distribution over graphs can be computed as

$$\pi(G \mid W) = \frac{\pi(W \mid G)\pi(G)}{\sum_{i=1}^{D_c} \pi(W \mid G)\pi(G)} \tag{21}$$

where D_c is the total number of decomposable graphs. The marginal likelihood $\pi(W \mid G)$ can be computed using (19), and we assume that all decomposable graphs are equally likely, that is $\pi(G) = 1/D_c$. This prior, as noted by Armstrong et al. (2009), will tend to favour graphs of medium size, where the size of a graph is given by the number of its edges. We use it here for computational convenience but more flexible priors can be defined if necessary; see Bornn and Caron (2011).

Sampling the graph from its posterior distribution in low dimension can be done by enumerating all possible decomposable graphs. For higher dimensions (greater than 4), it is more computationally efficient to appeal to the well developed MCMC methods for sampling decomposable graphs. The Monte Carlo sampling algorithm suggested in Giudici and Green (1999) to sample from the posterior distribution starts with a graph G^c and proceeds by randomly selecting the index of two vertices such that the addition or the deletion of an edge between these two vertices would result in a decomposable graph. The proposed resulting graph is denoted by G^p, it is then accepted according to the following MH acceptance probability

$$\min\left\{1, \frac{\pi(W \mid G^p)\pi(G^p)}{\pi(W \mid G^c)\pi(G^c)}\right\}. \tag{22}$$

The PXDA algorithm in this case proceeds in similar ways to the one outlined previously, except at each iteration, a graph structure G is sampled from its posterior distribution conditional on the latent variables. Then, given the sampled graph, the covariance matrix is sampled from a hyper-inverse Wishart exactly as in the previous section.

To summarize, at each iteration, the parameters are updated as follows:

- Sample $Z_{ij} \sim \pi\left(Z_{ij}|Z_{i,-j}, \beta, R\right)$ for all i, j as in section 3.
- Sample D using (8) and compute $W = ZD$.
- Propose a new decomposable graph G^p by adding or deleting an edge and set $G = G^p$ with probability given in (22).
- Sample $\Sigma \mid G$ according to (15) with $b_n = 2 + n$ and $K_n = W'W + I_J - M'\Xi^{-1}M$.
- Sample $\gamma \mid \Sigma$ according to (11).
- Compute $R = D^{-1}\Sigma D^{-1}$, $\beta = \gamma D^{-1}$ where $d_i = \left(\sigma_{ii}\right)^{1/2}$.

5 Simulation Results

5.1 Synthetic Data

The aim of the first simulation is to demonstrate the ability of the algorithm to recover the true graph structure underlying observed binary variables given a design matrix of covariates. We generate $n = 500$ independent data points from the model where X is an $n \times p$ design matrix with $p = 2$ whose entries were drawn from a uniform distribution in $[-0.5, 0.5]$, $J = 5$ and the regression coefficients β were set to

$$\beta = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 \\ 2 & 2 & 2 & 2 & 2 \end{pmatrix} \tag{23}$$

The correlation matrix R used to generate the Gaussian observations was set at

$$R = \begin{pmatrix} 1.000 & 0.000 & 0.000 & -0.491 & 0.000 \\ 0.000 & 1.000 & -0.296 & 0.000 & 0.116 \\ 0.000 & -0.296 & 1.000 & 0.000 & -0.392 \\ -0.491 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.116 & -0.392 & 0.000 & 1.000 \end{pmatrix} \tag{24}$$

corresponding to the inverse correlation

$$R^{-1} = \begin{pmatrix} 1.318 & 0.000 & 0.000 & 0.647 & 0.000 \\ 0.000 & 1.096 & 0.325 & 0.000 & 0.000 \\ 0.000 & 0.325 & 1.278 & 0.000 & 0.464 \\ 0.647 & 0.000 & 0.000 & 1.318 & 0.000 \\ 0.000 & 0.000 & 0.464 & 0.000 & 1.182 \end{pmatrix} \tag{25}$$

