Handbook of Statistics, Vol. 25
ISSN: 0169-7161

DOI 10.1016/S0169-7161(05)25029-0

Modeling and Analysis for Categorical Response Data

Siddhartha Chib

1. Introduction

In this chapter we discuss how Bayesian methods are used to model and analyze categorical response data. As in other areas of statistics, the growth of Bayesian ideas in the categorical data setting has been rapid, aided by developments in *Markov chain Monte Carlo* (MCMC) methods (Gelfand and Smith, 1990; Smith and Roberts, 1993; Tanner and Wong, 1987; Tierney, 1994; Chib and Greenberg, 1995), and by the framework for fitting and criticizing categorical data models developed in Albert and Chib (1993). In this largely self-contained chapter we summarize the various possibilities for the Bayesian modeling of categorical response data and provide the associated inferential techniques for conducting the prior–posterior analyses.

The rest of the chapter is organized as follows. We begin by including a brief overview of MCMC methods. We then discuss how the output of the MCMC simulations can be used to calculate the marginal likelihood of the model for the purpose of comparing models via Bayes factors. After these preliminaries, the chapter turns to the analysis of categorical response models in the cross-section setting, followed by extensions to multivariate and longitudinal responses.

1.1. Elements of Markov chain Monte Carlo

Suppose that $\boldsymbol{\psi} \in \Re^d$ denotes a vector of random variables of interest (typically consisting of a set of parameters $\boldsymbol{\theta}$ and other variables) in a particular Bayesian model and let $\pi(\boldsymbol{\psi}|\boldsymbol{y}) \propto \pi(\boldsymbol{\psi})p(\boldsymbol{y}|\boldsymbol{\psi})$ denote the posterior density, where $\pi(\boldsymbol{\psi})$ is the prior density and $p(\boldsymbol{y}|\boldsymbol{\psi})$ is the sampling density. Suppose that we are interested in calculating the posterior mean $\boldsymbol{\eta} = \int_{\Re^d} \boldsymbol{\psi}\pi(\boldsymbol{\psi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\psi}$ and that this integral cannot be computed analytically, or numerically, because the dimension of the integration precludes the use of quadrature-based methods. To tackle this problem we can rely on Monte Carlo sampling methods. Instead of concentrating on the computation of the above integral we can proceed in a more general way. We construct a method to produce a sequence of draws from the posterior density, namely

$$\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)} \sim \pi(\boldsymbol{\psi}|\boldsymbol{y}).$$

Provided M is large enough, this sample from the posterior density can be used to summarize the posterior density. We can estimate the mean by taking the average of these simulated draws. We can estimate the quantiles of the posterior density from the quantiles of the sampled draws. Other summaries of the posterior density can be obtained in a similar manner. Under suitable laws of large numbers these estimates converge to the posterior quantities as the simulation-size becomes large.

The sampling of the posterior distribution is, therefore, the central concern in Bayesian computation. One important breakthrough in the use of simulation methods was the realization that the sampled draws need not be independent, that simulation-consistency can be achieved with correlated draws. The fact that the sampled variates can be correlated is of immense practical and theoretical importance and is the defining characteristic of Markov chain Monte Carlo methods, popularly referred to by the acronym MCMC, where the sampled draws form a Markov chain. The idea behind these methods is simple and extremely general. In order to sample a given probability distribution, referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution, is the target distribution. Once the Markov chain has been constructed, a sample of draws from the target distribution is obtained by simulating the Markov chain a large number of times and recording its values. Within the Bayesian framework, where both parameters and data are treated as random variables, and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods.

Markov chain sampling methods originate with the work of Metropolis et al. (1953) in statistical physics. A vital extension of the method was made by Hastings (1970) leading to a method that is now called the Metropolis–Hastings algorithm (see Tierney, 1994 and Chib and Greenberg, 1995 for detailed summaries). This algorithm was first applied to problems in spatial statistics and image analysis (Besag, 1974). A resurgence of interest in MCMC methods started with the papers of Geman and Geman (1984) who developed an algorithm, a special case of the Metropolis method that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987) who proposed a MCMC scheme involving data augmentation to sample posterior distributions in missing data problems, and Gelfand and Smith (1990) where the value of the Gibbs sampler was demonstrated for general Bayesian problems with continuous parameter spaces.

The Gibbs sampling algorithm is one of the simplest Markov chain Monte Carlo algorithms and is easy to describe. Suppose that $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ denote some grouping (blocking) of $\boldsymbol{\psi}$ and let

$$\pi_1(\boldsymbol{\psi}_1|\boldsymbol{y}, \boldsymbol{\psi}_2) \propto p(\boldsymbol{y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2), \tag{1.1}$$

$$\pi_2(\boldsymbol{\psi}_2|\boldsymbol{y}, \boldsymbol{\psi}_1) \propto p(\boldsymbol{y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) \tag{1.2}$$

denote the associated conditional densities, often called the full conditional densities. Then, one cycle of the Gibbs sampling algorithm is completed by sampling each of the full conditional densities, using the most current values of the conditioning block. The Gibbs sampler in which each block is revised in fixed order is defined as follows.

ALGORITHM (*Gibbs Sampling*).

1 Specify an initial value $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\psi}_1^{(0)}, \boldsymbol{\psi}_2^{(0)})$:

2 Repeat for $j = 1, 2, \ldots, n_0 + G$.

Generate $\boldsymbol{\psi}_1^{(j)}$ from $\pi_1(\boldsymbol{\psi}_1|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})$

Generate $\boldsymbol{\psi}_2^{(j)}$ from $\pi_2(\boldsymbol{\psi}_2|\boldsymbol{y}, \boldsymbol{\psi}_1^{(j)})$

3 Return the values $\{\boldsymbol{\psi}^{(n_0+1)}, \boldsymbol{\psi}^{(n_0+2)}, \ldots, \boldsymbol{\psi}^{(n_0+G)}\}$.

In some problems it turns out that the full conditional density cannot be sampled directly. In such cases, the intractable full conditional density is sampled via the Metropolis–Hastings (M–H) algorithm. For specificity, suppose that the full conditional density $\pi(\boldsymbol{\psi}_1|\boldsymbol{y}, \boldsymbol{\psi}_2)$ is intractable. Let

$$q_1\big(\boldsymbol{\psi}_1, \boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2\big)$$

denote a proposal density that generates a candidate $\boldsymbol{\psi}_1'$, given the data and the values of the remaining blocks. Then, in the first step of the jth iteration of the MCMC algorithm, given the values $(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_2^{(j-1)})$ from the previous iteration, the transition to the next iterate of $\boldsymbol{\psi}_1$ is achieved as follows.

ALGORITHM (*Metropolis–Hastings step for sampling an intractable* $\pi_1(\boldsymbol{\psi}_1|\boldsymbol{y}, \boldsymbol{\psi}_2)$).

1 Propose a value for $\boldsymbol{\psi}_1$ by drawing:

$$\boldsymbol{\psi}_1' \sim q_1(\boldsymbol{\psi}_1^{(j-1)}, \cdot|\boldsymbol{y},\ \boldsymbol{\psi}_2^{(j-1)})$$

2 Calculate the probability of move $\alpha(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})$ given by

$$\min\left\{1, \frac{\pi(\boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})q_1(\boldsymbol{\psi}_1', \boldsymbol{\psi}_1^{(j-1)}|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})}{\pi(\boldsymbol{\psi}_1^{(j-1)}|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})q_1(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)})}\right\}.$$

3 Set

$$\boldsymbol{\psi}_1^{(j)} = \begin{cases} \boldsymbol{\psi}_1' & \text{with prob } \alpha\big(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)}\big), \\ \boldsymbol{\psi}_1^{(j-1)} & \text{with prob } 1 - \alpha\big(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1'|\boldsymbol{y}, \boldsymbol{\psi}_2^{(j-1)}\big). \end{cases}$$

A similar approach is used to sample $\boldsymbol{\psi}_2$ if the full conditional density of $\boldsymbol{\psi}_2$ is intractable. These algorithms extend straightforwardly to problems with more than two blocks.

1.2. Computation of the marginal likelihood

Posterior simulation by MCMC methods does not require knowledge of the normalizing constant of the posterior density. Nonetheless, if we are interested in comparing alter-

native models, then knowledge of the normalizing constant is essential. This is because the formal Bayesian approach for comparing models is via *Bayes factors*, or ratios of *marginal likelihoods*. The marginal likelihood of a particular model is the normalizing constant of the posterior density and is defined as

$$m(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \tag{1.3}$$

the integral of the likelihood function with respect to the prior density. If we have two models $\mathcal{M}_k$ and $\mathcal{M}_l$, then the Bayes factor is the ratio

$$\mathrm{B}_{kl} = \frac{m(\boldsymbol{y}|\mathcal{M}_k)}{m(\boldsymbol{y}|\mathcal{M}_l)}. \tag{1.4}$$

Computation of the marginal likelihood is, therefore, of some importance in Bayesian statistics (DiCiccio et al., 1997, Chen and Shao, 1998, Roberts, 2001). It is possible to avoid the direct computation of the marginal likelihood by the methods of Green (1995) and Carlin and Chib (1995), but this requires formulating a more general MCMC scheme in which parameters and models are sampled jointly. We do not consider these approaches in this chapter. Unfortunately, because MCMC methods deliver draws from the posterior density, and the marginal likelihood is the integral with respect to the prior, the MCMC output cannot be used directly to average the likelihood. To deal with this problem, a number of methods have appeared in the literature. One simple and widely applicable method is due to Chib (1995) which we briefly explain as follows.

Begin by noting that $m(\boldsymbol{y})$ by virtue of being the normalizing constant of the posterior density can be expressed as

$$m(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})}, \tag{1.5}$$

for any given point $\boldsymbol{\theta}^*$ (generally taken to be a high density point such as the posterior mode or mean). Thus, provided we have an estimate $\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y})$ of the posterior ordinate, the marginal likelihood can be estimated on the log scale as

$$\log m(\boldsymbol{y}) = \log p\big(\boldsymbol{y}|\boldsymbol{\theta}^*\big) + \log \pi\big(\boldsymbol{\theta}^*\big) - \log \hat{\pi}\big(\boldsymbol{\theta}^*|\boldsymbol{y}\big). \tag{1.6}$$

In the context of both single and multiple block MCMC chains, good estimates of the posterior ordinate are available. For example, when the MCMC simulation is run with B blocks, Chib (1995) employs the marginal–conditional decomposition

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y}) = \pi\big(\boldsymbol{\theta}_1^*|\boldsymbol{y}\big) \times \cdots \times \pi\big(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \boldsymbol{\Theta}_{i-1}^*\big) \times \cdots \times \pi\big(\boldsymbol{\theta}_B^*|\boldsymbol{y}, \boldsymbol{\Theta}_{B-1}^*\big), \tag{1.7}$$

where the typical term is of the form

$$\begin{aligned}&\pi\big(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \boldsymbol{\Theta}_{i-1}^*\big)\\ &\quad= \int \pi\big(\boldsymbol{\theta}_i^*|\boldsymbol{y},\ \boldsymbol{\Theta}_{i-1}^*, \boldsymbol{\Theta}^{i+1}, z\big)\pi\big(\boldsymbol{\Theta}^{i+1}, z|\ \boldsymbol{y}, \boldsymbol{\Theta}_{i-1}^*\big)\,\mathrm{d}\boldsymbol{\Theta}^{i+1}\,\mathrm{d}z.\end{aligned}$$

In the latter expression $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i)$ and $\boldsymbol{\Theta}^i = (\boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_B)$ denote the list of blocks up to i and the set of blocks from i to B, respectively, and z denotes any latent

data that is included in the sampling. This is the *reduced conditional ordinate*. It is important to bear in mind that in finding the reduced conditional ordinate one must integrate only over $(\boldsymbol{\Theta}^{i+1}, z)$ and that the integrating measure is conditioned on $\boldsymbol{\Theta}^*_{i-1}$.

