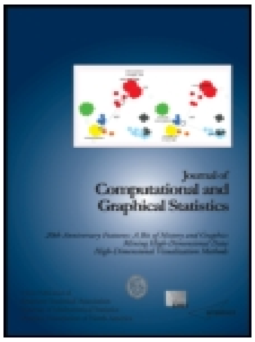

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables

Xiao Zhang, W. John Boscardin & Thomas R Belin

To cite this article: Xiao Zhang, W. John Boscardin & Thomas R Belin (2006) Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables, Journal of Computational and Graphical Statistics, 15:4, 880-896, DOI: 10.1198/106186006X160050

To link to this article: https://doi.org/10.1198/106186006X160050

Published online: 01 Jan 2012.

Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables

Xiao ZHANG, W. John BOSCARDIN, and Thomas R. BELIN

Hierarchical model specifications using latent variables are frequently used to reflect correlation structure in data. Motivated by the structure of a Bayesian multivariate probit model, we demonstrate a parameter-extended Metropolis-Hastings algorithm for sampling from the posterior distribution of a correlation matrix. Our sampling algorithms lead directly to two readily interpretable families of prior distributions for a correlation matrix. The methodology is illustrated through a simulation study and through an application with repeated binary outcomes on individuals from a study of a suicide prevention intervention.

Key Words: Metropolis-Hastings algorithm; Multivariate probit model.

1. INTRODUCTION

In longitudinal studies, repeated observations of a response variable and a set of covariates are made on individuals across occasions. Because repeated measurements are made on the same individual, the response variables are usually correlated within an individual. This correlation structure is clearly important for analyzing the data, but it can be difficult to model, especially for categorical responses. The generalized estimating equations (GEE) approach (Liang and Zeger 1986) has become a widely used strategy for analyzing correlated categorical outcomes, but this approach does not facilitate direct inference about the dependence.

More than a century ago, Pearson (1900) assumed that ordered categorical data were discretized versions of underlying approximately normally distributed variables and estimated the product moment correlation between these normal variables. This idea motivated wide use of probit regression models.

Xiao Zhang is Postdoctoral Fellow, Department of Biostatistics, University of Alabama at Birmingham School of Public Health, 1665 University Boulevard, Birmingham, AL 35294. W. John Boscardin is Associate Professor of Biostatistics, UCLA Schools of Medicine and Public Health, 51-254 Center for Health Sciences, Los Angeles, CA 90095-1772 (E-mail: *jbosco@ucla.edu*). Thomas R. Belin is Professor, Department of Biostatistics, UCLA School of Public Health, 51-254 Center for Health Sciences, Los Angeles, CA 90095-1772.

*Journal of Computational and Graphical Statistics, Volume 15, Number 4, Pages 880–896
DOI: 10.1198/106186006X160050*

Recent advances in Bayesian computational methods have extended the range of applications for these models. Albert and Chib (1993) developed a Markov chain Monte Carlo (MCMC) algorithm for univariate ordinal outcome data. Chib and Greenberg (1998) then extended this method to treat correlated multivariate outcomes. Briefly, they used a multivariate version of the probit model which assumes a latent continuous vector underlying each subject's ordinal data vector. The latent vector is assumed to follow a multivariate normal distribution with covariance matrix R. However, to avoid parameter identification problems, R must, in fact, be a *correlation* matrix. Sampling from the conditional distribution of a correlation matrix given other model parameters is challenging for two reasons. First, we need to put a reasonable prior distribution on the correlation matrix. Second, the conditional distribution is not generally available in closed form. Chib and Greenberg (1998) used a multivariate truncated normal distribution as the prior distribution for R and proposed a Metropolis-Hastings algorithm to sample R. Their approach has two potential disadvantages: (1) the proposal matrices at each step of the sampler are not necessarily correlation matrices, and (2) prior specification for the elements of the correlation matrix is not straightforward. The multivariate truncated normal prior distribution on R does not naturally incorporate useful prior information in the posterior analysis. Through reparameterization, Liu (2001) was able to avoid the difficulty of sampling a correlation matrix in the special case of a Jeffreys' prior on R. Nandram and Chen (1994) also avoided direct treatment of the correlation matrix using a joint reparameterization of R and the cutpoints for the ordinal data. Edwards and Allenby (2003) adopted a strategy of ignoring the parameter identification and then using the method of McCulloch and Rossi (1994) to deal with the identification restrictions. Barnard, McCulloch, and Meng (2000) modeled the correlation matrix of the random effects in a hierarchical linear model and used the Griddy Gibbs sampler (Ritter and Tanner 1992) to draw each of the components of a correlation matrix one at a time given the others assuming either a jointly uniform distribution or a marginally uniform prior distribution for the correlation matrix. Wong, Carter, and Kohn (2003) proposed a hierarchical prior on the partial correlation matrix and used a Metropolis-Hastings algorithm to sample each element. Liechty, Liechty, and Müller (2004) sampled the correlations using mixture priors to allow clustering into groups of positive correlations and negative correlations.

In this article, we develop families of flexible prior distributions for correlation matrices and a natural Metropolis-Hastings algorithm for sampling of the correlation matrix. We add extra variance parameters into the model by specifying a joint prior distribution on both the correlation matrix and a diagonal variance matrix, and then sample these two matrices together in a Metropolis-Hastings step. We demonstrate our methodology in the context of a multivariate probit model, but both the prior distributions and the algorithm can be applied in very general settings. The rest of the article proceeds as follows: Section 2 reviews details of the multivariate probit model and the Wishart distribution which are relevant to our sampling algorithms. Section 3 presents our parameter-extended Metropolis-Hastings (PX-MH) algorithms for sampling from the posterior distribution of the correlation matrix. Prior distributions for correlation matrices are discussed in Section 4. Section 5 uses a simulation study to illustrate the PX-MH algorithm for different prior distributions. Section 6 analyzes the correlation structure of binary data from a suicide prevention study (Rotheram-Borus et al. 1996). Some discussion and concluding remark are offered in Section 7.