For the prior (4) on β, we set $\Psi = 10^5 I_p$. We ran the MCMC algorithm with model selection as described in Section 4.3 on the simulated data and $N = 50,000$ Monte Carlo samples were collected. Of the 822 possible decomposable structures for a graph with five nodes, the algorithm explored 271 graphs. From Figure 1, we see that the true graph structure underlying the data was the most probable structure sampled by the algorithm and accounts for approximately 15% of the posterior mass. In Figure 2, we show an image of the matrix of posterior edge marginal probabilities. These probabilities were approximated by the empirical frequencies an edge between any two nodes was included in the sampled graphs. When compared to the image of the true inverse correlation matrix underlying the data, all the non-zero entries in R^{-1} correspond to edges that have considerably higher posterior marginal probabilities. In order to assess convergence, multiple runs were undertaken with different overdispersed starting values and very similar results were obtained in all cases.

In order to explore the benefits of performing model selection in comparison to fitting a saturated model, we ran a simulation comparing the two alternatives. In this simulation 50 different data sets were generated using the same graph structure as above, but sampling a different correlation matrix each time. For simplicity, no covariates were added in this case. For each data set, we fit both the saturated model and use the model selection procedure. To evaluate the posterior estimators provided in each case, the entropy loss was computed for each Monte Carlo sample using

$$L(\hat{R}, R) = \operatorname{trace}(\hat{R}R^{-1}) - \log|\hat{R}R^{-1}| - J$$

where $\hat{R}$ is the Monte Carlo sample estimate of the correlation matrix.

Figure 3 shows the box plots of the average entropy loss recorded for each data set. We can see that using the model selection procedure resulted in a significantly smaller entropy loss (paired t-test pval <0.01).

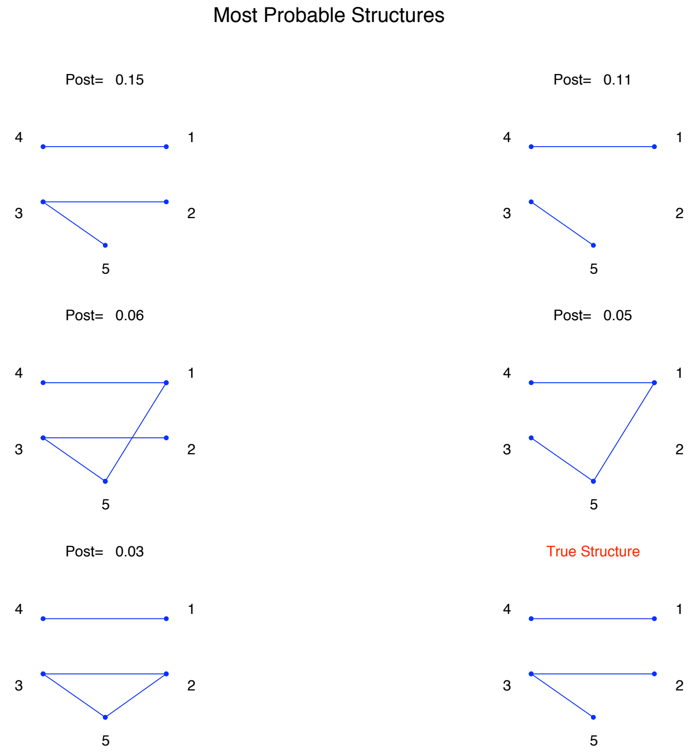

Figure 1: Most probable graphical structures and estimated posterior probabilities for simulated data.