Consider first the case where the normalizing constant of each full conditional density is known. Then, the first term of (1.7) can be estimated by the Rao–Blackwell method. To estimate the typical reduced conditional ordinate, one conducts a MCMC run consisting of the full conditional distributions

$$\begin{aligned}&\{\pi(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\Theta}^*_{i-1}, \boldsymbol{\Theta}^{i+1}, z); \ldots; \boldsymbol{\pi}(\boldsymbol{\theta}_B|\boldsymbol{y}, \boldsymbol{\Theta}^*_{i-1}, \boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_{B-1}, z);\\ &\pi(z|\boldsymbol{y}, \boldsymbol{\Theta}^*_{i-1}, \boldsymbol{\Theta}^i)\},\end{aligned} \tag{1.8}$$

where the blocks in $\boldsymbol{\Theta}_{i-1}$ are set equal to $\boldsymbol{\Theta}^*_{i-1}$. By MCMC theory, the draws on $(\boldsymbol{\Theta}^{i+1}, z)$ from this run are from the distribution $\pi(\boldsymbol{\Theta}^{i+1}, z|\boldsymbol{y}, \boldsymbol{\Theta}^*_{i-1})$ and so the reduced conditional ordinate can be estimated as the average

$$\hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\Theta}^*_{i-1}) = M^{-1}\sum_{j=1}^{M}\pi(\boldsymbol{\theta}^*_i|\boldsymbol{y}, \boldsymbol{\Theta}^*_{i-1}, \boldsymbol{\Theta}^{i+1,(j)}, z^{(j)})$$

over the simulated values of $\boldsymbol{\Theta}^{i+1}$ and z from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (1.7) is estimated in the same way though, conveniently, with fewer and fewer distributions appearing in the reduced runs. Given the marginal and reduced conditional ordinates, the marginal likelihood on the log scale is available as

$$\begin{aligned}&\log \hat{m}(\boldsymbol{y}|\mathcal{M})\\ &\quad= \log p(\boldsymbol{y}|\boldsymbol{\theta}^*, \mathcal{M}) + \log \pi(\boldsymbol{\theta}^*) - \sum_{i=1}^{B}\log \hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\Theta}^*_{i-1}),\end{aligned} \tag{1.9}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}^*)$ is the density of the data marginalized over the latent data z.

Consider next the case where the normalizing constant of one or more of the full conditional densities is not known. In that case, the posterior ordinate is estimated by a modified method developed by Chib and Jeliazkov (2001). If sampling is conducted in one block by the M–H algorithm with proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\boldsymbol{y})$ and probability of move

$$\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}^{(j-1)}|\boldsymbol{y})}{\pi(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{y})q(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\boldsymbol{y})}\right\}$$

then it can be shown that the posterior ordinate is given by

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{E_1\{\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\boldsymbol{y})\}}{E_2\{\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\boldsymbol{y})\}},$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and the denominator expectation E_2 is with respect to the proposal density $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\boldsymbol{y})$. This

leads to the simulation consistent estimate

$$\hat{\pi}\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y})}{J^{-1}\sum_{j=1}^{M}\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)}|\boldsymbol{y})}, \tag{1.10}$$

where $\boldsymbol{\theta}^{(g)}$ are the given draws from the posterior distribution while the draws $\boldsymbol{\theta}^{(j)}$ in the denominator are from $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y})$, given the fixed value $\boldsymbol{\theta}^*$.

In general, when sampling is done with B blocks, the typical reduced conditional ordinate is given by

$$\begin{aligned}&\pi\left(\boldsymbol{\theta}_i^*|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*\right)\\ &\quad= \frac{E_1\{\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*,\boldsymbol{\Theta}^{i+1},z)q_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*,\boldsymbol{\Theta}^{i+1},z)\}}{E_2\{\alpha(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*,\boldsymbol{\Theta}^{i+1},z)\}},\end{aligned} \tag{1.11}$$

where E_1 is the expectation with respect to $\pi(\boldsymbol{\Theta}^i,z|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*)$ and E_2 that with respect to the product measure $\pi(\boldsymbol{\Theta}^{i+1},z|\boldsymbol{y},\boldsymbol{\theta}_i^*)q_i(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*,\boldsymbol{\Theta}^{i+1})$. The quantity $\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\boldsymbol{\Theta}_{i-1}^*,\boldsymbol{\Theta}^{i+1},z)$ is the M–H probability of move. The two expectations are estimated from the output of the reduced runs in an obvious way.

2. Binary responses

Suppose that y_i is a binary $\{0,1\}$ response variable and $\boldsymbol{x}_i$ is a k vector of covariates. Suppose that we have a random sample of n observations and the Bayesian model of interest is

$$\begin{aligned}y_i|\boldsymbol{\beta} &\sim \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right), \quad i \leqslant n,\\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0,\boldsymbol{B}_0),\end{aligned}$$

where Φ is the cdf of the $N(0,1)$ distribution. We consider the case of the logit link below. On letting $p_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$, the posterior distribution of $\boldsymbol{\beta}$ is proportional to

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta})\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)},$$

which does not belong to any named family of distributions. To deal with this model, and others involving categorical data, Albert and Chib (1993) introduce a technique that has led to many applications. The Albert–Chib algorithm capitalizes on the simplifications afforded by introducing latent or auxiliary data into the sampling.

Instead of the specification above, the model is re-specified as

$$\begin{aligned}z_i|\boldsymbol{\beta} &\sim \mathcal{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta},1\right),\\ y_i &= I[z_i > 0], \quad i \leqslant n,\\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0,\boldsymbol{B}_0).\end{aligned} \tag{2.1}$$

This specification is equivalent to the binary probit regression model since

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \Pr(z_i > 0|\boldsymbol{x}_i,\ \boldsymbol{\beta}) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right),$$

as required.

Albert and Chib (1993) exploit this equivalence and propose that the latent variables $\boldsymbol{z} = \{z_1, \ldots, z_n\}$, one for each observation, be included in the MCMC algorithm along with the regression parameter $\boldsymbol{\beta}$. In other words, they suggest using MCMC methods to sample the joint posterior distribution

$$\pi(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \mathcal{N}\left(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1\right) \Pr(y_i|z_i, \boldsymbol{\beta}).$$

The term $\Pr(y_i|z_i, \boldsymbol{\beta})$ is obtained by reasoning as follows: $\Pr(y_i = 0|z_i, \beta)$ is one if $z_i < 0$ (regardless of the value taken by $\boldsymbol{\beta}$) and zero otherwise – but this is the definition of $I(z_i < 0)$. Similarly, $\Pr(y_i = 1|z_i, \boldsymbol{\beta})$ is one if $z_i > 0$, and zero otherwise – but this is the definition of $I(z_i > 0)$. Collecting these two cases we have therefore that

$$\pi(\boldsymbol{z}, \boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \mathcal{N}\left(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1\right)\left\{I(z_i < 0)^{1-y_i} + I(z_i > 0)^{y_i}\right\}.$$

The latter posterior density is now sampled by a two-block Gibbs sampler composed of the full conditional distributions

$$\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta}; \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z}.$$

Even though the parameter space has been enlarged, the introduction of the latent variables simplifies the problem considerably. The second conditional distribution, i.e., $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z}$, is the same as the distribution $\boldsymbol{\beta}|\boldsymbol{z}$ since $\boldsymbol{z}$ implies $\boldsymbol{y}$, and hence $\boldsymbol{y}$ has no additional information for $\boldsymbol{\beta}$. The distribution $\boldsymbol{\beta}|\boldsymbol{z}$ is easy to derive by standard Bayesian results for a continuous response. The first conditional distribution, i.e., $\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta}$, factors into n distributions $z_i|y_i, \boldsymbol{\beta}$ and from the above is seen to be truncated normal given the value of y_i. Specifically,

$$p(z_i|y_i, \boldsymbol{\beta}) \propto \mathcal{N}\left(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1\right)\left\{I(z_i < 0)^{1-y_i} + I(z_i > 0)^{y_i}\right\}$$

which implies that if $y_i = 0$, the posterior density of z_i is proportional to $\mathcal{N}(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)I[z < 0]$, a truncated normal distribution with support $(-\infty, 0)$, whereas if $y_i = 1$, the density is proportional to $\mathcal{N}(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)I[z > 0]$, a truncated normal distribution with support $(0, \infty)$. These truncated normal distributions are simulated by applying the *inverse cdf* method (Ripley, 1987). Specifically, it can be shown that the draw

$$\mu + \sigma\Phi^{-1}\left[\Phi\left(\frac{a-\mu}{\sigma}\right) + U\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)\right], \tag{2.2}$$

where Φ^{-1} is the inverse cdf of the $\mathcal{N}(0, 1)$ distribution and $U \sim \text{Uniform}(0, 1)$, is from a $\mathcal{N}(\mu, \sigma^2)$ distribution truncated to the interval (a, b).

The Albert–Chib algorithm for the probit binary response model may now be summarized as follows.

ALGORITHM 1 (*Binary probit link model*).

1 Sample

$$z_i|y_i, \boldsymbol{\beta} \propto \mathcal{N}\big(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1\big)\big\{I(z_i < 0)^{1-y_i} + I(z_i > 0)^{y_i}\big\}, \quad i \leqslant n$$

2 Sample

$$\boldsymbol{\beta}|z \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\,\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i z_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

3 Go to 1

EXAMPLE. Consider the data in Table 1, taken from Fahrmeir and Tutz (1994), which is concerned with the occurrence or nonoccurence of infection following birth by cesarean section. The response variable y is one if the cesarean birth resulted in an infection, and zero if not. The available covariates are three indicator variables: x_1 is an indicator for whether the cesarean was nonplanned; x_2 is an indicator for whether risk factors were present at the time of birth and x_3 is an indicator for whether antibiotics were given as a prophylaxis. The data in the table contains information from 251 births. Under the column of the response, an entry such as 11/87 means that there were 98 deliveries with covariates (1, 1, 1) of whom 11 developed an infection and 87 did not.

Let us model the binary response by a *probit* model, letting the probability of infection for the ith birth be given as

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \Phi\big(\boldsymbol{x}_i'\boldsymbol{\beta}\big),$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})'$ is the covariate vector, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is the vector of unknown coefficients and Φ is the cdf of the standard normal random variable. Let us assume that our prior information about $\boldsymbol{\beta}$ can be represented by a multivariate normal density with mean centered at zero for each parameter, and variance given by $5\boldsymbol{I}_4$, where $\boldsymbol{I}_4$ is the four-dimensional identity matrix.

Algorithm 1 is run for 5000 cycles beyond a burn-in of 100 iterations. In Table 2 we provide the prior and posterior first two moments, and the 2.5th (lower) and 97.5th (upper) percentiles, of the marginal densities of $\boldsymbol{\beta}$. Each of the quantities is computed from the posterior draws in an obvious way, the posterior standard deviation in the table is

Table 1
Cesarean infection data

y (1/0)	x_1	x_2	x_3
11/87	1	1	1
1/17	0	1	1
0/2	0	0	1
23/3	1	1	0
28/30	0	1	0
0/9	1	0	0
8/32	0	0	0

Table 2
Cesarean data: Prior–posterior summary based on 5000 draws (beyond a burn-in of 100 cycles) from the Albert–Chib algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	3.162	−1.100	0.210	−1.523	−0.698
β_1	0.000	3.162	0.609	0.249	0.126	1.096
β_2	0.000	3.162	1.202	0.250	0.712	1.703
β_3	0.000	3.162	−1.903	0.266	−2.427	−1.393

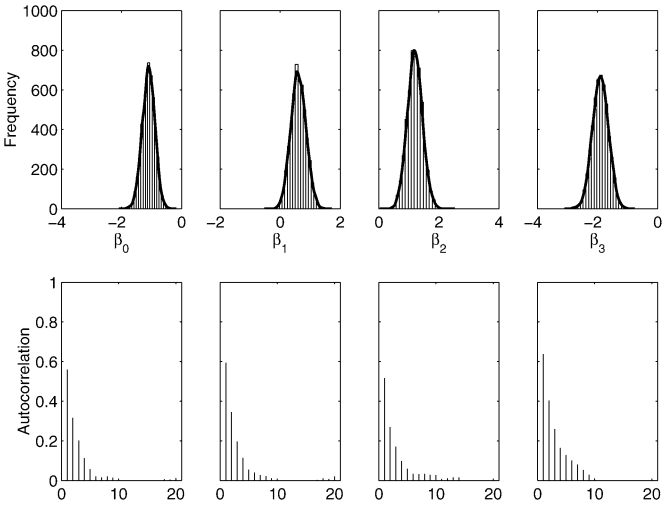

Fig. 1. Cesarean data with Albert–Chib algorithm: Marginal posterior densities (top panel) and autocorrelation plot (bottom panel).

the standard deviation of the sampled variates and the posterior percentiles are just the percentiles of the sampled draws. As expected, both the first and second covariates increase the probability of infection while the third covariate (the antibiotics prophylaxis) reduces the probability of infection.

To get an idea of the form of the posterior density we plot in Figure 1 the four marginal posterior densities. The density plots are obtained by smoothing the histogram of the simulated values with a Gaussian kernel. In the same plot we also report the autocorrelation functions (correlation against lag) for each of the sampled parameter values. The autocorrelation plots provide information of the extent of serial dependence in the sampled values. Here we see that the serial correlations decline quickly to zero indicating that the sampler is mixing well.