2. MODELING FRAMEWORK AND STATISTICAL BACKGROUND

2.1 MULTIVARIATE PROBIT MODEL

We begin by reviewing the multivariate probit model as described by Chib and Greenberg (1998). Suppose we have N subjects measured at each of q occasions or measured on each of q attributes. Let $Y_1, \ldots, Y_N$ be multivariate outcome variables with $Y_i = (Y_{i1}, \ldots, Y_{iq})^T$ for $i = 1, \ldots, N$, and let $X_{ij} = (X_{ij1}, \ldots, X_{ijl})^T$ be an $l \times 1$ vector of observed covariates for each subject i and each measurement occasion $j = 1, \ldots, q$. It is straightforward to extend our methodology to the case where Y_i is comprised of ordinal data with more than two levels, but for simplicity, we restrict our attention in this article to the case where the components of Y_i are binary. We assume the following model structure. Each Y_{ij} is Bernoulli distributed with success probability π_{ij} assumed to follow a probit model, that is, $\pi_{ij} = \Phi(X_{ij}^T\beta)$, where $\Phi(.)$ is the cumulative standard normal distribution function and β is a $l \times 1$ vector of unknown regression parameters.

Let $\mathrm{X}_i = [X_{i1}, \ldots, X_{iq}]^T$ be the design matrix for the ith subject. We assume the latent vector $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$ follows a multivariate normal distribution with mean equal to $\mathrm{X}_i\beta$ and covariance matrix equal to R, which is assumed common across subjects. Then $P(Z_{ij} > 0) = \Phi(X_{ij}^T\beta/\mathrm{R}_{jj}^{1/2})$, where R_{jj} is the variance of Z_{ij}. From the latent variable perspective, the probit model assumes that $P(Y_{ij} = 1) = P(Z_{ij} > 0)$, which requires R_{jj} to be equal to 1 for $j = 1, \ldots, q$. Therefore, for the multivariate probit model to be identified, R, the covariance matrix of Z_{ij}, must in fact be a correlation matrix. The matrix R is sometimes called the *tetrachoric* or *polychoric correlation* of the Y_i (Drasgow 1986).

Suppose the joint prior distribution for β and R is $p(\beta, R) = p(\beta)p(R)$ and $p(\beta) = \mathrm{N}_l(\beta; b, C)$. The prior distribution for the correlation matrix R, $p(R)$, will be discussed later. Letting $Y = (Y_1, \ldots, Y_N)$ and $Z = (Z_1, \ldots, Z_N)$, we have

$$p(\beta, R, Z|Y) \quad \propto \quad p(\beta) \times p(R) \times \prod_{i=1}^{N}[I_i \times \phi(Z_i; \mathrm{X}_i\beta, R)],$$

where ϕ is the standard normal density function, and I_i indicates compatibility of the latent vector Z_i with the binary vector Y_i through the expression

$$\begin{aligned} I_i \quad = \quad & 1_{(Z_{i1}>0, Z_{i2}>0, \ldots, Z_{iq}>0)} 1_{(Y_{i1}=1, Y_{i2}=1, \ldots, Y_{iq}=1)} \\ & + 1_{(Z_{i1}\le 0, Z_{i2}>0, \ldots, Z_{iq}>0)} 1_{(Y_{i1}=0, Y_{i2}=1, \ldots, Y_{iq}=1)} \\ & + \cdots \\ & + 1_{(Z_{i1}\le 0, Z_{i2}\le 0, \ldots, Z_{iq}\le 0)} 1_{(Y_{i1}=0, Y_{i2}=0, \ldots, Y_{iq}=0)}, \end{aligned}$$

where $1_{(.)}$ is the indicator function. To implement the sampler, the full conditional distributions are as follows:

- Using standard Bayesian linear model results, $\beta | R, Z, Y$ has a multivariate normal distribution:

  $$\beta | R, Z, Y \sim \mathrm{N}_l(\hat{\beta}, V_\beta),$$

  where

  $$V_\beta = \left(\sum_{i=1}^{N} \mathrm{X}_i{}^T R^{-1} \mathrm{X}_i + C^{-1} \right)^{-1}$$

  and

  $$\hat{\beta} = V_\beta \left(\sum_{i=1}^{N} \mathrm{X}_i{}^T R^{-1} Z_i + C^{-1} b \right).$$

- $Z_{ij} | \beta, R, Z_{ik}, k \neq j, Y_{ij}$ have normal distributions truncated at the left or right by zero:

  $$\begin{aligned} p(Z_{ij} | \beta, R, Z_{ik}, k \neq j, Y_{ij}) & \\ & \propto I_{ij} \times p(Z_{ij} | \beta, R, Z_{ik}, k \neq j) \\ & \propto I_{ij} \times \phi(Z_{ij}; \widetilde{\mu_{ij}}, \widetilde{R_{ij}}), \end{aligned}$$

  where $I_{ij} = 1_{(Y_{ij}=1, Z_{ij}>0)} + 1_{(Y_{ij}=0, Z_{ij} \leq 0)}$ indicates compatibility of Y_{ij} and Z_{ij}, and $\widetilde{\mu_{ij}}$ and $\widetilde{R_{ij}}$ are the conditional mean and variance of Z_{ij} given $Z_{ik}, k \neq j$:

  $$\begin{aligned} \widetilde{\mu_{i,j}} &= X_{i,j}^T \beta + R_{j,-j} R_{-j,-j}^{-1} (Z_{i,-j} - X_{i,-j} \beta) \\ \widetilde{R_{i,j}} &= R_{j,j} - R_{j,-j} R_{-j,-j}^{-1} R_{-j,j}; \end{aligned}$$

  in these expressions, $R_{i,-i}$ refers to the ith row of R without its ith column element, $R_{-i,-i}$ is the R matrix without its ith row and ith column, $X_{i,-j}$ is the matrix X_i without jth row, $Z_{i,-j}$ is the vector Z_i without its jth element, and $R_{-i,i}$ is the transpose of $R_{i,-i}$.

  In the event that Y_{ij} is missing, then Z_{ij} follows a univariate normal distribution without truncation:

  $$\begin{aligned} & p(Z_{ij} | \beta, R, Z_{ik}, k \neq j) \\ & \propto \quad \phi(Z_{ij}; \widetilde{\mu_{ij}}, \widetilde{R_{ij}}) \end{aligned}$$

- $p(R | \beta, Z, Y)$ is proportional to $p(R) \times \prod_{i=1}^{N} \phi(Z_i; \mathrm{X}_i \beta, R)$. It is not easy to directly draw simulations from this distribution, which does not have the form of any standard density function.

