Panel Data Modeling and Inference: A Bayesian Primer*

Siddhartha Chib
John M. Olin School of Business
Washington University in St. Louis
Campus Box 1133, 1 Brookings Dr.
St. Louis, MO 63130, USA

May 2004

1 Introduction

In this chapter we discuss how Bayesian methods are used to model and analyze panel data. As in other areas of econometrics and statistics, the growth of Bayesian ideas in the panel data setting has been aided by the revolutionary developments in *Markov chain Monte Carlo* (MCMC) methods. These methods, applied creatively, allow for the sophisticated modeling of continuous, binary, censored, count and multinomial responses under weak assumptions. The purpose of this largely self-contained chapter is to summarize the various modeling possibilities and to provide the associated inferential techniques for conducting the prior-posterior analyses.

The apparatus we outline in this chapter relies on some powerful and easily implementable Bayesian precepts (for a textbook discussion of Bayesian methods, see Congdon (2001)). One theme around which much of the discussion is organized is *hierarchical prior modeling* (*Lindley and Smith* [1972]) which allows the researcher to model cluster-specifc heterogeneity (and its dependence on cluster-specific covariates) through random effects and random-coefficients in various interesting ways. Another theme is the use of the general approaches of *Albert and Chib* [1993] and *Chib* [1992] for dealing with binary, ordinal and censored outcomes. A third theme is the

*To appear in the *Econometrics of Panel Data* (3rd edition), edited by Lazlo Matyas and Patrick Sevestre, Kluwer Press.

use of flexible and robust families of parametric distributions to represent sampling densities and prior distributions. A fourth theme is the comparison of alternative clustered data models via marginal likelihoods and Bayes factors, calculated via the method of *Chib* [1995]. A final theme is the use of MCMC methods (*Gelfand and Smith* [1990], *Tierney* [1994], *Chib and Greenberg* [1995]) to sample the posterior distribution, to calculate the predictive density and the posterior distribution of the residuals, and to estimate the marginal likelihood.

Because implementation of the Bayesian paradigm is inextricably tied to MCMC methods, we include a brief overview of MCMC methods and of certain basic results that prove useful in the derivation of the conditional densities that form the basis for model fitting by MCMC simulation methods. Methods for producing random variates from a few common distributions are also included. After these preliminaries, the chapter turns to the analysis of panel data models for continuous outcomes followed by a discussion of models and methods for binary, censored, count and multinomial outcomes. The last half of the chapter deals with the problems of an endogenous covariate, informative missingness, prediction, residual analysis and model comparison.

1.1 Hierarchical Prior Modeling

The Bayesian approach to panel data modeling relies extensively on the idea of a hierarchical prior which is used to model the heterogeneity in subject-specific coefficients and the distribution of the errors and the random effects. Suppose that for the ith cluster (subject) in the sample we are interested in modeling the distribution of $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ on a continuous response y. Also suppose that $\mathbf{W}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{in_i})'$ is a $n_i \times q$ matrix of observations on q covariates $\mathbf{w}_{it}$ whose effect on y is assumed to be cluster-specific. In particular, suppose that for the ith subject at the tth time point one writes

$$y_{it} = \mathbf{w}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}\ , i = 1, 2, ..., N;\ t = 1, 2, ..., n_i \tag{1}$$

or equivalently for all observations in the ith cluster

$$\begin{aligned} &\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, i = 1, 2, ..., N \\ &\boldsymbol{\varepsilon}_i \sim P \end{aligned}$$

where $\boldsymbol{\beta}_i$ is the cluster-specific coefficient vector and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, ..., \varepsilon_{in_i})$ is the error distributed *marginally* with mean zero according to the distribution P (to be modeled below).

In the context of observational data, one is concerned about the presence of unobserved confounders (variables that simultaneously affect the covariates $\mathbf{w}_{it}$ and the error ε_{it}). Under such endogeneity of the covariates, $E(\boldsymbol{\varepsilon}_i|\mathbf{W}_i, \boldsymbol{\beta}_i)$ is not zero and the cluster-specific effects are not identified without additional assumptions and the availability of instruments. To make progress, and to avoid the latter situation, it is common to assume that the covariates $\mathbf{w}_{it}$ are strictly exogenous in the sense that ε_i is uncorrelated with $\mathbf{W}_i$ and $\boldsymbol{\beta}_i$, which implies that ε_{it} is uncorrelated with past, current and future values of $\mathbf{w}_{it}$, given $\boldsymbol{\beta}_i$, or in other words, that the distribution of $\boldsymbol{\varepsilon}_i$ given $(\mathbf{W}_i, \boldsymbol{\beta}_i)$ is P. In the Bayesian context, this strict exogeneity assumption is not required and analysis can proceed under the weaker *sequential exogeneity* assumption wherein ε_{it} is uncorrelated with $\mathbf{w}_{it}$ given past values of $\mathbf{w}_{it}$ and β_i. Most of our analysis, in fact, is conducted under this assumption, although we do not make it explicit in the notation. There are situations, of course, where even the assumption of sequential exogeneity is not tenable. We consider one such important case below where a time-varying binary covariate (a non-randomly assigned "treatment") is correlated with the error. We show how the Bayesian analysis is conducted when an instrument is available to model the marginal distribution of the treatment.