2.1. Marginal likelihood of the binary probit

The marginal likelihood of the binary probit model is easily calculated by the method of Chib (1995). Let $\boldsymbol{\beta}^* = E(\boldsymbol{\beta}|\boldsymbol{y})$, then from the expression of the marginal likelihood

in (1.6), we have that

$$\ln m(\boldsymbol{y}) = \ln \pi\big(\boldsymbol{\beta}^*\big) + \ln p\big(\boldsymbol{y}|\boldsymbol{\beta}^*\big) - \ln \pi\big(\boldsymbol{\beta}^*|\boldsymbol{y}\big),$$

where the first term is $\ln \mathcal{N}_k(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \boldsymbol{B}_0)$, the second is

$$\sum_{i=1}^{n} y_i \ln \Phi\big(\boldsymbol{x}_i'\boldsymbol{\beta}^*\big) + (1-y_i)\big(1 - \ln \Phi\big(\boldsymbol{x}_i'\boldsymbol{\beta}^*\big)\big)$$

and the third, which is

$$\ln \int \pi(\boldsymbol{\beta}|z)\pi(z|\boldsymbol{y})\,\mathrm{d}z$$

is estimated by averaging the conditional normal density of $\boldsymbol{\beta}$ in Step 2 of Algorithm 1 at the point $\boldsymbol{\beta}^*$ over the simulated values $\{z^{(g)}\}$ from the MCMC run. Specifically, letting $\hat{\pi}(\boldsymbol{\beta}^*|\boldsymbol{y})$ denote our estimate of the ordinate at $\boldsymbol{\beta}^*$ we have

$$\hat{\pi}\big(\boldsymbol{\beta}^*|\boldsymbol{y}\big) = M^{-1}\sum_{i=1}^{M} \mathcal{N}_k\left(\boldsymbol{\beta}^*|\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{x}_i z_i^{(g)}\right), \boldsymbol{B}_n\right).$$

2.2. Other link functions

The preceding algorithm is for the probit link function. Albert and Chib (1993) showed how it can be readily extended to other link functions that are generated from the scale mixture of normals family of distributions. For example, we could let

$$\begin{aligned} z_i|\boldsymbol{\beta}, \lambda_i &\sim \mathcal{N}\big(\boldsymbol{x}_i'\boldsymbol{\beta}, \lambda_i^{-1}\big), \\ y_i &= I[z_i > 0], \quad i \leqslant n, \\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0), \\ \lambda_i &\sim \mathcal{G}\left(\frac{\upsilon}{2}, \frac{\upsilon}{2}\right) \end{aligned} \tag{2.3}$$

in which case $\Pr(y_i = 1|\boldsymbol{\beta}) = F_{t,\upsilon}(\boldsymbol{x}_i'\beta)$, where $F_{t,\upsilon}$ is the cdf of the standard-t distribution with υ degrees of freedom. Albert and Chib (1993) utilized this setup with $\upsilon = 8$ to approximate the logit link function. In this case, the posterior distribution of interest is given by

$$\begin{aligned} \pi(\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \mathcal{N}\big(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \lambda_i^{-1}\big)\mathcal{G}\left(\lambda_i\Big|\frac{\upsilon}{2}, \frac{\upsilon}{2}\right) \\ \times \big\{I(z_i < 0)^{1-y_i} + I(z_i > 0)^{y_i}\big\} \end{aligned}$$

which can be sampled in two-blocks as

$$(\boldsymbol{z}, \boldsymbol{\lambda})|\boldsymbol{y}, \boldsymbol{\beta};\ \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\lambda},$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$. The first block is sampled by the method of composition by sampling $z_i|y_i, \boldsymbol{\beta}$ marginalized over λ_i from the truncated student-t density

$\mathcal{T}_\nu(z_i|\boldsymbol{x}_i'\boldsymbol{\beta},1)\{I(z_i<0)^{1-y_i}+I(z_i>0)^{y_i}\}$, followed by sampling $\lambda_i|z_i,\boldsymbol{\beta}$ from an updated Gamma distribution. The second distribution is normal with parameters obtained by standard Bayesian results.

ALGORITHM 2 (*Student-t binary model*).

1 Sample

(a)

$$z_i|y_i,\boldsymbol{\beta}\propto\mathcal{T}_\nu\big(z_i|\boldsymbol{x}_i'\boldsymbol{\beta},1\big)\big\{I(z_i<0)^{1-y_i}+I(z_i>0)^{y_i}\big\}$$

(b)

$$\lambda_i|z_i,\boldsymbol{\beta}\sim\mathcal{G}\left(\frac{\nu+1}{2},\frac{\nu+(z_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right),\quad i\leqslant n$$

2 Sample

$$\boldsymbol{\beta}|z,\boldsymbol{\lambda}\sim\mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sum_{i=1}^n\lambda_i\boldsymbol{x}_iz_i\right),\boldsymbol{B}_n=\left(\boldsymbol{B}_0^{-1}+\sum_{i=1}^n\lambda_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

3 Goto 1

Chen et al. (1999) present a further extension of this idea to model a skewed link function while Basu and Mukhopadhyay (2000) and Basu and Chib (2003) model the distribution of λ_i by a Dirichlet process prior leading to a semiparametric Bayesian binary data model. The logistic normal link function is considered by Allenby and Lenk (1994) with special reference to data arising in marketing.

2.3. Marginal likelihood of the student-t binary model

The marginal likelihood of the student-*t* binary model is also calculated by the method of Chib (1995). It is easily seen from the expression of the marginal likelihood in (1.6) that the only quantity that needs to be estimated is $\pi(\boldsymbol{\beta}^*|\boldsymbol{y})$, where $\boldsymbol{\beta}^*=E(\boldsymbol{\beta}^*|\boldsymbol{y})$. An estimate of this ordinate is obtained by averaging the conditional normal density of $\boldsymbol{\beta}$ in Step 2 of Algorithm 2 at the point $\boldsymbol{\beta}^*$ over the simulated values $\{z^{(g)}\}$ and $\{\lambda_i^{(g)}\}$ from the MCMC run. Specifically, letting $\hat{\pi}(\boldsymbol{\beta}^*|\boldsymbol{y})$ denote our estimate of the ordinate at $\boldsymbol{\beta}^*$ we have

$$\hat{\pi}\big(\boldsymbol{\beta}^*|\boldsymbol{y}\big)=M^{-1}\sum_{i=1}^M\mathcal{N}_k\left(\boldsymbol{B}_n^{(g)}\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sum_{i=1}^n\lambda_i^{(g)}\boldsymbol{x}_iz_i^{(g)}\right),\boldsymbol{B}_n^{(g)}\right),$$

where

$$\boldsymbol{B}_n^{(g)}=\left(\boldsymbol{B}_0^{-1}+\sum_{i=1}^n\lambda_i^{(g)}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}.$$

3. Ordinal response data

Albert and Chib (1993) extend their algorithm to the ordinal categorical data case where y_i can take one of the values $\{0, 1, \ldots, J\}$ according to the probabilities

$$\Pr(y_i \leqslant j|\boldsymbol{\beta}, \boldsymbol{c}) = F\big(c_j - \boldsymbol{x}_i'\boldsymbol{\beta}\big), \quad j = 0, 1, \ldots, J, \tag{3.1}$$

where F is some link function, say either Φ or $F_{t,v}$. In this model the $\{c_j\}$ are category specific cut-points such that $c_0 = 0$ and $c_J = \infty$. The remaining cut-points $\boldsymbol{c} = (c_1, \ldots, c_{J-1})$ are assumed to satisfy the order restriction $c_1 \leqslant \cdots \leqslant c_{J-1}$ which ensures that the cumulative probabilities are nondecreasing. Given data $y_1, \ldots, y_n$ from this model, the likelihood function is

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{c}) = \prod_{j=0}^{J} \prod_{i:y_i=j} \big[F\big(c_j - \boldsymbol{x}_i'\boldsymbol{\beta}\big) - F\big(c_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}\big)\big] \tag{3.2}$$

and the posterior density is proportional to $\pi(\boldsymbol{\beta}, \boldsymbol{c})\, p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{c})$, where $\pi(\boldsymbol{\beta}, \boldsymbol{c})$ is the prior distribution.

Simulation of the posterior distribution is again feasible by the tactical introduction of latent variables. For the link function $F_{t,v}$, we introduce the latent variables $\boldsymbol{z} = (z_1, \ldots, z_n)$, where $z_i|\boldsymbol{\beta}, \lambda_i \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, \lambda_i^{-1})$. A priori, we observe $y_i = j$ if the latent variable z_i falls in the interval $[c_{j-1}, c_j)$. Define $\delta_{ij} = 1$ if $y_i = j$, and 0 otherwise, then the posterior density of interest is

$$\pi(\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{c}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}, \boldsymbol{c}) \prod_{i=1}^{N} \mathcal{N}\big(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \lambda_i^{-1}\big)\mathcal{G}\left(\lambda_i \middle| \frac{v}{2}, \frac{v}{2}\right) \times \left\{\sum_{l=0}^{J} I\,(c_{l-1} < z_i < c_l)^{\delta_{il}}\right\},$$

where $c_{-1} = -\infty$. Now the basic Albert and Chib MCMC scheme draws cut-points, the latent data and the regression parameters in sequence. Albert and Chib (2001) simplified the latter step by transforming the cut-points so as to remove the ordering constraint. The transformation is defined by the one-to-one map

$$a_1 = \log c_1; \qquad a_j = \log(c_j - c_{j-1}), \quad 2 \leqslant j \leqslant J - 1. \tag{3.3}$$

The advantage of working with $\boldsymbol{a}$ instead of $\boldsymbol{c}$ is that the parameters of the tailored proposal density in the M–H step for $\boldsymbol{a}$ can be obtained by an unconstrained optimization and the prior $\pi(\boldsymbol{a})$ on $\boldsymbol{a}$ can be an unrestricted multivariate normal. Next, given $y_i = j$ and the cut-points the sampling of the latent data z_i is from $\mathcal{T}_v(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\, I\,(c_{j-1} < z_i < c_j)$, marginalized over $\{\lambda_i\}$. The sampling of the parameters $\boldsymbol{\beta}$ is as in Algorithm 2.

ALGORITHM 3 (*Ordinal student link model*).

```
1 M-H
  (a) Sample
```

$$\pi(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{\beta}) \propto \pi(\boldsymbol{a}) \underbrace{\prod_{j=0}^{J}\prod_{i:y_i=j}\left[F_{t,\nu}\left(c_j-\boldsymbol{x}_i'\boldsymbol{\beta}\right)-F_{t,\nu}\left(c_{j-1}-\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right]}_{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a})}$$

with the M-H algorithm by calculating

$$\boldsymbol{m} = \arg\max_{\boldsymbol{a}} \log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a})$$

and $\boldsymbol{V} = \{-\partial \log f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a})/\partial\boldsymbol{a}\partial\boldsymbol{a}'\}^{-1}$ the negative inverse of the hessian at $\boldsymbol{m}$, proposing

$$\boldsymbol{a}' \sim f_T(\boldsymbol{a}|\boldsymbol{m},\boldsymbol{V},\xi),$$

for some choice of ξ, calculating

$$\alpha(\boldsymbol{a},\boldsymbol{a}'|\boldsymbol{y},\boldsymbol{\beta}) = \min\left\{\frac{\pi(\boldsymbol{a}')p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a}')}{\pi(\boldsymbol{a})p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{a})}\frac{\mathcal{T}_\nu(\boldsymbol{a}|\boldsymbol{m},\boldsymbol{V},\xi)}{\mathcal{T}_\nu(\boldsymbol{a}'|\boldsymbol{m},\boldsymbol{V},\xi)},1\right\}$$

and moving to $\boldsymbol{a}'$ with probability $\alpha(\boldsymbol{a},\boldsymbol{a}'|\boldsymbol{y},\boldsymbol{\beta})$, then transforming the new $\boldsymbol{a}$ to $\boldsymbol{c}$ via the inverse map $c_j = \sum_{i=1}^{j}\exp(a_i)$, $1 \leqslant j \leqslant J-1$.