The remainder of this article is focused on developing direct and general methods for sampling R and specifying prior distributions for R.

2.2 Marginal Distributions for Correlation Matrices

There are not many well-known families of density functions for correlation matrices. Box and Tiao (1973) noted that the Jeffreys' prior for R is $p(R) \propto |R|^{-(q+1)/2}$. Chib and Greenberg (1998) put a multivariate normal prior on elements of R, truncated to the space of proper correlation matrices. It is somewhat difficult to place informative priors on R in this framework. Barnard et al. (2000) considered used two alternative prior specifications for R: either a jointly uniform prior, $p(R) \propto 1$, or a marginally uniform prior, $p(r_{ij}) \propto 1$.

In line with the derivations of Barnard et al. (2000), whose models will emerge as special cases of our parameter-extended methodology, we now motivate more flexible solutions for placing a prior density on the correlation matrix by decomposition of the covariance matrix. If W is a positive definite covariance matrix, then we can write $W = D^{\frac{1}{2}} R D^{\frac{1}{2}}$ where R is the corresponding correlation matrix (i.e., $r_{ij} = w_{ij}/\sqrt{w_{ii} w_{jj}}$), and $D = \mathrm{diag}(d_1, \ldots, d_q)$ is a diagonal matrix of the variances with $d_i = w_{ii}$, implying that $D^{\frac{1}{2}}$ is a diagonal matrix of standard deviations. The Jacobian of the transformation from W to (R, D) is

$$J_{W \to R,D} = \prod_{i=1}^{q} d_i^{\frac{q-1}{2}}. \tag{2.1}$$

Thus, we can calculate the joint density function of (R, D) if we know the density function of W.

We now investigate marginal distributions for the correlation matrix R. Suppose $W_{q \times q}$ follows a Wishart distribution with degrees of freedom m and scale matrix Σ, that is, $W \sim \mathrm{Wishart}_q(m, \Sigma)$. This implies

$$p(W) = c^{-1} |W|^{\frac{m-q-1}{2}} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1} W\right),$$

where $\mathrm{etr}(\cdot)$ represents the operator $\exp(\mathrm{tr}(\cdot))$ and c is the normalizing constant. The joint distribution of R and D can be characterized in terms of the distribution of W and the Jacobian of the transformation $(W \to R, D)$ as follows:

$$\begin{aligned} p(R, D) &= c^{-1} \left(\prod_{i=1}^{q} w_{ii}\right)^{\frac{q-1}{2} + \frac{m-q-1}{2}} |R|^{\frac{m-q-1}{2}} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1} D^{\frac{1}{2}} R D^{\frac{1}{2}}\right) \\ &= c^{-1} |R|^{\frac{m-q-1}{2}} |D|^{\frac{m}{2}-1} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1} D^{\frac{1}{2}} R D^{\frac{1}{2}}\right). \end{aligned}$$

By integrating out D, the marginal density of R is

$$p(R) = \int c^{-1} |R|^{\frac{m-q-1}{2}} |D|^{\frac{m}{2}-1} \mathrm{etr}\left(-\frac{1}{2}\Sigma^{-1} D^{\frac{1}{2}} R D^{\frac{1}{2}}\right) dd_1 \ldots dd_q. \tag{2.2}$$

In general, there is no closed form for the above integral, and thus we cannot obtain an explicit density function for the correlation matrix R. However, in the special case when Σ is equal to any nonsingular diagonal matrix, we have that D and R are independent with the elements of D having scaled chi-squared distributions and $p(R) \propto |R|^{\frac{1}{2}(m-q-1)}$, where the normalizing constant is analytically tractable (Gupta and Nagar 2000, theorem 3.3.24). Similarly, we can obtain parallel results for the case where $W_{q\times q}$ follows an inverse-Wishart distribution with degrees of freedom m and scale matrix Σ, that is, $\text{Wishart}_q{}^{-1}(m, \Sigma)$. In this case, for Σ equal to any nonsingular diagonal matrix, we can get the following closed form for the marginal density of R:

$$p(R) \propto |R|^{\frac{1}{2}(m-1)(q-1)-1}\left(\prod_i |R_{ii}|\right)^{-\frac{m}{2}},$$

where R_{ii} is the ith principle submatrix of R and the normalizing constant can be explicitly calculated.

3. PARAMETER-EXTENDED METROPOLIS-HASTINGS (PX-MH) ALGORITHM FOR SAMPLING THE CORRELATION MATRIX

Although there is no closed form for the marginal density of R in general, a key insight from Section 2.2 is that we can obtain an explicit joint density for R and D based on the density of $W = D^{\frac{1}{2}}RD^{\frac{1}{2}}$. Therefore, by including an extra variance parameter D into our model and specifying a joint prior distribution on R and D, we can use a Metropolis-Hastings algorithm to sample R and D together.

We pursue this strategy for the multivariate probit model of Section 2.1. We specify the prior for the parameters as $p(\beta, R, D) = p(\beta)p(R, D)$, with interest now focused on the joint posterior distribution $p(\beta, R, D, Z|Y)$. The full conditional distributions for β and Z are the same as in Section 2.1. The full conditional distribution for R is replaced by jointly sampling (R, D) from

$$p(R, D|\beta, Z, Y) \propto p(R, D)p(Z|\beta, R) = p(R, D)\prod_{i=1}^{N}\phi(Z_i; \mathrm{X}_i\beta, R). \tag{3.1}$$

The two difficulties of putting a prior on R and sampling R from its posterior distribution are now easily solved. Prior distributions for R are discussed in Section 4. Briefly, we can use a Wishart prior distribution on W and then calculate the induced prior on (R, D) using the Jacobian of the transformation in Equation (2.1). Sampling (R, D) is accomplished through a Metropolis-Hastings algorithm which proposes a random covariance matrix centered around the current value of $W = D^{\frac{1}{2}}RD^{\frac{1}{2}}$. We call this the parameter-extended Metropolis-Hastings algorithm (PX-MH) by analogy with earlier application of the idea of parameter expansion (Liu, Rubin, and Wu 1998), where inference is facilitated by a statistical computing strategy that incorporates unidentified parameters (D, in the present case) instead of using a standard identifiability constraint. Our setting is somewhat different in that the model is specified through an informative joint prior distribution on (R, D). However, as in parameter expansion, the data are not informative about D. Both algorithms differ from the data augmentation paradigm where the data are informative about the extra parameters, although the amount of information does not increase with sample size (Gelman 2004).