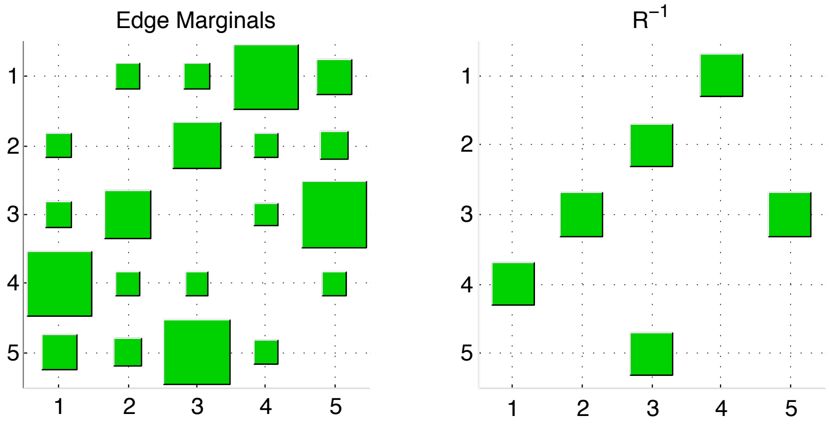

Figure 2: Matrix of posterior edge marginal probabilities, simulated data. Here the size of the square is indicative of the magnitude of the posterior edge marginal probability and therefore the larger the square the higher the probability that an edge exists between the corresponding nodes.

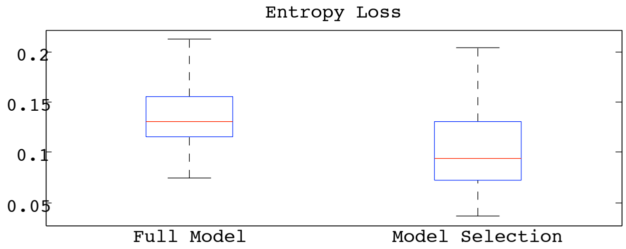

Figure 3: Boxplot of the entropy loss over 50 datasets obtained using a fixed saturated correlation matrix (left) versus model selection (right).

5.2 "Women and Mathematics" Data Set

We further illustrate our method by applying the model selection algorithm to the "Women and Mathematics" data set (Fowlkes et al. 1988). This data was analyzed by Tarantola (2004) and by Madigan and Raftery (1994) using the MC^3 algorithm, to learn the structure of a discrete graph. The primary aim is to study the attitude of 1190 New Jersey high school students toward mathematics. The six binary variables collected are:

X_1 Lecture Attendance (Attend/Did not Attend)

X_2 Gender (Male/Female)

X_3 School Type (Urban/Suburban)

X_4 "I will be needing Mathematics in my future work" (Agree/Disagree)

X_5 Subject Preference (Math/Science vs. Liberal Arts)

X_6 Future Plans (College/Job)

We ran the algorithm for $N = 200,000$ iterations. Multiple runs with different starting values and very similar results were observed in all cases. Our method identifies six structures with probabilities higher that 3% (0.03), accounting for 38% of the total probability mass. The most probable structures are depicted in figure 4.

All of the probable graphs chosen by our algorithm show that X_1, lecture attendance, was independent conditional on all the other variables in the graph. Similarly, all of the probable models include edges between $[X_5, X_2, X_4]$, showing association between gender, subject preference and whether they will need mathematics in their future. These results are consistent with those found by others. Our method also detects with high probability the clique $[X_6, X_5, X_4]$, suggesting a conditional association between future plans, subject preference, and whether the respondents thought there will be needing mathematics in their future. Moreover, a relationship between X_6, future plans and X_3, type of school chosen, was detected. Comparing our results with those of Tarantola (2004), we note that while their most probable graph does not correspond to the most probable one obtained here, some of their most probable graphs are identical to the ones we obtained. Furthermore, if we compare the edge marginal probability, we see that in their hierarchical models, with a certain value of the hyperparameters ($\lambda_0 = 64$ in the non-hierarchical