(b) Sample

$$z_i|y_i,\boldsymbol{\beta},\boldsymbol{c} \propto \mathcal{T}_\nu\left(z_i|\boldsymbol{x}_i'\boldsymbol{\beta},1\right)\left\{\sum_{l=0}^{J} I\left(c_{l-1}<z_i<c_l\right)^{\delta_{il}}\right\}, \quad i \leqslant n$$

(c) Sample

$$\lambda_i|y_i,z_i,\boldsymbol{\beta},\boldsymbol{c} \sim \mathcal{G}\left(\frac{\nu+1}{2},\frac{\nu+(z_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right), \quad i \leqslant n$$

2 Sample

$$\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{\lambda} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sum_{i=1}^{n}\lambda_i\boldsymbol{x}_i z_i\right),\boldsymbol{B}_n=\left(\boldsymbol{B}_0^{-1}+\sum_{i=1}^{n}\lambda_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

3 Goto 1

3.1. Marginal likelihood of the student-t ordinal model

The marginal likelihood of the student-t ordinal model can be calculated by the method of Chib (1995) and Chib and Jeliazkov (2001). Let $\boldsymbol{\beta}^* = E(\boldsymbol{\beta}|\boldsymbol{y})$ and $\boldsymbol{c}^* = E(\boldsymbol{c}^*|\boldsymbol{y})$ and let $\boldsymbol{a}^*$ be the transformed $\boldsymbol{c}^*$. Then from the expression of the marginal likelihood in (1.6) we can write

$$\ln m(\boldsymbol{y}) = \ln\pi\left(\boldsymbol{\beta}^*,\boldsymbol{a}^*\right)+\ln p\left(\boldsymbol{y}|\boldsymbol{\beta}^*,\boldsymbol{c}^*\right)-\ln\pi\left(\boldsymbol{a}^*|\boldsymbol{y}\right)-\ln\pi\left(\boldsymbol{\beta}^*|\boldsymbol{y},\boldsymbol{a}^*\right).$$

The marginal posterior ordinate $\pi(\boldsymbol{a}^*|\boldsymbol{y})$ from (1.11) can be written as

$$\pi\big(\boldsymbol{a}^*|\boldsymbol{y}\big) = \frac{E_1\{\alpha(\boldsymbol{a}, \boldsymbol{a}^*|\boldsymbol{y}, \boldsymbol{\beta})\mathcal{T}_\nu(\boldsymbol{a}'|\boldsymbol{m}, \boldsymbol{V}, \xi)\}}{E_2\{\alpha(\boldsymbol{a}^*, \boldsymbol{a}|\boldsymbol{y}, \boldsymbol{\beta})\}},$$

where E_1 is the expectation with respect to $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ and E_2 is the expectation with respect to $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{a}^*)\mathcal{T}_\nu(\boldsymbol{a}|\boldsymbol{m}, \boldsymbol{V}, \xi)$. The E_1 expectation can be estimated by draws from the main run and the E_2 expectation from a reduced run in which $\boldsymbol{a}$ is set to $\boldsymbol{a}^*$. For each value of $\boldsymbol{\beta}$ in this reduced run, $\boldsymbol{a}$ is also sampled from $\mathcal{T}_\nu(\boldsymbol{a}|\boldsymbol{m}, \boldsymbol{V}, \xi)$; these two draws (namely the draws of $\boldsymbol{\beta}$ and $\boldsymbol{a}$) are used to average $\alpha(\boldsymbol{a}^*, \boldsymbol{a}|\boldsymbol{y}, \boldsymbol{\beta})$. The final ordinate $\pi(\boldsymbol{\beta}^*|\boldsymbol{y}, \boldsymbol{a}^*)$ is estimated by averaging the normal conditional density in Step 2 of Algorithm 3 over the values of $\{z_i\}$ and $\{\lambda_i\}$ from this same reduced run.

4. Sequential ordinal model

The ordinal model presented above is appropriate when one can assume that a single unobservable continuous variable underlies the ordinal response. For example, this representation can be used for modeling ordinal letter grades in a mathematics class by assuming that the grade is an indicator of the student's unobservable continuous-valued intelligence. In other circumstances, however, a different latent variable structure may be required to model the ordinal response. For example, McCullagh (1980), and Farhmeir and Tutz (1994) consider a data set where the relative tonsil size of children is classified into the three states: "present but not enlarged", "enlarged", and "greatly enlarged". The objective is to explain the ordinal response as a function of whether the child is a carrier of the Streptococcus pyogenes. In this setting, it may not be appropriate to model the ordinal tonsil size in terms of a single latent variable. Rather, it may be preferable to imagine that each response is determined by two continuous latent variables where the first latent variable measures the propensity of the tonsil to grow abnormally and pass from the state "present but not enlarged" to the state "enlarged". The second latent variable measures the propensity of the tonsil to grow from the "enlarged" to the "greatly enlarged" states. Thus, the final "greatly enlarged" state is realized only when the tonsil has passed through the earlier levels. This latent variable representation leads to a *sequential* model because the levels of the response are achieved in a sequential manner.

Suppose that one observes independent observations $y_1, \ldots, y_n$, where each y_i is an ordinal categorical response variable with J possible values $\{1, \ldots, J\}$. Associated with the ith response y_i, let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ denote a set of k covariates. In the sequential ordinal model, the variable y_i can take the value j only after the levels $1, \ldots, j-1$ are reached. In other words, to get to the outcome j, one must pass through levels $1, 2, \ldots, j-1$ and stop (or fail) in level j. The probability of stopping in level j $(1 \leqslant j \leqslant J-1)$, conditional on the event that the jth level is reached, is given by

$$\Pr(y_i = j|y_i \geqslant j, \boldsymbol{\delta}, \boldsymbol{c}) = F\big(c_j - \boldsymbol{x}_i'\boldsymbol{\delta}\big), \tag{4.1}$$

where $\boldsymbol{c} = (c_1, \ldots, c_{J-1})$ are unordered cutpoints and $\boldsymbol{x}_i'\boldsymbol{\delta}$ represents the effect of the covariates. This probability function is referred to as the discrete time hazard function

(Tutz, 1990; Farhmeir and Tutz, 1994). It follows that the probability of stopping at level j is given by

$$\begin{aligned}\Pr(y_i = j|\boldsymbol{\delta}, \boldsymbol{c}) &= \Pr(y_i = j|y_i \geqslant j, \boldsymbol{\delta}, \boldsymbol{c}) \Pr(y_i \geqslant j) \\ &= F\big(c_j - \boldsymbol{x}_i'\boldsymbol{\delta}\big) \prod_{k=1}^{j-1} \big\{1 - F\big(c_k - \boldsymbol{x}_i'\boldsymbol{\delta}\big)\big\}, \quad j \leqslant J-1, \quad (4.2)\end{aligned}$$

whereas the probability that the final level J is reached is

$$\Pr(y_i = J|\boldsymbol{\delta}, \boldsymbol{c}) = \prod_{k=1}^{J-1} \big\{1 - F\big(c_k - \boldsymbol{x}_i'\boldsymbol{\delta}\big)\big\}, \quad (4.3)$$

since the event $y_i = J$ occurs only if all previous $J-1$ levels are passed.

The sequential model is formally equivalent to the continuation-ratio ordinal models, discussed by Agresti (1990, Chapter 9), Cox (1972) and Ten Have and Uttal (2000). The sequential model is useful in the analysis of discrete-time survival data (Farhmeir and Tutz (1994), Chapter 9; Kalbfleish and Prentice, 1980). More generally, the sequential model can be used to model nonproportional and nonmonotone hazard functions and to incorporate the effect of time dependent covariates. Suppose that one observes the time to failure for subjects in the sample, and let the time interval be subdivided (or grouped) into the J intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{J-1}, \infty)$. Then one defines the event $y_i = j$ if a failure is observed in the interval $[a_{j-1}, a_j)$. The discrete hazard function (4.1) now represents the probability of failure in the time interval $[a_{j-1}, a_j)$ given survival until time a_{j-1}. The vector of cutpoint parameters $\boldsymbol{c}$ represents the baseline hazard of the process.

Albert and Chib (2001) develop a full Bayesian analysis of this model based on the framework of Albert and Chib (1993). Suppose that F is the distribution function of the standard normal distribution, leading to the sequential ordinal model. Corresponding to the ith observation, define latent variables $\{w_{ij}\}$, where $w_{ij} = \boldsymbol{x}_i'\boldsymbol{\delta} + e_{ij}$ and the e_{ij} are independently distributed from $N(0, 1)$. We observe $y_i = 1$ if $w_{i1} \leqslant c_1$, and observe $y_i = 2$ if the first latent variable $w_{i1} > c_1$ and the second variable $w_{i2} \leqslant c_2$. In general, we observe $y_i = j\,(1 \leqslant j \leqslant J-1)$ if the first $j-1$ latent variables exceed their corresponding cutoffs, and the jth variable does not: $y_i = j$ if $w_{i1} > c_1, \ldots, w_{ij-1} > c_{j-1}, w_{ij} \leqslant c_j$. In this model, the latent variable w_{ij} represents one's propensity to continue to the $(j+1)$st level in the sequence, given that the individual has already attained level j.

This latent variable representation can be simplified by incorporating the cutpoints $\{\gamma_j\}$ into the mean function. Define the new latent variable $z_{ij} = w_{ij} - c_j$. Then it follows that $z_{ij}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{x}_{ij}'\boldsymbol{\beta}, 1)$, where $\boldsymbol{x}_{ij}' = (0, 0, \ldots, -1, 0, 0, \boldsymbol{x}_i')$ with -1 in the jth column, and $\boldsymbol{\beta} = (c_1, \ldots, c_{J-1}, \boldsymbol{\delta}')'$. The observed data are then generated

according to

$$y_i = \begin{cases} 1 & \text{if } z_{i1} \leqslant 0, \\ 2 & \text{if } z_{i1} > 0, z_{i2} \leqslant 0, \\ \vdots & \vdots \\ J-1 & \text{if } z_{i1} > 0, \ldots, z_{iJ-2} > 0, z_{iJ-1} \leqslant 0, \\ J & \text{if } z_{i1} > 0, \ldots, z_{iJ-1} > 0. \end{cases} \tag{4.4}$$

Following the Albert and Chib (1993) approach, this model can be fit by MCMC methods by simulating the joint posterior distribution of $(\{z_{ij}\}, \boldsymbol{\beta})$. The latent data $\{z_{ij}\}$, conditional on $(\boldsymbol{y}, \boldsymbol{\beta})$, are independently distributed as truncated normal. Specifically, if $y_i = j$, then the latent data corresponding to this observation are represented as $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ij_i})$ where $j_i = \min\{j, J-1\}$ and the simulation of $\boldsymbol{z}_i$ is from a sequence of truncated normal distributions. The posterior distribution of the parameter vector $\boldsymbol{\beta}$, conditional on the latent data $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, has a simple form. Let $\boldsymbol{X}_i$ denote the covariate matrix corresponding to the ith subject consisting of the j_i rows $\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{ij_i}$. If $\boldsymbol{\beta}$ is assigned a multivariate normal prior with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{B}_0$, then the posterior distribution of $\boldsymbol{\beta}$, conditional on $\boldsymbol{z}$, is multivariate normal.

ALGORITHM 4 (*Sequential ordinal probit model*).

`1 Sample` $z_i|y_i = j, \boldsymbol{\beta}$

(a) `If` $y_i = 1$, `sample`

$$z_{i1}|y_i, \boldsymbol{\beta} \propto \mathcal{N}\big(\boldsymbol{x}'_{i1}\boldsymbol{\beta}, 1\big) I(z_{i1} < 0)$$

(b) `If` $y_i = j$ $(2 \leqslant j \leqslant J-1)$, `sample`

$$z_{ik}|y_i, \boldsymbol{\beta} \propto \mathcal{N}\big(\boldsymbol{x}'_{ik}\boldsymbol{\beta}, 1\big) I(z_{ik} > 0), \quad k = 1, \ldots, j-1$$
$$z_{ij}|y_i, \boldsymbol{\beta} \propto \mathcal{N}\big(\boldsymbol{x}'_{ij}\boldsymbol{\beta}, 1\big) I(z_{ij} < 0)$$

(c) `If` $y_i = J$, `sample`

$$z_{ik}|y_i, \boldsymbol{\beta} \propto \mathcal{N}\big(\boldsymbol{x}'_{ik}\boldsymbol{\beta}, 1\big) I(z_{ik} > 0), \quad k = 1, \ldots, J-1$$

`2 Sample`

$$\boldsymbol{\beta}|\boldsymbol{z} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}'_i \boldsymbol{z}_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}'_i \boldsymbol{X}_i\right)^{-1}\right)$$

`3 Goto 1`

Albert and Chib (2001) present generalizations of the sequential ordinal model and also discuss the computation of the marginal likelihood by the method of Chib (1995).

5. Multivariate responses

One of the considerable achievements of the recent Bayesian literature is in the development of methods for multivariate categorical data. A leading case concerns that

of multivariate binary data. From the frequentist perspective, this problem is tackled by Carey et al. (1993) and Glonek and McCullagh (1995) in the context of the multivariate logit model, and from a Bayesian perspective by Chib and Greenberg (1998) in the context of the multivariate probit (MVP) model (an analysis of the multivariate logit model from a Bayesian perspective is supplied by O'Brien and Dunson, 2004). In many respects, the multivariate probit model, along with its variants, is the canonical modeling approach for correlated binary data. This model was introduced more than twenty five years ago by Ashford and Sowden (1970) in the context of bivariate binary responses and was analyzed under simplifying assumptions on the correlation structure by Amemiya (1972), Ochi and Prentice (1984) and Lesaffre and Kaufmann (1992). The general version of the model was considered to be intractable. Chib and Greenberg (1998), building on the framework of Albert and Chib (1993), resolve the difficulties.