PX-MH ALGORITHM

Set initial value of $(R^{(0)}, D^{(0)})$ through setting $W^{(0)} = D^{(0)\frac{1}{2}} R^{(0)} D^{(0)\frac{1}{2}}$ to an initial covariance matrix.

Then, at iteration t

1. Generate (R^*, D^*) by generating $W^* = D^{*\frac{1}{2}} R^* D^{*\frac{1}{2}}$ from $\text{Wishart}_q(m, W^{(t)})$.
2. Take

$$(R^{(t+1)}, D^{(t+1)}) = \begin{cases} (R^*, D^*) & \text{with probability } \alpha \\ (R^{(t)}, D^{(t)}) & \text{otherwise.} \end{cases}$$

where $\alpha = \min\left\{ \frac{p(R^*, D^* | \beta, Z, Y)}{p(R^{(t)}, D^{(t)} | \beta, Z, Y)} \frac{f(W^{(t)} | W^*)}{f(W^* | W^{(t)})}, 1 \right\}$. Here, $p(R, D|\beta, Z, Y)$ is the joint posterior density of (R, D) given in Equation (3.1) and $f(.|W^{(t)})$, the proposal density, is equal to the product of the Jacobian term given in Equation (2.1) and the Wishart density with m degrees of freedom and scale matrix equal to $W^{(t)}$. The process is then iterated with inference based on the now familiar Markov chain Monte Carlo framework (Robert and Casella 2000). The degrees of freedom m can be set to adjust the acceptance rate. Smaller values of m correspond to larger distances between the proposed value and the current value and thus lower acceptance rates.

4. PRIOR DISTRIBUTIONS FOR CORRELATION MATRICES

In this section, we propose several prior distribution families for correlation matrices, which can allow useful prior information and facilitate the Metropolis-Hastings step for sampling correlation matrices. The jointly uniform and marginally uniform prior distributions used by Barnard et al. (2000) and the Jeffreys' prior distribution are special cases.

Based on Section 2.2, we know the joint density of R and D is given by the product of the Jacobian of the transformation $(W \rightarrow R, D)$ and the density of a $\text{Wishart}_q(W; m, \Sigma)$ distribution.

We characterize this joint distribution as a parameter-extended Wishart, or PXW distribution, which we denote as $(R, D) \sim \text{PXW}(m, \Sigma)$. The PXW distribution can serve as a prior distribution for R. We can choose the structure of Σ to reflect our prior belief about R, with higher values of m representing stronger prior belief.

Daniels and Kass (1999, 2001) explored hierarchical priors on the covariance as well as hierarchical priors on the correlations which shrink the correlations toward 0. For the PXW prior, if Σ is equal to any nonsingular diagonal matrix, it will cause the off-diagonal elements of the correlation matrix to shrink towards 0, with high values of m leading to more shrinkage. This is appropriate if a priori, we believe R is near the identity matrix. Two special cases of note: the joint uniform prior of Barnard et al. (2000) corresponds to $\text{PXW}(q + 1, I)$, and the Jeffreys' prior $p(R) \propto |R|^{-(q+1)/2}$ corresponds to $\text{PXW}(0, I)$.

As also noted in Section 2.2, the same ideas can be extended to the inverse-Wishart family. If W follows a $\text{Wishart}_q^{-1}(W; m, \Sigma)$ distribution, then the joint density of R and D can be written as the product of the Jacobian of the transformation ($W \rightarrow R, D$) and the $\text{Wishart}_q^{-1}(W; m, \Sigma)$ density. We characterize this joint distribution as a parameter-extended inverse-Wishart, or PXIW distribution, which we denote as $(R, D) \sim \text{PXIW}(m, \Sigma)$. The marginal uniform distribution of (Barnard et al. 2000) corresponds to $\text{PXIW}(q + 1, I)$.

Both the PXW and PXIW families can be extended hierarchically (Boscardin and Weiss 2001). Briefly, the idea is to first choose a reasonable parametric family $\Sigma(\theta)$ that might capture the important features of R. For example, $\Sigma(\cdot)$ might be an autoregressive matrix of order one, in which case θ is a two-dimensional vector of the lag-one correlation ρ and the variance parameter σ^2. We can then specify the joint prior distribution of R and D using $\text{PXW}(m, \Sigma(\theta))$ or $\text{PXIW}(m, \Sigma(\theta))$. The model is completed with prior distributions on m and θ. In most settings, m will be a degrees-of-freedom parameter that can either be fixed, which might be viewed as a tuning parameter for the amount of smoothing performed, or allowed to vary. Similar models have appeared in the pre-MCMC literature (Chen 1979; Dickey, Lindley, and Press 1985), but computational limitations only allowed for very specific cases to be considered. In the next section, we use this hierarchical prior centered around first-order autoregressive (AR(1)) and compound symmetric (CS) parametric families.

In the following sections, we use the PXW prior for correlation matrices instead of the PXIW prior. Because conjugacy is not obtained in either case, the choice hinges on intuitive notions about the scale on which prior assumptions are most meaningful.

5. A SIMULATION STUDY

To illustrate the performance of the PX-MH algorithm, we carried out a simulation study for longitudinal binary datasets with five time points. The sample size for each simulated dataset was equal to 50. We set the covariate vector $X_{ij} = (1, j - 3)^T$ for each subject i and time point j with regression parameters $\beta = (\beta_0, \beta_1) = (0.5, 1.0)$. For the correlation matrix R of the latent variables Z_i, we used an ante-dependence structure with four correlation parameters $\rho = (\rho_1, \rho_2, \rho_3, \rho_4) = (0.7, 0.9, 0.5, 0.3)$ and let $r_{ij} = \prod_{k=i}^{j} \rho_k$ for $i < j$. Following a multivariate probit model, we generated a multivariate normal variable $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$ with mean vector equal to $X_{ij}^T \beta$ and covariance matrix equal to R. We then set $Y_{ij} = 1$ if $Z_{ij} > 0$, otherwise $Y_{ij} = 0$ for $i = 1, \ldots, 50$ and $j = 1, 2, 3, 4, 5$.