In practice, even when the assumption of sequential exogeneity of the covariates $\mathbf{w}_{it}$ holds, it is quite possible that there exist covariates $\mathbf{a}_i : r \times 1$ (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

$$\underbrace{\begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iq} \end{pmatrix}}_{\boldsymbol{\beta}_i} = \underbrace{\begin{pmatrix} \mathbf{a}_i' & \mathbf{0}' & \cdots & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{a}_i' & \cdots & \cdots & \mathbf{0}' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \cdots & \mathbf{a}_i' \end{pmatrix}}_{\mathbf{A}_i} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_{11} \\ \boldsymbol{\beta}_{22} \\ \vdots \\ \vdots \\ \boldsymbol{\beta}_{qq} \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{pmatrix}}_{\mathbf{b}_i}$$

or in vector-matrix form

$$\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{b}_i$$

where $\mathbf{A}_i$ is a $q \times k$ matrix given as $\mathbf{I}_q \otimes \mathbf{a}_i'$, $k = r \times q$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{22}, ..., \boldsymbol{\beta}_{qq})$ is a $k \times 1$ dimensional vector, and $\mathbf{b}_i$ is the mean zero random effects vector

(uncorrelated with $\mathbf{A}_i$ and ε_i) that is distributed according to the distribution Q. This is the second-stage of the model. It may be noted that the matrix $\mathbf{A}_i$ can be the identity matrix of order q or the zero matrix of order q. Thus, the effect of $\mathbf{a}_i$ on β_{i1} (the intercept) is measured by $\boldsymbol{\beta}_{11}$, that on β_{i2} is measured by $\boldsymbol{\beta}_{22}$ and that on β_{iq} by $\boldsymbol{\beta}_{qq}$.

In the same way, the hierarchical approach can be used to model the distributions P and Q. One way is to assume that each of these distributions belong to the (hierarchical) scale mixture of normals family. Formally, to model the distribution of $\boldsymbol{\varepsilon}_i$, we could, for example, let

$$\begin{aligned}\boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i, \boldsymbol{\Omega}_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)\\ \lambda_i &\sim G\end{aligned}$$

where $\boldsymbol{\Omega}_i$ is a positive-definite matrix depending perhaps on a set of unknown parameters ϕ, σ^2 is an unknown positive scale parameter, and λ_i is the random scale parameter that is drawn independently across clusters from some distribution G (say with known parameters). If for example, we assume that

$$G = \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right)$$

where $\mathcal{G}$ denotes the gamma distribution, then the distribution of $\boldsymbol{\varepsilon}_i$ marginalized over λ_i is multivariate-t with density proportional to

$$|\Omega|^{1/2}\left(1 + \frac{1}{\nu\sigma^2}\boldsymbol{\varepsilon}'\Omega_i^{-1}\boldsymbol{\varepsilon}\right)^{-(\nu+n_i)/2}.$$

Similarly, to model the random effects vector $\mathbf{b}_i$ we could let

$$\begin{aligned}\mathbf{b}_i|\eta_i, \mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0}, \eta_i^{-1}\mathbf{D}\right)\\ \eta_i &\sim F\end{aligned}$$

where $\mathbf{D}$ is a full matrix and η_i is a positive random variable drawn independently across clusters from a distribution F.

The Bayesian hierarchical model is completed through the specification of prior densities on all the non-cluster-specific coefficients. In general terms, we let

$$(\boldsymbol{\beta}, \mathbf{D}, \sigma^2, \boldsymbol{\phi}) \sim \pi$$

where π is some suitable parametric distribution. Interestingly, it is possible to model the prior distribution in stages by putting a prior on the parameters (hyperparameters) of π. Note that the latter distribution is a

prior distribution on parameters from the different stages of the hierarchical model.