The MVP model provides a relatively straightforward way of modeling multivariate binary data. In this model, the marginal probability of each response is given by a probit function that depends on covariates and response specific parameters. Associations between the binary variables are incorporated by assuming that the vector of binary outcomes are a function of correlated Gaussian variables, taking the value one if the corresponding Gaussian component is positive and the value zero otherwise. This connection with a latent Gaussian random vector means that the regression coefficients can be interpreted independently of the correlation parameters (unlike the case of log-linear models). The link with Gaussian data is also helpful in estimation. By contrast, models based on marginal odds ratios Connolly and Liang (1988) tend to proliferate nuisance parameters as the number of variables increase and they become difficult to interpret and estimate. Finally, the Gaussian connection enables generalizations of the model to ordinal outcomes (see Chen and Dey, 2000), mixed outcomes (Kuhnert and Do, 2003), and spatial data (Fahrmeir and Lang, 2001; Banerjee et al., 2004).

5.1. Multivariate probit model

Let y_{ij} denote a binary response on the ith observation unit and jth variable, and let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$, $1 \leqslant i \leqslant n$, denote the collection of responses on all J variables. Also let $\boldsymbol{x}_{ij}$ denote the set of covariates for the jth response and $\boldsymbol{\beta}_j \in R^{k_j}$ the conformable vector of covariate coefficients. Let $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J') \in R^k$, $k = \sum k_j$ denote the complete set of coefficients and let $\boldsymbol{\Sigma} = \{\sigma_{jk}\}$ denote a $J \times J$ correlation matrix. Finally let

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}' & \boldsymbol{0}' & \cdots & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{x}_{i2}' & \boldsymbol{0}' & \boldsymbol{0}' \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0}' & \boldsymbol{0}' & \cdots & \boldsymbol{x}_{iJ}' \end{pmatrix}$$

denote the $J \times k$ covariate matrix on the ith subject. Then, according to the multivariate probit model, the marginal probability that $y_{ij} = 1$ is given by the probit form

$$\Pr(y_{ij} = 1|\boldsymbol{\beta}) = \Phi\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j\right)$$

and the joint probability of a given outcome y_i, conditioned on parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and covariates $\boldsymbol{x}_{ij}$, is

$$\Pr(y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{A_{iJ}} \dots \int_{A_{i1}} \mathcal{N}_J(t|0, \boldsymbol{\Sigma})\, \mathrm{d}t, \tag{5.1}$$

where $\mathcal{N}_J(t|0, \boldsymbol{\Sigma})$ is the density of a J-variate normal distribution with mean vector 0 and correlation matrix $\boldsymbol{\Sigma}$ and A_{ij} is the interval

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j) & \text{if } y_{ij} = 1, \\ [\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j, \infty) & \text{if } y_{ij} = 0. \end{cases} \tag{5.2}$$

Note that each outcome is determined by its own set of k_j covariates $\boldsymbol{x}_{ij}$ and covariate effects $\boldsymbol{\beta}_j$.

The multivariate discrete mass function presented in (5.1) can be specified in terms of latent Gaussian random variables. This alternative formulation also forms the basis of the computational scheme that is described below. Let $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iJ})$ denote a J-variate normal vector and let

$$\boldsymbol{z}_i \sim \mathcal{N}_J(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{5.3}$$

Now let y_{ij} be 1 or 0 according to the sign of z_{ij}:

$$y_{ij} = I(z_{ij} > 0), \quad j = 1, \dots, J. \tag{5.4}$$

Then, the probability in (5.1) may be expressed as

$$\int_{B_{iJ}} \dots \int_{B_{i1}} \mathcal{N}_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{z}_i, \tag{5.5}$$

where B_{ij} is the interval $(0, \infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$. It is easy to confirm that this integral reduces to the form given above. It should also be noted that due to the threshold specification in (5.4), the scale of z_{ij} cannot be identified. As a consequence, the matrix $\boldsymbol{\Sigma}$ must be in correlation form (with units on the main diagonal).

5.2. Dependence structures

One basic question in the analysis of correlated binary data is the following: How should correlation between binary outcomes be defined and measured? The point of view behind the MVP model is that the correlation is modeled at the level of the latent data which then induces correlation amongst the binary outcomes. This modeling perspective is both flexible and general. In contrast, attempts to model correlation directly (as in the classical literature using marginal odds ratios) invariably lead to difficulties, partly because it is difficult to specify pair-wise correlations in general, and partly because the binary data scale is not natural for thinking about dependence.

Within the context of the MVP model, alternative dependence structures are easily specified and conceived, due to the connection with Gaussian latent data. Some of the possibilities are enumerated below.

- Unrestricted form. Here $\boldsymbol{\Sigma}$ is fully unrestricted except for the unit constraints on the diagonal. The unrestricted $\boldsymbol{\Sigma}$ matrix has $J^* = J(J-1)/2$ unknown correlation parameters that must be estimated.
- Equicorrelated form. In this case, the correlations are all equal and described by a single parameter ρ. This form can be a starting point for the analysis when one is dealing with outcomes where are all the pair-wise correlations are believed to have the same sign. The equicorrelated model may be also be seen as arising from a random effect formulation.
- Toeplitz form. Under this case, the correlations depend on a single parameter ρ but under the restriction that $\text{Corr}(z_{ik}, z_{il}) = \rho^{|k-l|}$. This version can be useful when the binary outcomes are collected from a longitudinal study where it is plausible that the correlation between outcomes at different dates will diminish with the lag. In fact, in the context of longitudinal data, the $\boldsymbol{\Sigma}$ matrix can be specified in many other forms with analogy with the correlation structures that arise in standard time series ARMA modeling.

5.3. Student-t specification

Now suppose that the distribution on the latent z_i is multivariate-t with specified degrees of freedom ν. This gives rise to a model that may be called the multivariate-t link model. Under the multivariate-t assumption, $z_i|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{T}_\nu(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with density

$$\mathcal{T}_\nu(z_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2}\left\{1 + \frac{1}{\nu}(z_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(z_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}^{-(\nu+J)/2}. \quad (5.6)$$

As before, the matrix $\boldsymbol{\Sigma}$ is in correlation form and the observed outcomes are defined by (5.4). The model for the latent z_i may be expressed as a scale mixture of normals by introducing a random variable $\lambda_i \sim \mathcal{G}(\frac{\nu}{2}, \frac{\nu}{2})$ and letting

$$z_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \lambda_i \sim \mathcal{N}_J\left(\boldsymbol{X}_i\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Sigma}\right). \quad (5.7)$$

Conditionally on λ_i, this model is equivalent to the MVP model.

5.4. Estimation of the MVP model

Consider now the question of inference in the MVP model. We are given a set of data on n subjects with outcomes $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$ and interest centers on the parameters of the model $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y})$, given some prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ on the parameters.

To simplify the MCMC implementation for this model Chib and Greenberg (1998) follow the general approach of Albert and Chib (1993) and employ the latent variables $z_i \sim \mathcal{N}_J(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with the observed data given by $y_{ij} = I(z_{ij} > 0)$, $j = 1, \ldots, J$. Let $\boldsymbol{\sigma} = (\sigma_{12}, \sigma_{31}, \sigma_{32}, \ldots, \sigma_{J,J})$ denote the $J(J-1)/2$ distinct elements of $\boldsymbol{\Sigma}$. It can be shown that the admissible values of $\boldsymbol{\sigma}$ (that lead to a positive definite $\boldsymbol{\Sigma}$ matrix) form a convex solid body in the hypercube $[-1, 1]^p$. Denote this set by C. Now let $\boldsymbol{z} = (z_1, \ldots, z_n)$ denote the latent values corresponding to the observed data $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$.

Then, the algorithm proceeds with the sampling of the posterior density

$$\pi(\boldsymbol{\beta}, \boldsymbol{\sigma}, z|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}, \boldsymbol{\sigma}) f(z|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \Pr(\boldsymbol{y}|z, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$
$$\propto \pi(\boldsymbol{\beta}, \boldsymbol{\sigma}) \prod_{i=1}^{n} \big(\phi_J(z_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \Pr(\boldsymbol{y}_i|z_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})\big), \quad \boldsymbol{\beta} \in \Re^k, \boldsymbol{\sigma} \in C,$$

where, akin to the result in the univariate binary case,

$$\Pr(\boldsymbol{y}_i|z_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J} \big\{ I(z_{ij} \leqslant 0)^{1-y_{ij}} + I(z_{ij} > 0)^{y_{ij}} \big\}.$$

Conditioned on $\{z_i\}$ and $\boldsymbol{\Sigma}$, the update for $\boldsymbol{\beta}$ is straightforward, while conditioned on $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, z_{ij} can be sampled one at a time conditioned on the other latent values from truncated normal distributions, where the region of truncation is either $(0, \infty)$ or $(-\infty, 0)$ depending on whether the corresponding y_{ij} is one or zero. The key step in the algorithm is the sampling of $\boldsymbol{\sigma}$, the unrestricted elements of $\boldsymbol{\Sigma}$, from the full conditional density $\pi(\boldsymbol{\sigma}|z, \boldsymbol{\beta}) \propto p(\boldsymbol{\sigma}) \prod_{i=1}^{n} \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$. This density, which is truncated to the complicated region C, is sampled by a M–H step with tailored proposal density $q(\boldsymbol{\sigma}|z, \boldsymbol{\beta}) = \mathcal{T}_v(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)$ where

$$\boldsymbol{m} = \arg\max_{\boldsymbol{\sigma} \in C} \sum_{i=1}^{n} \ln \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$
$$\boldsymbol{V} = -\left\{ \frac{\partial^2 \sum_{i=1}^{n} \ln \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} \right\}^{-1}_{\boldsymbol{\sigma} = \boldsymbol{m}}$$

are the mode and curvature of the target distribution, given the current values of the conditioning variables.

ALGORITHM 5 (*Multivariate probit*).

1 Sample for $i \leqslant n$, $j \leqslant J$

$$z_{ij}|\boldsymbol{z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \propto \mathcal{N}(\mu_{ij}, v_{ij}) \big\{ I(z_{ij} \leqslant 0)^{1-y_{ij}} + I(z_{ij} > 0)^{y_{ij}} \big\}$$
$$\mu_{ij} = \mathrm{E}(z_{ij}|\boldsymbol{z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$
$$v_{ij} = \mathrm{Var}(z_{ij}|\boldsymbol{z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$

2 Sample

$$\boldsymbol{\beta}|z, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_k\left(\boldsymbol{B}_n \left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} z_i \right), \boldsymbol{B}_n \right),$$

where

$$\boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i^{-1} \right)^{-1}$$

3 Sample

$$\pi(\boldsymbol{\sigma}|z, \boldsymbol{\beta}) \propto p(\boldsymbol{\sigma}) \prod_{i=1}^{n} \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

by the M-H algorithm, calculating the parameters $(\boldsymbol{m}, \boldsymbol{V})$, proposing $\boldsymbol{\sigma}' \sim \mathcal{T}_\nu(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)$, calculating

$$\alpha(\boldsymbol{\sigma}, \boldsymbol{\sigma}'|\boldsymbol{y}, \boldsymbol{\beta}, \{z_i\}) = \min\left\{ \frac{\pi(\boldsymbol{\sigma}') \prod_{i=1}^{n} \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}') I[\boldsymbol{\sigma}' \in C]}{\pi(\boldsymbol{\sigma}) \prod_{i=1}^{n} \mathcal{N}_J(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})} \frac{\mathcal{T}_\nu(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)}{\mathcal{T}_\nu(\boldsymbol{\sigma}'|\boldsymbol{m}, \boldsymbol{V}, \xi)}, 1 \right\}$$

and moving to $\boldsymbol{\sigma}'$ with probability $\alpha(\boldsymbol{\sigma}, \boldsymbol{\sigma}'|\boldsymbol{y}, \boldsymbol{\beta}, \{z_i\})$

4 Goto 1

EXAMPLE. As an application of this algorithm consider a data set in which the multivariate binary responses are generated by a panel structure. The data is concerned with the health effects of pollution on 537 children in Stuebenville, Ohio, each observed at ages 7, 8, 9 and 10 years, and the response variable is an indicator of wheezing status. Suppose that the marginal probability of wheeze status of the ith child at the jth time point is specified as

$$\Pr(y_{ij} = 1|\boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij}), \quad i \leqslant 537,\ j \leqslant 4,$$

where $\boldsymbol{\beta}$ is constant across categories, x_1 is the age of the child centered at nine years, x_2 is a binary indicator variable representing the mother's smoking habit during the first year of the study, and $x_3 = x_1 x_2$. Suppose that the Gaussian prior on $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ is centered at zero with a variance of $10\boldsymbol{I}_k$ and let $\pi(\boldsymbol{\sigma})$ be the density of a normal distribution, with mean zero and variance $\boldsymbol{I}_6$, *restricted* to region that leads to a positive-definite correlation matrix, where $(\sigma_{21}, \sigma_{31}, \sigma_{32}, \sigma_{41}, \sigma_{42}, \sigma_{43})$. From 10,000 cycles of Algorithm 5 one obtains the following covariate effects and posterior distributions of the correlations.