We used the PX-MH algorithm to obtain posterior inference for the regression parameters and the correlation matrix based on six alternative prior distributions for the correlation matrix R. We used PXW prior distributions with the following prior guesses for R, none of which is correct: Identity, AR(1)(0.3), AR(1)(0.5), AR(1)(θ), CS(0.5), and CS(θ), where AR(1)(θ) and CS(θ) refer to the hierarchical versions of these parametric families where θ is allowed to vary. We used a Beta(3,1) prior distribution shifted and scaled to the interval $(-1, 1)$ for θ, implying a weakly informative distribution with prior mean equal to 0.5. We also explored analogous PXIW prior distribution, but since the PX-MH algorithm performed similarly with PXW and PXIW priors, we only show the Wishart-based results in the following simulation study.

To check the convergence of the MCMC algorithm, for one simulated dataset, we calculated Gelman and Rubin's potential scale reduction factor, $\sqrt{\hat{R}}$ for five dispersed chains with the first 1,000 iterations discarded as burn-in (Gelman and Rubin 1992). The degrees of freedom for the Wishart proposal distribution was set to 300, and this resulted in an acceptance rate of 15%. Although this is somewhat below the target rate of 23% given by Gelman, Roberts, and Gilks (1996), we observed that convergence did not appear to be quite as rapid when the degrees of freedom for the proposal distribution was increased to achieve a higher acceptance rate. For all the regression parameters $\beta = (\beta_0, \beta_1)$ and correlation parameters r_{ij}, the values of $\sqrt{\hat{R}}$ were all below 1.1 after 20,000 iterations and declined consistently through a further 80,000 iterations. The multivariate potential scale reduction factor (Brooks and Gelman 1998) was 1.13 after 20,000 iterations, improving to 1.06 at 100,000 iterations. Altering the proposal degrees of freedom did not result in improved mixing. For the simulation study, we thus used a total of 21,000 iterations with the first 1000 discarded for a burn-in period. The simulations were saved every 10th iteration to produce manageable sample sizes for analysis.

For each of 50 simulated datasets for each simulation condition, we used the PX-MH algorithm to sample the correlation matrix and calculate the posterior mean and the standard deviation of model parameters. Knowing the underlying true values enabled us to calculate bias and mean-squared error for individual model parameters.

To make a matrix-wise comparison of the estimated correlation matrix under different prior assumptions, we computed the Stein risk function, namely $L(\hat{R}, R) = \mathrm{tr}(\hat{R}R^{-1}) - \log|\hat{R}R^{-1}| - q$ where $\hat{R}$ is the estimated posterior correlation matrix and R represents the true correlation matrix (Yang and Berger 1994).

Table 1 presents results across the various prior distributions for regression parameters and the Stein risk function, averaged over 50 simulation replicates for each prior. From the table, we see that the regression parameter estimates succeed in reproducing the true underlying values. We review the values for the Stein risk function after considering results for individual parameter values in the next paragraph.

Table 2 presents detailed results for correlation-matrix parameters. The portion of the table above the diagonal shows the posterior mean and standard deviation, and the lower triangular portion of the table shows the root mean square error for each of the elements of the correlation matrix. We can see that the model with the identity prior always gives the smallest posterior mean correlations, consistent with shrinkage towards the identity, as well as the largest posterior standard deviations for each of the elements of the correlation matrix. The two hierarchical models, AR(1)(θ) and CS(θ), give higher means and smaller standard deviations than the other models. However, they tend to give larger values for the root mean square errors than the models with nonhierarchical priors. We can see that for some correlation elements, such as r_{15} r_{35}, and r_{45}, the model with AR(1)(0.3) prior gives the best estimation. However, for correlation elements, such as r_{14}, r_{24}, and r_{34}, the model with AR(1)(0.5) prior gives much better estimation than the other models. Because it is difficult in an element-by-element analysis to say which model gives better estimated correlation matrix than the other models, we carry out a matrix-wise comparison using the Stein loss function, which is shown in the fourth column of Table 1. We can see that for AR(1) structure,

Table 1. Posterior Means and Standard Deviations for the Regression Parameters (β_0 and β_1), the Stein Loss Function($L(R,\hat{R})$), and hyper-parameter θ under various priors in the simulation study.

Quantity	β_0	β_1	$L(R,\hat{R})$	θ
True	0.50	1.00	.	.
Identity	0.50	1.04	2.37	.
(std)	(0.14)	(0.11)	(0.61)	.
AR(1)(0.3)	0.50	1.04	1.45	.
(std)	(0.14)	(0.11)	(0.43)	.
AR(1)(0.5)	0.50	1.04	1.02	.
(std)	(0.14)	(0.11)	(0.34)	.
AR(1)(θ)	0.49	1.02	1.20	0.69
(std)	(0.15)	(0.11)	(0.66)	(0.15)
CS(0.5)	0.49	1.01	1.55	.
(std)	(0.15)	(0.11)	(0.36)	.
CS(θ)	0.49	1.00	1.80	0.63
(std)	(0.15)	(0.11)	(0.58)	(0.16)

the model with AR(1)(0.5) prior gives smallest value of the loss function and the model with the identity prior gives the largest value of the loss function. The loss function value for the hierarchical prior AR(θ) is better than those with identity and AR(1)(0.3) priors, but not as good as the model with AR(1)(0.5) prior. This result suggests that we could use a hierarchical prior distribution for the correlation matrix if we do not have any specific prior knowledge about it, but that a good fixed choice for the prior scale can outperform the hierarchical model. We also explore the compound symmetry (CS) structure for the prior of the correlation matrix. From Table 1, we can see that the values of the loss function are larger for the models with CS structure than those with AR(1) structure. Because AR(1) structure is closer to the true correlation structure (which is ante-dependence) than the compound symmetry structure, we see that better specification of the prior distribution gives better posterior inference. Through Table 1 and Table 2, we can see that the estimated parameters of the mean structure, β, are robust to the choice of prior distribution, while the correlation parameters are more sensitive to the prior specification.