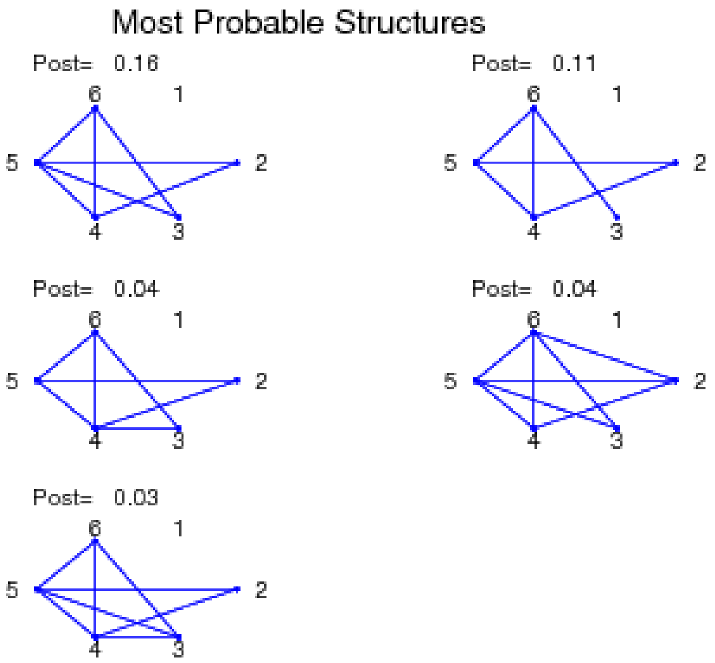

Figure 4: Most probable graphical structures and estimated posterior probabilities for the "Women and Mathematics" data set.

model and $f = 64$ in the hierarchical model), the edge marginals they obtained are very similar to the ones we have (see table 1). The graph they report in their paper corresponds to the hierarchical model with $f = 1$.

Edges	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)	(3,2)	(4,2)	(5,2)
This paper	0.11	0.10	0.14	0.13	0.13	0.04	0.90	1.00
Non Hierarchical $\lambda_0 = 64$	0.10	0.12	0.15	0.13	0.15	0.11	1.0	1.0
Hierarchical $f = 64$	0.12	0.15	0.17	0.15	0.16	0.12	1.0	1.0
Edges	(6,2)	(4,3)	(5,3)	(6,3)	(5,4)	(6,4)	(6,5)	
This paper	0.20	0.22	0.56	1.00	1.00	1.00	0.98	
Non Hierarchical $\lambda_0 = 64$	0.02	0.81	0.33	1.0	1.0	0.80	0.79	
Hierarchical $f = 64$	0.03	0.80	0.43	1.0	1.0	0.89	0.88	

Table 1: Table of edge marginal probabilities comparing results obtained by the algorithm developed in this paper and those by MC^3 algorithm as outlined in Tarantola (2004)

5.3 Six Cities Data Set

In the last example we consider the data set used by Chib and Greenberg (1998), a subset of the Six Cities study, a longitudinal study of the health effects of air pollution. The data under consideration consists of repeated binary measures of wheezing status of 537 children from Stuebenville, Ohio. The objective is to model the probability of wheeze status over time as a function of the mother's smoking status during the first year of the study and the child's age. The response variable of interest y_{ij} indicates the wheezing status of child i at $j = (1, 2, 3, 4)$ corresponding to age $(7, 8, 9, 10)$. The nature of the data is suggestive that there exists an association between the elements of the multivariate response variable. We are interested in assessing the effects of the mother's smoking status at onset (1=yes, 0=no) .

Table 2 breaks down the number of "cases" representing wheezing symptoms, by age group and mother smoking status. From looking at these results, it appears that the percentages of wheezing children is higher for smoker mothers at each age.

Mother Smoking Status	7	8	9	10
Smoker	32 (17.0%)	40 (21.3%)	36 (19.1%)	27 (14.4%)
Non Smoker	55 (15.8%)	51 (14.6%)	49 (14.0%)	36 (10.0%)
Total	87 (16.2%)	91 (16.9%)	85 (15.8%)	63 (11.7 %)

Table 2: Breakdown of wheezing cases by age group and smoking status of mother at onset. The percentages in brackets are percentages with respect to each age groups.