Table 3
Covariate effects in the Ohio wheeze data: MVP model with unrestricted correlations. In the table, NSE denotes the numerical standard error, lower is the 2.5th percentile and upper is the 97.5th percentile of the simulated draws. The results are based on 10000 draws from Algorithm 5

	Prior		Posterior				
	Mean	Std dev	Mean	NSE	Std dev	Lower	Upper
$\boldsymbol{\beta}_1$	0.000	3.162	−1.108	0.001	0.062	−1.231	−0.985
$\boldsymbol{\beta}_2$	0.000	3.162	−0.077	0.001	0.030	−0.136	−0.017
$\boldsymbol{\beta}_3$	0.000	3.162	0.155	0.002	0.101	−0.043	0.352
$\boldsymbol{\beta}_4$	0.000	3.162	0.036	0.001	0.049	−0.058	0.131

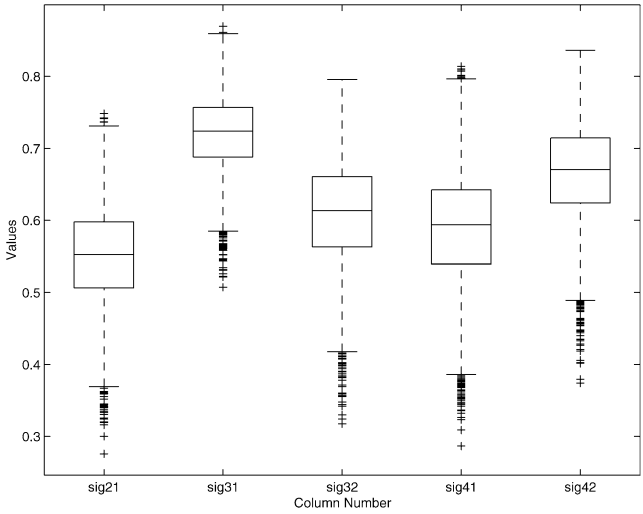

Fig. 2. Posterior boxplots of the correlations in the Ohio wheeze data: MVP model.

Notice that the summary tabular output contains not only the posterior means and standard deviations of the parameters but also the 95% credibility intervals, all computed from the sampled draws. It may be seen from Figure 2 that the posterior distributions of the correlations are similar suggesting that an equicorrelated correlation structure might be appropriate for these data.

5.5. *Marginal likelihood of the MVP model*

The calculation of the marginal likelihood of MVP model is considered by both Chib and Greenberg (1998) and Chib and Jeliazkov (2001). The main issue is the calculation of the posterior ordinate which may be estimated from

$$\pi\left(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^* | \boldsymbol{y}\right) = \pi\left(\boldsymbol{\sigma}^* | \boldsymbol{y}\right) \pi\left(\boldsymbol{\beta}^* | \boldsymbol{y}, \boldsymbol{\Sigma}^*\right),$$

where the first ordinate (that of the correlations) is not in tractable form. To estimate the marginal ordinate one can apply (1.11) leading to

$$\pi\left(\boldsymbol{\sigma}^* | \boldsymbol{y}\right) = \frac{E_1\{\alpha(\boldsymbol{\sigma}, \boldsymbol{\sigma}' | \boldsymbol{y}, \boldsymbol{\beta}, \{z_i\}) \mathcal{T}_v(\boldsymbol{\sigma}' | \boldsymbol{m}, \boldsymbol{V}, \xi)\}}{E_2\{\alpha(\boldsymbol{\sigma}^*, \boldsymbol{\sigma} | \boldsymbol{y}, \boldsymbol{\beta}, \{z_i\})\}},$$

where $\alpha(\boldsymbol{\sigma}, \boldsymbol{\sigma}' | \boldsymbol{y}, \boldsymbol{\beta}, \{z_i\})$ is the probability of move defined in Algorithm 5. The numerator expectation can be calculated from the draws $\{\boldsymbol{\beta}^{(g)}, \{z_i^{(g)}\}, \boldsymbol{\sigma}^{(g)}\}$ from the output of Algorithm 5. The denominator expectation is calculated from a reduced run consisting of the densities

$$\pi\left(\boldsymbol{\beta} | \boldsymbol{y}, \{z_i\}, \boldsymbol{\Sigma}^*\right); \quad \pi\left(\{z_i\} | \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*\right),$$

after $\boldsymbol{\Sigma}$ is fixed at $\boldsymbol{\Sigma}^*$. Then for each draw

$$\boldsymbol{\beta}^{(j)}, z^{(j)} \sim \pi\left(\boldsymbol{\beta}, z | \boldsymbol{y}, \boldsymbol{\Sigma}^*\right)$$

in this reduced run, we also draw $\boldsymbol{\sigma}^{(j)} \sim \mathcal{T}_\nu(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)$. These draws are used to average $\alpha(\boldsymbol{\sigma}^*, \boldsymbol{\sigma}|\boldsymbol{y}, \boldsymbol{\beta}, \{z_i\})$. In addition, the sampled variates $\{\boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)}\}$ from this same reduced run are also used to estimate $\pi(\boldsymbol{\beta}^*|\boldsymbol{y}, \boldsymbol{\Sigma}^*)$ by averaging the normal density in Step 2 of Algorithm 5.

5.6. *Fitting of the multivariate t-link model*

Algorithm 5 is easily modified for the fitting of the multivariate-t link version of the MVP model. One simple possibility is to include the $\{\lambda_i\}$ into the sampling since conditioned on the value of λ_i, the t-link binary model reduces to the MVP model. With this augmentation, one implements Steps 1–3 conditioned on the value of $\{\lambda_i\}$. The sampling is completed with an additional step involving the simulation of $\{\lambda_i\}$. A straightforward calculation shows that the updated full conditional distribution of λ_i is

$$\lambda_i|z_i, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{G}\left(\frac{\nu + J}{2}, \frac{\nu + (z_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(z_i - \boldsymbol{X}_i\boldsymbol{\beta})}{2}\right), \quad i \leqslant n,$$

which is easily sampled. The modified sampler thus requires little extra coding.

5.7. *Binary outcome with a confounded binary treatment*

In many problems, one central objective of model building is to isolate the effect of a binary covariate on a binary (or categorical) outcome. Isolating this effect becomes difficult if the effect of the binary covariate on the outcome is confounded with those of missing or unobserved covariates. Such complications are the norm when dealing with observational data. As an example, suppose that y_i is a zero-one variable that represents post-surgical outcomes for hip-fracture patients (one if the surgery is successful, and zero if not) and x_i is an indicator representing the extent of delay before surgery (one if the delay exceeds say two days, and zero otherwise). In this case, it is quite likely that delay and post-surgical outcome are both affected by patient factors, e.g., patient frailty, that are usually either imperfectly measured and recorded or absent from the data. In this case, we can model the outcome and the treatment assignment jointly to control for the unobservable confounders. The resulting model is similar to the MVP model and is discussed in detail by Chib (2003).

Briefly, suppose that the outcome model is given by

$$y_i = I\left(z_{1i}'\boldsymbol{\gamma}_1 + x_i\beta + \varepsilon_i\right), \quad i \leqslant n, \tag{5.8}$$

and suppose that the effect of the binary treatment x_i is confounded with that of unobservables that are correlated with ε_i, even after conditioning on z_{1i}. To model this confounding, let the treatment assignment mechanism be given by

$$x_i = I\left(z_i'\boldsymbol{\gamma} + u_i\right),$$

where the confounder u_i is correlated with ε_i and the covariate vector $z_i = (z_{1i}, z_{2i})$ contains the variables z_{2i} that are part of the intake model but not present in the outcome model. These covariates are called instrumental variables. They must be independent of (ε_i, u_i) given z_{1i} but correlated with the intake. Now suppose for simplicity that the joint

distribution of the unobservables (ε_i, u_i) is Gaussian with mean zero and covariance matrix $\boldsymbol{\Omega}$ that is in correlation form. It is easy to see, starting with the joint distribution of (ε_i, u_i) followed by a change of variable, that this set-up is a special case of the MVP model. It can be estimated along the lines of Algorithm 5.

Chib (2003) utilizes the posterior draws to calculate the average treatment effect (ATE) which is defined as $E(y_{i1}|\boldsymbol{z}_{1i}, \boldsymbol{\gamma}_1, \beta) - E(y_{i0}|\boldsymbol{z}_{1i}, \boldsymbol{\gamma}_1, \beta)$, where y_{ij} is the outcome when $x_i = j$. These are the so-called potential outcomes. From (5.8) it can be seen that the two potential outcomes are $y_{i0} = I(\boldsymbol{z}_{1i}'\boldsymbol{\gamma}_1 + \varepsilon_i)$ and $y_{i1} = I(\boldsymbol{z}_{1i}'\boldsymbol{\gamma}_1 + \beta + \varepsilon_i)$ and, therefore, under the normality assumption, the difference in expected values of the potential outcomes conditioned on $\boldsymbol{z}_{1i}$ and $\boldsymbol{\psi} = (\boldsymbol{\gamma}_1, \beta)$ is

$$E(y_{i1}|\boldsymbol{z}_{1i}, \boldsymbol{\psi}) - E(y_{i0}|\boldsymbol{z}_{1i}, \boldsymbol{\psi}) = \Phi\big(\boldsymbol{z}_{1i}'\boldsymbol{\gamma}_1 + \beta\big) - \Phi\big(\boldsymbol{z}_{1i}'\boldsymbol{\gamma}_1\big).$$

Thus, the ATE is

$$\text{ATE} = \int \big\{\Phi\big(\boldsymbol{z}_1'\boldsymbol{\gamma}_1 + \beta\big) - \Phi\big(\boldsymbol{z}_1'\boldsymbol{\gamma}_1\big)\big\}\pi(\boldsymbol{z}_1)\,\mathrm{d}\boldsymbol{z}_1, \tag{5.9}$$

where $\pi(\boldsymbol{z}_1)$ is the density of $\boldsymbol{z}_1$ (independent of $\boldsymbol{\psi}$ by assumption). Calculation of the ATE, therefore, entails an integration with respect to $\boldsymbol{z}_1$. In practice, a simple idea is to perform the marginalization using the empirical distribution of $\boldsymbol{z}_1$ from the observed sample data. If we are interested in more than the difference in expected value of the potential outcomes, then the posterior distribution of the ATE can be obtained by evaluating the integral as a Monte Carlo average, for each sampled value of $(\boldsymbol{\gamma}_1, \beta)$ from the MCMC algorithm.

6. Longitudinal binary responses

Consider now the situation in which one is interested in modeling longitudinal (univariate) binary data. Suppose that for the ith individual at time t, the probability of observing the outcome $y_{it} = 1$, conditioned on parameters $\boldsymbol{\beta}_1$ and random effects $\boldsymbol{\beta}_{2i}$, is given by

$$\Pr(y_{it} = 1|\boldsymbol{\beta}_i) = \Phi(\boldsymbol{x}_{1it}\boldsymbol{\beta}_1 + \boldsymbol{w}_{it}\boldsymbol{\beta}_{2i}),$$

where Φ is the cdf of the standard normal distribution, x_{1it} is a k_1 vector of covariates whose effect is assumed to be constant across subjects (clusters) and $\boldsymbol{w}_{it}$ is an additional and *distinct* q vector of covariates whose effects are cluster-specific. The objective is to learn about the parameters and the random effects given n subjects each with a set of measurements $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ observed at n_i points in time. The presence of the random effects ensures that the binary responses are correlated. Modeling of the data and the estimation of the model is aided once again by the latent variable framework of Albert and Chib (1993) and the hierarchical prior modeling of Lindley and Smith (1972).

For the ith cluster, we define the vector of latent variable

$$\boldsymbol{z}_i = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{W}_i\boldsymbol{\beta}_{2i} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{I}_{n_i}),$$

and let

$$y_{it} = I[z_{it} > 0],$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{in_i})'$, $\boldsymbol{X}_{1i}$ is a $n_i \times k_1$, and $\boldsymbol{W}_i$ is $n_i \times q$.