In the last column of Table 1, we also show posterior means and the standard deviations for the hyper-parameter θ under the AR(1)(θ) and CS(θ) models. There is no underlying "true value" for θ, since the true correlation matrix does not belong to either of these parametric families. Thus, $\tilde{\theta} = E[\theta|Y]$ can be thought of as the averaged value in the assumed parametric family. Here, we find that $\tilde{\theta} = 0.69$ for the AR(1)(θ) model and $\tilde{\theta} = 0.63$ for the CS(θ) model. This suggests that slightly higher fixed choices for θ may have been more appropriate than 0.5.

Table 2. Detailed Simulation-Study Findings for the Correlation Matrix Entries. Above diagonal: posterior means and standard deviations. Below diagonal: posterior root mean square errors.

True	0.70	0.63	0.32	0.09
Identity	0.39 (0.22)	0.28 (0.26)	0.15 (0.3)	0.03 (0.33)
AR(1)(0.3)	0.51 (0.2)	0.36 (0.25)	0.20 (0.29)	0.07 (0.33)
AR(1)(0.5)	0.61 (0.18)	0.45 (0.23)	0.28 (0.28)	0.14 (0.32)
AR(1)(θ)	0.70 (0.17)	0.58 (0.21)	0.41 (0.26)	0.32 (0.31)
CS(0.5)	0.62 (0.17)	0.58 (0.20)	0.54 (0.22)	0.53 (0.25)
CS(θ)	0.68 (0.17)	0.65 (0.19)	0.61 (0.21)	0.61 (0.23)
	True	0.90	0.45	0.14
0.389	Identity	0.59 (0.17)	0.25 (0.25)	0.06 (0.33)
0.283	AR(1)(0.3)	0.67 (0.15)	0.33 (0.24)	0.13 (0.32)
0.213	AR(1)(0.5)	0.72 (0.13)	0.41 (0.22)	0.21 (0.31)
0.214	AR(1)(θ)	0.79 (0.12)	0.52 (0.21)	0.39 (0.29)
0.205	CS(0.5)	0.72 (0.13)	0.55 (0.19)	0.55 (0.22)
0.212	CS(θ)	0.76 (0.13)	0.61 (0.19)	0.63 (0.22)
		True	0.50	0.15
0.439	0.363	Identity	0.30 (0.22)	0.06 (0.32)
0.370	0.331	AR(1)(0.3)	0.42 (0.20)	0.15 (0.32)
0.296	0.302	AR(1)(0.5)	0.52 (0.19)	0.26 (0.30)
0.261	0.336	AR(1)(θ)	0.63 (0.17)	0.45 (0.27)
0.219	0.335	CS(0.5)	0.54 (0.18)	0.53 (0.23)
0.238	0.405	CS(θ)	0.61 (0.17)	0.62 (0.22)
			True	0.30
0.346	0.351	0.335	Identity	0.05 (0.33)
0.342	0.275	0.291	AR(1)(0.3)	0.30 (0.31)
0.337	0.220	0.261	AR(1)(0.5)	0.47 (0.27)
0.417	0.170	0.290	AR(1)(θ)	0.63 (0.24)
0.509	0.225	0.257	CS(0.5)	0.49 (0.24)
0.585	0.194	0.319	CS(θ)	0.58 (0.23)
				True
0.358	0.317	0.360	0.434	Identity
0.342	0.247	0.345	0.327	AR(1)(0.3)
0.335	0.225	0.340	0.330	AR(1)(0.5)
0.420	0.275	0.440	0.436	AR(1)(θ)
0.478	0.221	0.455	0.339	CS(0.5)
0.555	0.269	0.531	0.409	CS(θ)

6. EXAMPLE

We illustrate our PX-MH algorithm on data from a suicide prevention study (Rotheram-Borus et al. 1996). The suicide prevention dataset consists of 140 female adolescents presenting to the emergency room at a New York City medical center after a suicide attempt between March 1991 and February 1994. The mean age of these participants was 15 years with an age range of 12 to 18 years. Participants were offered one of two treatment protocols: either the standard emergency room procedures as of the start of the study, or a specialized program designed by the investigators that aimed through staff training and a video presentation to improve patient understanding of the eventual course of treatment. In both cases, subjects were offered the opportunity to attend frequent counseling sessions. Participants were followed after 3 months, 6 months, 12 months, and 18 months. We examine the polychoric correlation of longitudinal measures of suicidal ideation, a discrete variable representing whether the patient's answers to a standardized measurement procedure indicated strong thoughts of suicide.

We modeled $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})$, the longitudinal vector of ideation indicators at 0, 3, 6, 12, and 18 months, where the 0-month measure was obtained just after initiation of treatment. We allowed different means at each of the five time points across four different groups of subjects defined by crossing treatment status (standard protocol vs. specialized intervention) and frequency of attendance at counseling sessions (frequent vs. infrequent). Thus, the vector β is of length 20, and the 20-by-5 design matrix X_i reflects the five subgroup-specific means for the ith subject. The model is described as follows:

1. $p(Y_{ij} = 1|Z_{ij}) = 1_{(Z_{ij}>0)}$
2. $Z_i|\beta, R \sim \mathrm{N}_5(X_i\beta, R)$
3. $p(\beta, R) = p(\beta)p(R)$.

We use the PX-MH algorithm for this model with two alternative prior assumptions: the PXW$(m_0 = 8, I)$ prior (labeled PXW_I) and PXW$(m_0 = 8, AR(1)(\rho = 0.5))$ prior (labeled PXW_AR1). The PXW_I prior distribution reflects prior belief in no correlation; this is not what we expect to see, but this choice of prior distribution is still useful for comparison. The PXW_AR1 prior distribution reflects a belief in moderate correlations that attenuate with the gap between measurements. This is a more realistic representation of our prior guess for the correlation structure. As in the example of the previous section we ran 21,000 iterations with the first 1,000 treated as burn-in iterations. We checked the convergence behavior by generating five dispersed starting values and calculating Gelman-Rubin potential scale reduction factors, $\sqrt{\hat{R}}$, which were all less than 1.10. Using $m = 300$ as the degrees of freedom for the proposal Wishart distribution, the Metropolis-Hastings acceptance rate was close to 18%.