Similar to Lawrence et al. (2008), we use the mother's smoking status at onset as a covariate. For the prior (4) on β, we set $\Psi = 10^5 I_p$. We fit the probit model using a saturated correlation matrix as in Lawrence et al. (2008) but also estimating the structure of association from the data. We generated $N = 50,000$ samples and used $3,000$ samples as burn-in. Table 3 presents the posterior means for β and R. These results are compared to those presented in Lawrence et al. (2008).

From the results in Table 3, we note that the results obtained using our proposed method and prior do not differ significantly from those presented in Lawrence et al. (2008). The estimates for the regression coefficients are robust to the correlation structure as well as the prior specification. The marginal posterior correlations are slightly smaller under the marginal uniform priors which is consistent with the fact that Jeffreys' prior tends to push marginal correlations to ±1. Furthermore, in order to compare the performance of the algorithm from a computational standpoint we look at the plots of the autocorrelation function in Figure 5 and we compare that to Figure 1 in Lawrence et al. (2008). We can see that the proposed algorithm fairs well in comparison and shows significant autocorrelation to only about 10 to 15 lags compared to 25 to 30 lags for the algorithm in Lawrence et al. (2008) which in turn performs significantly better than the algorithm in Chib and Greenberg (1998).

Because our method performs correlation selection, whereby the graph is sampled alongside the other parameters, we can obtain a posterior distribution over graph structures. Figure 6 depicts the most probable graphs that were obtained. The saturated model used by Chib and Greenberg (1998) and Lawrence et al. (2008) was the second most probable graph with posterior probability of 0.26. The most probable graph structure (43% of posterior mass) was one where the wheezing status of the child age 9 is conditionally independent of the wheezing status at age 7.

Finally, in order to assess the fit of the model to the data, we use posterior predictive checks suggested in Gelman et al. (2003). Posterior predictive checks are useful for detecting systematic differences between the model and the observed data. They are done by generating replicated

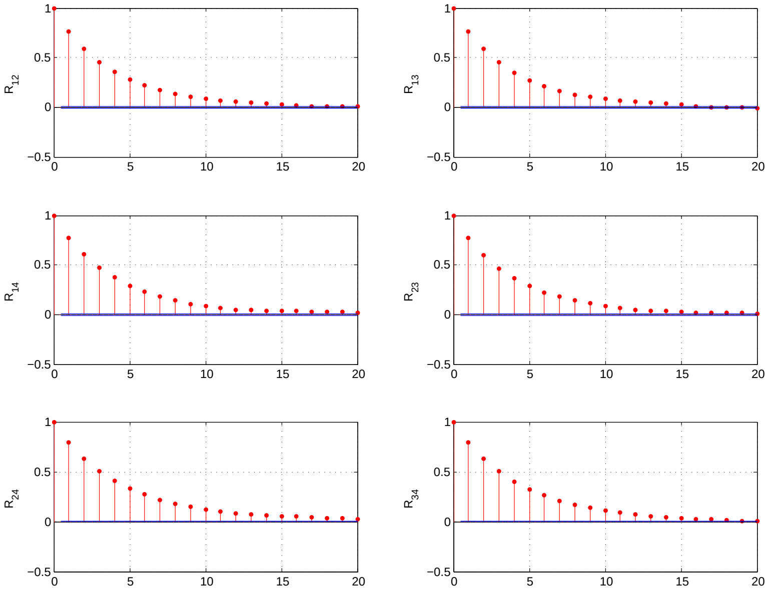

Figure 5: Autocorrelation function for the marginal correlation parameters obtained from the Six Cities example under the saturated assumption.