In this setting, the effects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2i}$ are estimable (without any further assumptions) under the *sequential exogeneity* assumption wherein ε_{it} is uncorrelated with $(\boldsymbol{x}_{1it}, \boldsymbol{w}_{it})$ given past values of $(\boldsymbol{x}_{1it}, \boldsymbol{w}_{it})$ and $\boldsymbol{\beta}_i$. In practice, even when the assumption of sequential exogeneity of the covariates $(\boldsymbol{x}_{1it}, \boldsymbol{w}_{it})$ holds, it is quite possible that there exist covariates $\boldsymbol{a}_i$: $r \times 1$ (with an intercept included) that are correlated with the random-coefficients $\boldsymbol{\beta}_i$. These subject-specific covariates may be measurements on the subject at baseline (time $t = 0$) or other time-invariant covariates. In the Bayesian hierarchical approach this dependence on subject-specific covariates is modeled by a hierarchical prior. One quite general way to proceed is to assume that

$$\underbrace{\begin{pmatrix} \beta_{21i} \\ \beta_{22i} \\ \vdots \\ \beta_{2qi} \end{pmatrix}}_{\boldsymbol{\beta}_{2i}} = \underbrace{\begin{pmatrix} \boldsymbol{a}_i' & \mathbf{0}' & \cdots & \cdots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{a}_i' & \cdots & \cdots & \mathbf{0}' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \cdots & \boldsymbol{a}_i' \end{pmatrix}}_{\boldsymbol{A}_i = \boldsymbol{I} \otimes \boldsymbol{a}_i'} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_{21} \\ \boldsymbol{\beta}_{22} \\ \vdots \\ \boldsymbol{\beta}_{2q} \end{pmatrix}}_{\boldsymbol{\beta}_2} + \underbrace{\begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{pmatrix}}_{\boldsymbol{b}_i}$$

or in vector–matrix form

$$\boldsymbol{\beta}_{2i} = \boldsymbol{A}_i \boldsymbol{\beta}_2 + \boldsymbol{b}_i,$$

where $\boldsymbol{A}_i$ is a $q \times k_2$ matrix, $k = r \times q$, $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{22}, \ldots, \boldsymbol{\beta}_{2q})$ is a $(k_2 \times 1)$-dimensional vector, and $\boldsymbol{b}_i$ is the mean zero random effects vector (uncorrelated with $\boldsymbol{A}_i$ and ε_i) that is distributed according to the distribution (say)

$$\boldsymbol{b}_i | \boldsymbol{D} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D}).$$

This is the second stage of the model. It may be noted that the matrix $\boldsymbol{A}_i$ can be the identity matrix of order q or the zero matrix of order q. In this hierarchical model, if $\boldsymbol{A}_i$ is not the zero matrix then identifiability requires that the matrices $\boldsymbol{X}_{1i}$ and $\boldsymbol{W}_i$ have no covariates in common. For example, if the first column of $\boldsymbol{W}_i$ is a vector of ones, then $\boldsymbol{X}_{1i}$ cannot include an intercept. If $\boldsymbol{A}_i$ is the zero matrix, however, $\boldsymbol{W}_i$ is typically a subset of $\boldsymbol{X}_{1i}$. Thus, the effect of $\boldsymbol{a}_i$ on β_{21i} (the intercept) is measured by $\boldsymbol{\beta}_{21}$, that on β_{21i} is measured by $\boldsymbol{\beta}_{22}$ and that on β_{2qi} by $\boldsymbol{\beta}_{2q}$.

Inserting the model of the cluster-specific random coefficients into the first stage yields

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{W}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i,$$

where

$$\boldsymbol{X}_i = (\boldsymbol{X}_{1i}\ \boldsymbol{W}_i \boldsymbol{A}_i) \quad \text{with } \boldsymbol{\beta} = (\boldsymbol{\beta}_1 \boldsymbol{\beta}_2),$$

as is readily checked. We can complete the model by making suitable assumptions about the distribution of $\boldsymbol{\varepsilon}_i$. One possibility is to assume that

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{I}_{n_i})$$

which leads to the probit link model. The prior distribution on the parameters can be specified as

$$\boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, B_0), \qquad \boldsymbol{D}^{-1} \sim \mathcal{W}_q(\rho_0, \boldsymbol{R}),$$

where the latter denotes a Wishart density with degrees of freedom ρ_0 and scale matrix $\boldsymbol{R}$.

The MCMC implementation in this set-up proceeds by including the $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ in the sampling (Chib and Carlin, 1999). Given $\boldsymbol{z}$ the sampling resembles the steps of the continuous response longitudinal model. To improve the efficiency of the sampling procedure, the sampling of the latent data is done marginalized over $\{\boldsymbol{b}_i\}$ from the conditional distribution of $z_{it}|\boldsymbol{z}_{i(-t)}, y_{it}, \boldsymbol{\beta}, \boldsymbol{D}$, where $\boldsymbol{z}_{i(-t)}$ is the vector $\boldsymbol{z}_i$ excluding z_{it}.

ALGORITHM 6 (*Gaussian–Gaussian panel probit*).

1 Sample

(a)

$$z_{it}|\boldsymbol{z}_{i(-t)}, y_{it}, \boldsymbol{\beta}, \boldsymbol{D} \propto \mathcal{N}(\mu_{it}, v_{it})\left\{I(z_{it} \leqslant 0)^{1-y_{it}} + I(z_{it} > 0)^{y_{it}}\right\}$$
$$\mu_{it} = \mathrm{E}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{D})$$
$$v_{it} = \mathrm{Var}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{D})$$

(b)

$$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{D} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\right), \boldsymbol{B}\right)$$

$$\boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}; \quad \boldsymbol{V}_i = \boldsymbol{I}_{n_i} + \boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'$$

(c)

$$\boldsymbol{b}_i|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{D} \sim \mathcal{N}_q\big(\boldsymbol{D}_i\boldsymbol{W}_i'(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{D}_i\big), \quad i \leqslant N$$
$$\boldsymbol{D}_i = \big(\boldsymbol{D}^{-1} + \boldsymbol{W}_i'\boldsymbol{W}_i\big)^{-1}$$

2 Sample

$$\boldsymbol{D}^{-1}|\boldsymbol{y}, \boldsymbol{\beta}, \{\boldsymbol{b}_i\} \sim \mathcal{W}_q\{\rho_0 + n, \boldsymbol{R}_n\}$$

where

$$\boldsymbol{R}_n = \left(\boldsymbol{R}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{b}_i\boldsymbol{b}_i'\right)^{-1}$$

3 Goto 1

The model with errors distributed as student-*t*

$$\begin{aligned}\boldsymbol{\varepsilon}_i|\lambda_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \lambda_i \boldsymbol{I}_{n_i}),\\ \eta_i &\sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)\end{aligned}$$

and random effects distributed as student-*t*

$$\begin{aligned}\boldsymbol{b}_i|\boldsymbol{D}, \boldsymbol{\eta}_i &\sim \mathcal{N}_q\left(\mathbf{0}, \eta_i^{-1}\boldsymbol{D}\right),\\ \eta_i &\sim G\left(\frac{\nu_b}{2}, \frac{\nu_b}{2}\right)\end{aligned}$$

can also be tackled in ways that parallel the developments in the previous section. We present the algorithm for the student–student binary response panel model without comment.

ALGORITHM 7 (*Student–Student binary panel*).

1 Sample

(a)

$$\begin{aligned}z_{it}|\boldsymbol{z}_{i(-t)}, y_{it}, \boldsymbol{\beta}, \boldsymbol{D} &\propto \mathcal{N}(\mu_{it}, v_{it})\left\{I(z_{it} \leqslant 0)^{1-y_{it}} + I(z_{it} > 0)^{y_{it}}\right\}\\ \mu_{it} &= \mathrm{E}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{D}, \lambda_i, \eta_i)\\ v_{it} &= \mathrm{Var}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{D}, \lambda_i, \eta_i)\end{aligned}$$

(b)

$$\boldsymbol{\beta}|z, \boldsymbol{D}\{\lambda_i\}, \{\eta_i\} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\right), \boldsymbol{B}_n\right)$$

where

$$\boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}; \quad \boldsymbol{V}_i = \lambda_i^{-1}\boldsymbol{I}_{n_i} + \eta_i^{-1}\boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'$$

(c)

$$\boldsymbol{b}_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{D}, \lambda_i, \eta_i \sim \mathcal{N}_q\left(\boldsymbol{D}_i\boldsymbol{W}_i'\lambda_i(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{D}_i\right)$$

where

$$\boldsymbol{D}_i = \left(\eta_i\boldsymbol{D}^{-1} + \lambda_i\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}$$

2 Sample

(a)

$$\lambda_i|\boldsymbol{y}, \boldsymbol{\beta}, \{\boldsymbol{b}_i\} \sim \mathcal{G}\left(\frac{\nu + n_i}{2}, \frac{\nu + \boldsymbol{e}_i'\boldsymbol{e}_i}{2}\right), \quad i \leqslant n$$

where $\boldsymbol{e}_i = (\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{W}_i\boldsymbol{\beta}_i)$

(b)

$$\eta_i|\boldsymbol{b}_i, \boldsymbol{D} \sim \mathcal{G}\left(\frac{\nu_b + q}{2}, \frac{\nu_b + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i}{2}\right), \quad i \leqslant n$$

3 Sample

$$\boldsymbol{D}^{-1}|z, \boldsymbol{\beta}, \{\boldsymbol{b}_i\}, \{\lambda_i\}, \{\eta_i\} \sim \mathcal{W}_q\{\rho_0 + n, \boldsymbol{R}_n\}$$

where

$$\boldsymbol{R}_n = \left(\boldsymbol{R}_0^{-1} + \sum_{i=1}^{n} \eta_i \boldsymbol{b}_i \boldsymbol{b}_i'\right)^{-1}$$

4 Goto 1

Related modeling and analysis techniques can be developed for ordinal longitudinal data. In the estimation of such models the one change is a M–H step for the sampling of the cut-points, similar to the corresponding step in Algorithm 3. It is also possible in this context to let the cut-points be subject-specific; an example is provided by Johnson (2003).

6.1. Marginal likelihood of the panel binary models

The calculation of the marginal likelihood is again straightforward. For example, in the student–student binary panel model, the posterior ordinate is expressed as

$$\pi\left(\boldsymbol{D}^{-1*}, \boldsymbol{\beta}^*|\boldsymbol{y}\right) = \pi\left(\boldsymbol{D}^{-1*}|\boldsymbol{y}\right)\pi\left(\boldsymbol{\beta}^*|\boldsymbol{y}, \boldsymbol{D}^*\right),$$

where the first term is obtained by averaging the Wishart density in Algorithm 7 over draws on $\{\boldsymbol{b}_i\}$, $\{\lambda_i\}$ and z from the full run. To estimate the second ordinate, which is conditioned on $\boldsymbol{D}^*$, we run a reduced MCMC simulation with the full conditional densities

$$\pi\left(z|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{D}^*, \{\lambda_i\}\right); \pi\left(\boldsymbol{\beta}|z, \boldsymbol{D}^*, \{\lambda_i\}, \{\eta_i\}\right);$$
$$\pi\left(\{\boldsymbol{b}_i\}|z, \boldsymbol{\beta}, \boldsymbol{D}^*, \{\lambda_i\}, \{\eta_i\}\right); \pi\left(\{\lambda_i\}, \{\eta_i\}|z, \boldsymbol{\beta}, \boldsymbol{D}^*\right),$$

where each conditional utilizes the fixed value of $\boldsymbol{D}$. The second ordinate is now estimated by averaging the Gaussian density of $\boldsymbol{\beta}$ in Algorithm 7 at $\boldsymbol{\beta}^*$ over the draws on $(z, \{\boldsymbol{b}_i\}, \{\lambda_i\}, \{\eta_i\})$ from this reduced run.

7. Longitudinal multivariate responses

In some circumstances it is necessary to model a set of multivariate categorical outcomes in a longitudinal setting. For example, one may be interested in the factors that influence purchase into a set of product categories (for example, egg, milk, cola, etc.) when a typical consumer purchases goods at the grocery store. We may have a longitudinal sample of such consumers each observed over many different shopping occasions.

On each occasion, the consumer has a multivariate outcome vector representing the (binary) incidence into each of the product categories. The available data also includes information on various category specific characteristics, for example, the average price, display and feature values of the major brands in each category. The goal is to model the category incidence vector as a function of available covariates, controlling for the heterogeneity in covariates effects across subjects in the panel. Another example of multidimensional longitudinal data appears in Dunson (2003) motivated by studies of an item responses battery that is used to measure traits of an individual repeatedly over time. The model discussed in Dunson (2003) allows for mixtures of count, categorical, and continuous response variables.