Table 3 contains the posterior means and 90% credible intervals, using PXW_AR1, for the probability of suicidal ideation at each of the five time points and for every combination of intervention and frequency of attendance. We also list the empirical proportions of ideation in each cell of the table. For frequent attenders, the probability of suicidal ideation

Table 3. Posterior Estimated Probabilities and Empirical Proportions of Suicidal Ideation Under Standard and Specialized Programs for all Subjects With 90% Credible Intervals

Treatment	*Standard*				*Specialized*			
Attendance	*Nonfrequent*		*Frequent*		*Nonfrequent*		*Frequent*	
time point	*Posterior*	*Emp*	*Posterior*	*Emp*	*Posterior*	*Emp*	*Posterior*	*Emp*
0 Months	0.60	0.60	0.83	0.82	0.42	0.42	0.44	0.44
	(0.49,0.71)	(32/53)	(0.67,0.93)	(18/22)	(0.30,0.56)	(16/38)	(0.30,0.60)	(12/27)
3 Months	0.37	0.20	0.56	0.38	0.34	0.26	0.22	0.19
	(0.27,0.49)	(8/41)	(0.38,0.72)	(6/16)	(0.22,0.47)	(9/34)	(0.12,0.37)	(5/26)
6 Months	0.32	0.14	0.27	0.20	0.29	0.21	0.19	0.15
	(0.22,0.43)	(6/42)	(0.13,0.44)	(4/20)	(0.18,0.42)	(7/34)	(0.09,0.33)	(4/26)
12 Months	0.19	0.11	0.26	0.24	0.18	0.11	0.12	0.11
	(0.11,0.29)	(5/47)	(0.13,0.43)	(5/21)	(0.10,0.30)	(4/35)	(0.04,0.25)	(3/27)
18 Months	0.18	0.09	0.22	0.23	0.29	0.23	0.12	0.08
	(0.11,0.28)	(4/47)	(0.10,0.39)	(5/22)	(0.18,0.42)	(8/35)	(0.04,0.24)	(2/26)

under the specialized program was lower than that under the standard program, and sharply lower at the early time points, suggesting the effectiveness of the specialized intervention. Regarding frequency of attendance at follow-up counseling sessions, suicidal ideation rates were slightly but not convincingly lower for frequent versus nonfrequent attenders in the specialized intervention group, and rates for frequent attenders were higher than for non-frequent attenders with the standard emergency room protocol.

The patterns in the empirical proportions are fairly well tracked by the estimated probabilities in the specialized-treatment groups and the frequent attenders in the standard-treatment group; however, departures between observed and predicted rates are seen among infrequent attenders in the standard-treatment group.

Table 4 shows point estimates and standard errors for the correlation parameters under PXW_I and PXW_AR1 priors. We see the estimates for the correlation parameters appear to depend somewhat on the specification of prior distributions. The strongest correlation is r_{34}, suggesting that those with higher suicidal ideation at the 6-month time point tended to have higher levels at the 12-month time point. Other correlations at later time points were also substantial.

The posterior density plots and prior density plots for the correlation matrix with PXW_I and PXW_AR1 priors are drawn in Figure 1. The data contain at least some information about all of the correlation parameters, and quite a bit of information about several of them

Table 4. Posterior Means and Standard Deviations for the Elements of R in the Suicidal Ideation Data Under the Two Prior Specifications PXW_I and PXW_AR1

Quantity	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}
PXW_I Mean	0.24	0.24	0.00	0.43	0.49
(SD)	(0.12)	(0.14)	(0.13)	(0.13)	(0.11)
PXW_AR1 Mean	0.31	0.27	0.09	0.44	0.57
(SD)	(0.12)	(0.13)	(0.12)	(0.13)	(0.10)
Quantity	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
PXW_I Mean	0.54	0.53	0.70	0.34	0.44
(SD)	(0.11)	(0.10)	(0.09)	(0.11)	(0.12)
PXW_AR1 Mean	0.57	0.58	0.74	0.41	0.52
(SD)	(0.10)	(0.10)	(0.08)	(0.11)	(0.11)

(notably r_{25} and r_{34}). Inference appears to be fairly robust to the choice between these two priors.

7. DISCUSSION

In this article, we introduced the PX-MH algorithm for sampling a correlation matrix in an MCMC setting and presented general families of prior distributions for a correlation matrix. We used a simulation study and an application to repeated binary measures from a study of suicide prevention to illustrate properties of our methods.

The PX-MH algorithm has two distinct advantages for sampling a correlation matrix: (1) it directly proposes a correlation matrix in each sampling step; and (2) it allows flexible prior distributions on correlation matrices, which can incorporate informative prior information through the specification of the scale matrix and the value of the degrees-of-freedom parameter. Both of the families of distributions we introduced, the PXW prior distributions and the PXIW prior distributions, offer these advantages, providing flexibility to choose either a Wishart or an inverse-Wishart family. The PX-MH algorithm also allows hierarchical priors for correlation matrices.

Given that the examples we considered involved five-dimensional correlation matrices, a question arises regarding the performance of the procedure in higher-dimensional settings. We explored the PX-MH algorithm on simulated data with a 20-by-20 correlation matrix and we found that acceptable convergence behavior necessitated at least 100,000 iterations. A larger value of m, the degrees of freedom for the proposal Wishart density, is also needed to provide an appropriate acceptance rate. For example, with a 20-by-20 correlation matrix, we needed m to be at least 2,000 to obtain a roughly 10% acceptance rate. Meanwhile, higher values of m mean smaller moves from the current value. Thus, increasing m is associated with higher autocorrelation, thereby requiring more iterations