As another example of a hierarchical model, suppose that $\mathbf{X}_{1i}$ is an additional $n_i \times k_1$ matrix of observations on k_1 covariates whose effect on y is assumed to be non-cluster-specific. Now suppose that the model generating $\mathbf{y}_i$ is taken to

$$\mathbf{y}_i = \mathbf{X}_{1i}\boldsymbol{\beta}_1 + \mathbf{W}_i\boldsymbol{\beta}_{2i} + \boldsymbol{\varepsilon}_i, i = 1, 2, ..., N \tag{2}$$

where, as above, the distribution of the subject-specific $\boldsymbol{\beta}_{2i}$ is modeled as

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

with the remaining components of the model unchanged. In this hierarchical model, if $\mathbf{A}_i$ is not the zero matrix then identifiability requires that the matrices $\mathbf{X}_{1i}$ and $\mathbf{W}_i$ have no covariates in common. For example, if the first column of $\mathbf{W}_i$ is a vector of ones, then $\mathbf{X}_{1i}$ cannot include an intercept. If $\mathbf{A}_i$ is the zero matrix, however, $\mathbf{W}_i$ is typically a subset of $\mathbf{X}_{1i}$.

These two types of hierarchical Bayesian models play a large role in the Bayesian analysis of clustered data. Notice that both models share the same form. This is seen by inserting the model of the cluster-specific random coefficients into the first stage which yields

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i, \boldsymbol{\Omega}_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \\ \mathbf{b}_i|\eta_i, \mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0}, \eta_i^{-1}\mathbf{D}\right) \\ \lambda_i \sim G \ , \ \eta_i &\sim F \\ (\boldsymbol{\beta}, \mathbf{D}, \sigma^2) &\sim \pi \end{aligned}$$

where in the first type of hierarchical model

$$\mathbf{X}_i = \mathbf{W}_i\mathbf{A}_i$$

and in the second type of hierarchical model

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

as is readily checked. The latter model is therefore the canonical Bayesian hierarchical model for continuous clustered data.

1.2 Elements of Markov chain Monte Carlo

The basic idea behind MCMC methods is quite simple. Suppose that $\pi(\boldsymbol{\psi}|\mathbf{y}) \propto \pi(\boldsymbol{\psi})p(\mathbf{y}|\boldsymbol{\psi})$ is the posterior density for a set of parameters $\psi \in \Re^d$ in a particular Bayesian model defined by the prior density $\pi(\boldsymbol{\psi})$ and sampling density or likelihood function $p(\mathbf{y}|\boldsymbol{\psi})$ and that interest centers on the posterior mean $\eta = \int_{\Re^d} \psi\pi(\psi|\mathbf{y})d\psi$. Now suppose that this integral cannot be computed analytically and that the dimension of the integration exceeds three or four (which essentially rules out the use of standard quadrature-based methods). In such cases one calculates the integral by Monte Carlo sampling methods. The general idea is to abandon the immediate task at hand (which is the computation of the above integral) and to ask how the posterior density $\pi(\boldsymbol{\psi}|y)$ may be sampled. The reason for changing our focus is that if we were to have the draws

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

from the posterior density, then provided the sample is large enough, we estimate not just the above integral but also other features of the posterior density by taking those draws and forming the relevant sample-based estimates. For example, the sample average of the sampled draws is our simulation-based estimate of the posterior mean, while the quantiles of the sampled output are estimates of the posterior quantiles, with other summaries obtained in a similar manner. Under suitable laws of large numbers these estimates converge to the posterior quantities as the simulation-size becomes large. In short, the problem of computing an intractable integral is reduced to the problem of sampling the posterior density.

The sampling of the posterior distribution is, therefore, the central focus of 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 val-

ues. 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, Rosenbluth, Rosenbluth, Teller and Teller* [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 *Chib and Greenberg* [1995] for a detailed summary). 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 for some grouping of the parameters into sub-blocks, say ψ_1 and ψ_2 (the extension to more than two blocks is straightforward), the set of full conditional densities

$$\pi_1(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2) \propto p(\mathbf{y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) \tag{3}$$

$$\pi_2(\boldsymbol{\psi}_2|\mathbf{y}, \boldsymbol{\psi}_1) \propto p(\mathbf{y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) \tag{4}$$

are tractable (that is, of known form and readily sampled). 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 $j = 1, 2, ..., n_0 + G$.

   Generate $\boldsymbol{\psi}_1^{(j)}$ from $\pi_1(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)})$
   Generate $\boldsymbol{\psi}_2^{(j)}$ from $\pi_2(\boldsymbol{\psi}_2|\mathbf{y}, \boldsymbol{\psi}_1^{(j)})$

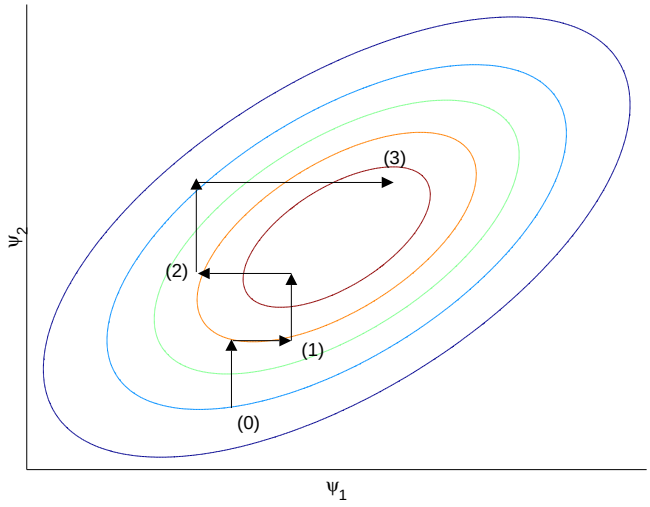

Figure 1: Gibbs algorithm with two blocks: Three complete iterations of the algorithm.