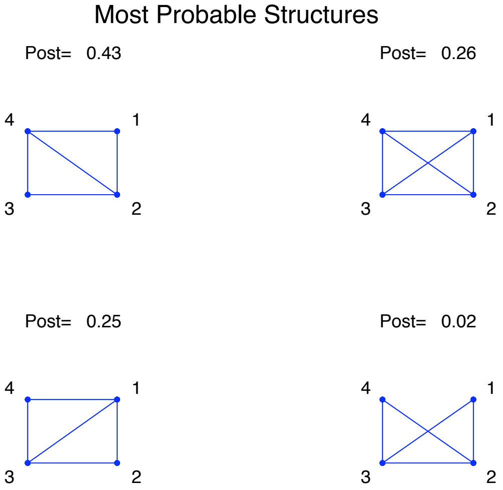

Figure 6: The four most probable graph structures, with their associated posterior probability

	Saturated		Model Selection
	Jeffreys	Marginally uniform	Marginally uniform
β_{11}	-0.983	-0.922	-0.922
β_{12}	0.011	0.032	0.031
β_{21}	-1.029	-0.929	-0.983
β_{22}	0.219	0.223	0.222
β_{31}	-1.054	-1.012	-1.013
β_{32}	0.166	0.181	0.182
β_{41}	-1.235	-1.180	-1.181
β_{42}	0.150	0.167	0.169
r_{12}	0.592	0.498	0.495
r_{13}	0.535	0.455	0.422
r_{14}	0.573	0.487	0.485
r_{23}	0.700	0.588	0.589
r_{24}	0.571	0.494	0.469
r_{34}	0.641	0.548	0.552

Table 3: Posterior means of the parameters of interest. The first two columns correspond to using the saturated model for the Jeffreys's prior of Lawrence et al. (2008) and the marginally uniform prior. The last column corresponds to using a marginally uniform prior and performing model selection.

data sets from the posterior predictive distribution of the statistical model. The replicated data is then compared with the observed data set with respect to some features of interest. For this example, we took $L = 25,000$ samples from the joint posterior distribution, and generated a data set Y^l for each (β^l, R^l), $l = 1, \dots, L$.

The resulting binary data set is then compared with the observed data with respect to the following three quantities

- % of kids that never experienced wheezing at any age.
- % of kids that experienced wheezing at every age.
- % of kids that did not experience wheezing at onset, however, once they have they have continued for more than 2 time points.

Each of the above measures, $T(y)$ was computed under the observed data and compared to the quantity obtained under the replicated artificial data. The Bayesian p-value is used to assess whether differences are statistically significant. The Bayesian p-value is given by $\Pr(T(Y^{rep}) > T(y)|y)$ and is estimated from simulations using $\sum_{l=1}^{L} \mathbb{I}\left(T(Y^l) > T(y)\right)/L$.

The p-values are computed for the case where the correlation structure is inferred from the data. The results in Table 4 suggest that the posterior predictive distribution produces data that are in reasonable agreement with the observed data. Indeed the p-values obtained are neither too close to 0 nor too close to 1.

Test Variable	$T(y)$	95% Intervals for $T(Y_{rep})$	p-value
% never exp. wheezing	66.1	[55.12 67.59]	0.08
% always exp. wheezing	3.3	[1.11 4.47]	0.21
% incident wheezing	2.4	[1.86 5.77]	0.93

Table 4: Posterior predictive statistical checks with 95% intervals and p-values, indicating that the posterior predictive distribution of the proposed model is not statistically different from the observed data

6 Discussion

This article has presented a Bayesian approach for inference in the multivariate probit model. The model allows us to impose a structure of association through a sparse inverse correlation matrix. We have proposed efficient MCMC algorithms to carry out the computations. The algorithms proposed herein to learn the conditional independence structure of association between the observed binary variables can be viewed as an extension to the binary case of methodology used for decomposable Gaussian graphical models. Simulation results and applications on two real data sets demonstrate the performance of the model and algorithm.

SUPPLEMENTAL MATERIALS

Matlab code: Matlab code is provided as well as all the necessary data, scripts and functions to replicate the figures in the paper.

References

Armstrong, H., Carter, C., Wong, K., and Kohn, R. (2009), “Bayesian Covariance Matrix Estimation Using a Mixture of Decomposable Graphical Models,” *Statistics and Computing*, 19, 303–316.