To illustrate a canonical situation of multivariate longitudinal categorical responses, we combine the MVP model with the longitudinal models in the previous section. Let $\boldsymbol{y}_{it}$ represent a J vector of binary responses and suppose that we have for each subject i at each time t, a covariate matrix X_{1it} in the form

$$\boldsymbol{X}_{1it} = \begin{pmatrix} \boldsymbol{x}'_{11it} & \boldsymbol{0}' & \cdots & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{x}'_{12it} & \boldsymbol{0}' & \boldsymbol{0}' \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0}' & \boldsymbol{0}' & \cdots & \boldsymbol{x}'_{1Jit} \end{pmatrix},$$

where each $\boldsymbol{x}_{1jit}$ has k_{1j} elements so that the dimension of $\boldsymbol{X}_{1it}$ is $J \times k_1$, where $k_1 = \sum_{j=1}^{J} k_{1j}$. Now suppose that there is an additional (distinct) set of covariates $\boldsymbol{W}_{it}$ whose effect is assumed to be both response and subject specific and arranged in the form

$$\boldsymbol{W}_{it} = \begin{pmatrix} \boldsymbol{w}'_{1it} & \boldsymbol{0}' & \cdots & \boldsymbol{0}' \\ \boldsymbol{0}' & \boldsymbol{w}'_{2it} & \boldsymbol{0}' & \boldsymbol{0}' \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0}' & \boldsymbol{0}' & \cdots & \boldsymbol{w}'_{Jit} \end{pmatrix},$$

where each w_{jit} has q_j elements so that the dimension of $\boldsymbol{W}_{it}$ is $J \times q$, where $q = \sum_{j=1}^{J} q_j$. Also suppose that in terms of the latent variables $\boldsymbol{z}_{it}$: $J \times 1$, the model generating the data is given by

$$\boldsymbol{z}_{it} = \boldsymbol{X}_{1it}\boldsymbol{\beta}_1 + \boldsymbol{W}_{it}\boldsymbol{\beta}_{2i} + \boldsymbol{\varepsilon}_{it}, \quad \boldsymbol{\varepsilon}_{it} \sim \mathcal{N}_J(\boldsymbol{0}, \Sigma),$$

such that

$$y_{jit} = I[z_{jit} > 0], \quad j \leqslant J,$$

and where

$$\boldsymbol{\beta}_1 = \begin{pmatrix} \boldsymbol{\beta}_{11} \\ \boldsymbol{\beta}_{12} \\ \vdots \\ \boldsymbol{\beta}_{1J} \end{pmatrix} : k_1 \times 1 \quad \text{and} \quad \boldsymbol{\beta}_{2i} = \begin{pmatrix} \boldsymbol{\beta}_{21i} \\ \boldsymbol{\beta}_{22i} \\ \vdots \\ \boldsymbol{\beta}_{2Ji} \end{pmatrix} : q \times 1.$$

To model the heterogeneity in effects across subjects we can assume that

$$\underbrace{\begin{pmatrix}\boldsymbol{\beta}_{21i}\\ \boldsymbol{\beta}_{22i}\\ \vdots\\ \boldsymbol{\beta}_{2Ji}\end{pmatrix}}_{\boldsymbol{\beta}_{2i}:\, q\times 1} = \underbrace{\begin{pmatrix}\boldsymbol{A}_{1i} & \mathbf{0} & \cdots & \cdots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{A}_{2i} & \cdots & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \boldsymbol{A}_{Ji}\end{pmatrix}}_{\boldsymbol{A}_i:\, q\times k_2} \underbrace{\begin{pmatrix}\boldsymbol{\beta}_{21}\\ \boldsymbol{\beta}_{22}\\ \vdots\\ \vdots\\ \boldsymbol{\beta}_{2J}\end{pmatrix}}_{\boldsymbol{\beta}_2:\, k_2\times 1} + \underbrace{\begin{pmatrix}\boldsymbol{b}_{i1}\\ \boldsymbol{b}_{i2}\\ \vdots\\ \boldsymbol{b}_{iJ}\end{pmatrix}}_{\boldsymbol{b}_i:\, q\times 1},$$

where each matrix $\boldsymbol{A}_{ji}$ is of the form $\boldsymbol{I}_{q_j} \otimes \boldsymbol{a}'_{ji}$ for some r_j category-specific covariates $\boldsymbol{a}_{ji}$. The size of the vector $\boldsymbol{\beta}_2$ is then $k_2 = \sum_{j=1}^{J} q_j \times r_j$. Substituting this model of the random-coefficients into the first stage yields

$$z_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \boldsymbol{W}_{it}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{it}, \quad \boldsymbol{\varepsilon}_{it} \sim \mathcal{N}_J(\mathbf{0}, \Sigma),$$

where

$$\boldsymbol{X}_{it} = (\boldsymbol{X}_{1it}, \boldsymbol{W}_{it}\boldsymbol{A}_i)$$

and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$: $k \times 1$. Further assembling all n_i observations on the ith subject for a total of $n_i^* = J \times n_i$ observations, we have

$$\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i^*}(\mathbf{0}, \boldsymbol{\Omega}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}),$$

where $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i})'$: $n_i^* \times 1$, $\boldsymbol{X}_i = (\boldsymbol{X}'_{i1}, \ldots, \boldsymbol{X}'_{in_i})'$: $n_i^* \times k$, and $\boldsymbol{W}_i = (\boldsymbol{W}'_{i1}, \ldots, \boldsymbol{W}'_{in_i})'$: $n_i^* \times q$. Apart from the increase in the dimension of the problem, the model is now much like the longitudinal model with univariate outcomes. The fitting, consequently, parallels that in Algorithms 6 and 7.

ALGORITHM 8 (*Longitudinal MVP*).

1 Sample

(a)

$$z_{it}|\boldsymbol{z}_{i(-t)}, y_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{D} \propto \mathcal{N}(\mu_{it}, v_{it})\left\{I(z_{it} \leqslant 0)^{1-y_{it}} + I(z_{it} > 0)^{y_{it}}\right\}$$
$$\mu_{it} = \mathrm{E}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{D})$$
$$v_{it} = \mathrm{Var}(z_{it}|\boldsymbol{z}_{i(-t)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{D})$$

(b)

$$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\right), \boldsymbol{B}_n\right)$$

where

$$\boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}; \quad \boldsymbol{V}_i = \boldsymbol{\Omega}_i + \boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'$$

(c)

$$\boldsymbol{b}_i|z_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathcal{N}_q\big(\boldsymbol{D}_i \boldsymbol{W}_i' \boldsymbol{\Omega}_i^{-1}(z_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{D}_i\big)$$

where

$$\boldsymbol{D}_i = \big(\boldsymbol{D}^{-1} + \boldsymbol{W}_i' \boldsymbol{\Omega}_i^{-1} \boldsymbol{W}_i\big)^{-1}$$

2 Sample

$$\boldsymbol{D}^{-1}|z, \boldsymbol{\beta}, \{\boldsymbol{b}_i\} \sim \mathcal{W}_q\{\rho_0 + n, \boldsymbol{R}_n\}$$

where

$$\boldsymbol{R}_n = \left(\boldsymbol{R}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{b}_i \boldsymbol{b}_i'\right)^{-1}$$

3 Goto 1

8. Conclusion

This chapter has summarized Bayesian modeling and fitting techniques for categorical responses. The discussion has dealt with cross-section models for binary and ordinal data and presented extensions of those models to multivariate and longitudinal responses. In each case, the Bayesian MCMC fitting technique is simple and easy to implement.

The discussion in this chapter did not explore residual diagnostics and model fit issues. Relevant ideas are contained in Albert and Chib (1995, 1997) and Chen and Dey (2000). We also did not take up the multinomial probit model as, for example, considered in McCulloch et al. (2001), or the class of semiparametric binary response models in which the covariate effects are modeled by regression splines or other related means and discussed by, for example, Wood et al. (2002), Holmes and Mallick (2003) and Chib and Greenberg (2004). Fitting of all these models relies on the framework of Albert and Chib (1993).

In the Bayesian context it is relatively straightforward to compare alternative models. In this chapter we restricted our attention to the approach based on marginal likelihoods and Bayes factors and showed how the marginal likelihood can be computed for the various models from the output of the MCMC simulations. Taken together, the models and methods presented in this chapter are a testament to the versatility of Bayesian ideas for dealing with categorical response data.

References

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Albert, J., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

Albert, J., Chib, S. (1995). Bayesian residual analysis for binary response models. *Biometrika* **82**, 747–759.
Albert, J., Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *J. Amer. Statist. Assoc.* **92**, 916–925.
Albert, J., Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57**, 829–836.
Allenby, G.M., Lenk, P.J. (1994). Modeling household purchase behavior with logistic normal regression. *J. Amer. Statist. Assoc.* **89**, 1218–1231.
Amemiya, T. (1972). Bivariate probit analysis: Minimum chi-square methods. *J. Amer. Statist. Assoc*. **69**, 940–944.
Ashford, J.R., Sowden, R.R. (1970). Multivariate probit analysis. *Biometrics* **26**, 535–546.
Banerjee, S., Carlin, B., Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.
Basu, S., Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.* **98**, 224–235.
Basu, S., Mukhopadhyay, S. (2000). Binary response regression with normal scale mixture links. In: Dey, D.K., Ghosh, S.K., Mallick, B.K. (Eds.), *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker, New York, pp. 231–242.
Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B* **36**, 192–236.
Carey, V., Zeger, S.L., Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.
Carlin, B.P., Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B* **57**, 473–484.
Chen, M.-H., Dey, D. (2000). A unified Bayesian analysis for correlated ordinal data models. *Brazilian J. Probab. Statist.* **14**, 87–111.
Chen, M.-H., Shao, Q.-M. (1998). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25**, 1563–1594.
Chen, M.-H., Dey, D., Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *J. Amer. Statist. Assoc.* **94**, 1172–1186.
Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc*. **90**, 1313–1321.
Chib, S. (2003). On inferring effects of binary treatments with unobserved confounders (with discussion). In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), *Bayesian Statistics, vol. 7*. Oxford University Press, London.
Chib, S., Carlin, B.P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statist. Comput.* **9**, 17–26.
Chib, S., Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *Amer. Statist*. **49**, 327–335.
Chib, S., Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.
Chib, S., Greenberg, E. (2004). Analysis of additive instrumental variable models. Technical Report, Washington University in Saint Louis.
Chib, S., Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *J. Amer. Statist. Assoc.* **96**, 270–281.
Connolly, M.A., Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501–506.
Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc., Ser. B* **34**, 187–220.
DiCiccio, T.J., Kass, R.E., Raftery, A.E., Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903–915.
Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *J. Amer. Statist. Assoc.* **98**, 555–563.
Farhmeir, L., Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.
Fahrmeir, L., Lang, S. (2001). Bayesian semiparametric regression analysis of multicategorical time-space data. *Ann. Inst. Statist. Math.* **53**, 11–30.
Gelfand, A.E., Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **12**, 609–628.

Glonek, G.F.V., McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc., Ser. B* **57**, 533–546.

Green, P.E. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

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

Holmes, C.C., Mallick, B.K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *J. Amer. Statist. Assoc.* **98**, 352–368.

Johnson, T.R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrica* **68**, 563–583.

Kalbfleish, J., Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kuhnert, P.M., Do, K.A. (2003). Fitting genetic models to twin data with binary and ordered categorical responses: A comparison of structural equation modelling and Bayesian hierarchical models. *Behavior Genetics* **33**, 441–454.

Lesaffre, E., Kaufmann, H.K. (1992). Existence and uniqueness of the maximum likelihood estimator for a multivariate probit model. *J. Amer. Statist. Assoc.* **87**, 805–811.

Lindley, D.V., Smith, A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc., Ser. B* **34**, 1–41.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc., Ser. B* **42**, 109–127.

McCulloch, R.E., Polson, N.G., Rossi, P.E. (2001). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99**, 173–193.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

O'Brien, S.M., Dunson, D.B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60**, 739–746.

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

Ripley, B. (1987). *Stochastic Simulation*. Wiley, New York.

Roberts, C.P. (2001). *The Bayesian Choice*. Springer-Verlag, New York.

Smith, A.F.M., Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B* **55**, 3–24.

Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–549.

Ten Have, T.R., Uttal, D.H. (2000). Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Appl. Statist.* **43**, 371–384.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Tutz, G. (1990). Sequential item response models with an ordered response. *British J. Math. Statist. Psychol.* **43**, 39–55.

Wood, S., Kohn, R., Shively, T., Jiang, W.X. (2002). Model selection in spline nonparametric regression. *J. Roy. Statist. Soc., Ser. B* **64**, 119–139.