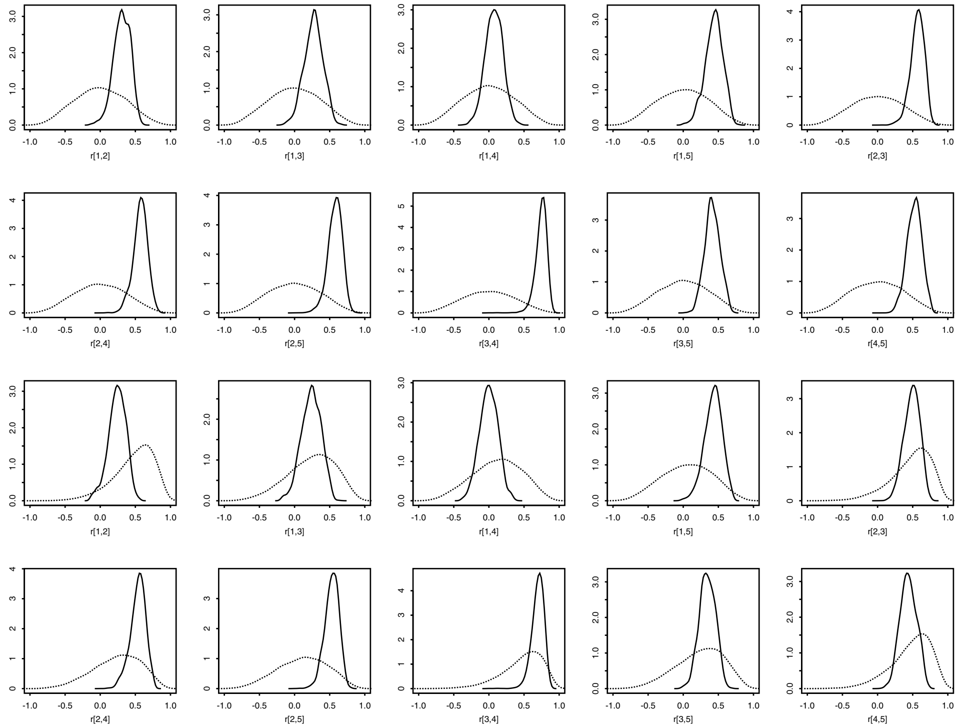

Figure 1. The first and second rows: the density plots for each correlation parameter with prior PXW_I centered around the identity matrix. The last two rows: the density plots for each correlation parameter with prior PXW_AR1 centered around the AR(1)(0.5) matrix. Horizonal axes run from -1 to 1 with tick marks every 0.5. The solid lines are the posterior density plots and the dashed line are the prior density plots. The density plots show that there is considerably more information about some correlation parameters (e.g., r_{25} and r_{34}), and the inference is fairly robust to the choice between these two priors.

to obtain convincing convergence. The increase in the correlation dimension also increases the dimension of the latent variable Z. Since we sample each component of Z conditional on all the others, higher dimensional Z will also slow convergence on account of higher autocorrelation. Despite these concerns, an advantage of the PX-MH algorithm is that we can also improve convergence through the strength of prior specification. Optimizing the tradeoffs involved in convergence of the algorithm is an open research area.

The PX-MH algorithm can be applied to different settings, such as multivariate probit models for ordinal data along the lines of the setting considered by Nandram and Chen (1994) and multinomial probit models analyzed by McCulloch, Polson, and Rossi (2000). Our work in these areas is summarized in separate reports.

ACKNOWLEDGMENTS

This work was supported by NIH grants MH60213 (all), NS30308 and AI28697 (WJB), MH58107 and MH68639 (TB).

[Received August 2005. Revised May 2006.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

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

Boscardin, W. J., and Weiss, R. E. (2001), "Models for the Covariance Matrix of Multivariate Longitudinal and Repeated Measures Data," in *Proceedings of the American Statistical Association*, Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

Brooks, S. P., and Gelman, A. (1998), "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7, 434–455.

Chen, C.-F. (1979), "Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis," *Journal of the Royal Statistical Society*, Series B, 41, 235–248.

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

Daniels, M. J., and Kass, R. E. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices and its Use in Hierarchical Models," *Journal of the American Statistical Association*, 94, 1254–1263.

——— (2001), "Shrinkage Estimators for Covariance Matrices," *Biometrics*, 57, 1173–1184.

Dickey, J. M., Lindley, D. V., and Press, S. J. (1985), "Bayesian Estimation of the Dispersion Matrix of a Multivariate Normal Distribution," *Communications in Statistics*, 14, 1019–1034.

Drasgow, F. (1986), "Polychoric and Polyserial Correlations," in *Encyclopedia of Statistical Sciences*, New York: Wiley.

Edwards, Y. D., and Allenby, Greg M. (2003), "Multivariate Analysis of Multiple Response Data," *Journal of Marketing Research*, August 2003, 321–334.

Gelman, A. (2004), "Parameterization and Bayesian Modeling," *Journal of the American Statistical Association*, 99, 537–545.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis Jumping Rules," in *Bayesian Statistics* (vol. 5), eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, Cambridge, MA: Oxford University Press, pp. 599–608.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511.

Gupta, A. K., and Nagar, D. K. (2000), *Matrix Variate Distributions*, Boca Raton: Chapman & Hall/CRC.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004), "Bayesian Correlation Estimation," *Biometrika*, 91, 1–14.

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

Liu, C., Rubin, D. B., and Wu, Y. (1998), "Parameter Expansion to Accelerate EM: The PX–EM Algorithm," *Biometrika*, 85, 755–770.

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

McCulloch, R., and Rossi, P. E. (1994), "An Exact Likelihod Analysis Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240.

Nandram, B., and Chen, M.-H. (1994), "Accelerating Gibbs Sampler Convergence in the Generalized Linear Models via a Reparameterization," *Journal of Statistical Computation and Simulation*, 81, 27–40.

Pearson, K. (1900), "Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable," *Philosophical Transactions of the Royal Society of London*, Series A, 195, 1–47.

Ritter, C., and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs

Sampler," *Journal of the American Statistical Association*, 87, 861–868.

Robert, C. P., and Casella, G. (2000), *Monte Carlo Statistical Methods*, New York: Springer.

Rotheram-Borus, M. J., Piacentini, J., Van Rossem, R., Graae, F., Cantwell, C., Castro-Blanco, D., Miller, S., and Feldman, J. (1996), "Enhancing Treatment Adherence with a Specilized Emergency Room Program for Adolescent Suicide Attempters," *Journal of the American Academy of Child and Adolescent Psychiatry*, 35, 654–663.

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

Yang, R., and Berger, J. O. (1994), "Estimation of a Covariance Matrix Using the Reference Prior," *The Annals of Statistics*, 22, 1195–1211.