Barnard, J., McCulloch, R., and Meng, X. (2000), “Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage,” *Statistica Sinica*, 10, 1281–1311.

Bornn, L. and Caron, F. (2011), “Bayesian Clustering in Decomposable Graphs,” *Bayesian Analysis*, to appear.

Carvalho, C. M., Massam, H., and West, M. (2007), “Simulation of Hyper-inverse Wishart Distributions in Graphical Models,” *Biometrika*, 94, 647–659.

Chib, S. and Greenberg, E. (1998), “Analysis of Multivariate Probit Models,” *Biometrika*, 85, 347–361.

Dawid, A. and Lauritzen, S. (1993), “Hyper Markov laws in the Statistical Analysis of Decomposable Graphical Models,” *The Annals of Statistics*, 21, 1272–1317.

Dempster, A. (1972), "Covariance Selection," *Biometrics*, 28, 157–175.

Dobra, D., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004), "Sparse Graphical Models for Exploring Gene Expression Data," *Journal of Multivariate Analysis*, 90, 196–212.

Fowlkes, E. B., Freeny, A. E., and Landwehr, J. M. (1988), "Evaluating Logistic Models for Large Contingency Tables," *Journal of the American Statistical Association*, 83, 611–622.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. (2003), *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*, Chapman and Hall/CRC, 2nd ed.

Geweke, J. (1991), "Efficient Simulation From the Multivariate Normal and Student–T Distributions Subject to Linear Constraints," *Computing Science and Statistics: Proceedings of the Twenty–Third Symposium on the Interface, Alexandria, VA: American Statistical Association, pp.*

Giudici, P. (1996), "Learning in Graphical Gaussian Models," *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5-9,1994*, 621–628.

Giudici, P. and Green, P. (1999), "Decomposable Graphical Gaussian Model Determination," *Biometrika*, 86, 785–801.

Imai, K. and van Dyk, D. (2005), "A Bayesian Analysis of the Multinomial Probit Model using Marginal Data Augmentation," *Journal of the American Statistical Association*, 124, 311–334.

Lauritzen, S. (1996), *Graphical Models*, Clarendon Press, Oxford.

Lawrence, E., Bingham, D., Liu, C., and Nair, V. (2008), "Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion," *Technometrics*, 50, 182–191.

Linardakis, M. and Dellaportas, P. (2003), "Assessment of Athens Metro Passenger Behaviour via a Multiranked Probit Model," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 185–200(16).

Liu, C. (2001), "Bayesian Analysis of Multivariate Probit Models - Discussion on The Art of Data Augmentation by Van Dyk and Meng," *Journal of Computational and Graphical Statistics*, 10, 75–81.

Liu, J. and Wu, Y. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Liu, X. and Daniels, M. (2006), "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization," *Journal of Computational and Graphical Statistics*, 15, 897–914(18).

Madigan, D. and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

McCulloch, R., Polson, N., and Rossi, P. (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193.

Nobile, A. (2000), "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint," *Journal of Econometrics*, 99, 335–345.

Robert, C. (1995), "Simulation of Truncated Normal Variables," *Statistics and Computing*, 5, 121–125.

Tarantola, C. (2004), "MCMC Model Determination for Discrete Graphical Models," *Statistical Modeling*, 4, 39–61.

van Dyk, D. and Meng, X. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 1, 1–50.

Webb, E. and Forster, J. (2008), "Bayesian Model Determination for Multivariate Ordinal and Binary Data," *Computational Statistics and Data Analysis*, 52, 2632–2649.

Wong, F., Carter, C., and Kohn, R. (2003), "Efficient Estimation of Covariance Selection Models," *Biometrika*, 90, 809–830.

Zhang, X., Boscardin, W., and Belin, T. (2006), "Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables," *Journal of Computational and Graphical Statistics*, 15, 880–896.