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

To illustrate the manner in which the blocks are revised, consider Figure 1.1 which traces out a possible trajectory of the sampling algorithm under the assumption that each block consists of a single component. The contours in the plot represent the joint distribution of $\boldsymbol{\psi}$ and the labels "(0)", "(1)" etc., denote the simulated values. Note that one iteration of the algorithm is completed after both components are revised. Also notice that each component is revised along the direction of the coordinate axes. This feature is a source of problems if the two components are highly correlated because then the contours become compressed and movements along the coordinate axes tend to produce only small moves.

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 $\pi(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2)$ is intractable. Let

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

denote a suitably chosen proposal density of making a transition from $\boldsymbol{\psi}_1$ to $\boldsymbol{\psi}_1'$, given the data and the values of the remaining blocks (see for example

Chib and Greenberg [1995]). Then, in the first step of the jth iteration of the MCMC algorithm, given the values $\boldsymbol{\psi}_2^{(j-1)}$ of the remaining block, the updated iterate of $\boldsymbol{\psi}_1$ is drawn as follows.

Algorithm: Metropolis-Hastings for sampling an intractable $\pi_1(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2)$

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

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

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

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

3. Set

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

A similar approach is used to sample $\boldsymbol{\psi}_2$ if the full conditional density of $\boldsymbol{\psi}_2$ is intractable. These algorithms are extended to more than two blocks in a straightforward manner (*Chib* [2001]).

1.3 Some Basic Bayesian Updates

We now summarize four results that appear in the development of the MCMC algorithms for the various models that are discussed below. These results provide, for the stated models, the posterior distribution of a set of parameters, conditional on the other parameters of the model. The results are stated in some generality and are specialized, as needed, in the subsequent discussion.

Result 1: Suppose that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\ ,\ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)\ ,\ i \le N$$
$$\mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D})$$
$$\boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \mathbf{B}_0)$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ is a vector of n_i observations on the dependent variable for subject i. Then marginalized over $\{\mathbf{b}_i\}$

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \{\lambda_i\}, \{\eta_i\}, \boldsymbol{\Omega}_i, \mathbf{D} \sim \mathcal{N}_k \left\{\hat{\boldsymbol{\beta}}, \mathbf{B}\right\} \tag{5}$$

where

$$\hat{\boldsymbol{\beta}} = \mathbf{B}\left(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{N} \mathbf{X}_i \mathbf{V}_i^{-1} \mathbf{y}_i\right) , \tag{6}$$

$$\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \tag{7}$$

and

$$\mathbf{V}_i = \sigma^2 \lambda_i^{-1} \boldsymbol{\Omega}_i + \eta_i^{-1} \mathbf{W}_i \mathbf{D} \mathbf{W}_i' \tag{8}$$

Result 2: Suppose that

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \\ \mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D}) \end{aligned}$$

Then

$$\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \sigma^2, \lambda_i, \eta_i, \mathbf{D} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right) \tag{9}$$

where

$$\hat{\mathbf{b}}_i = \sigma^{-2}\lambda_i \mathbf{D}_i \mathbf{W}_i' \boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{10}$$

and

$$\mathbf{D}_i = (\eta_i \mathbf{D}^{-1} + \sigma^{-2}\lambda_i \mathbf{W}_i' \boldsymbol{\Omega}_i^{-1} \mathbf{W}_i)^{-1}. \tag{11}$$

Result 3: Suppose that

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \\ \mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i\mathbf{D}) \, , \ i \leq N \\ \mathbf{D}^{-1} \sim \mathcal{W}_q(\rho_0, \mathbf{R}_0) \end{aligned}$$

where $\mathcal{W}_T(\rho, \mathbf{R})$ is the Wishart distribution with density

$$c\frac{|\mathbf{W}|^{(\nu-T-1)/2}}{|\mathbf{R}|^{\nu/2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{R}^{-1}\mathbf{W}\right)\right\}, \quad |\mathbf{W}| > 0,$$

$$c = \left(2^{\rho T/2} \pi^{T(T-1)/4} \prod_{i=1}^{T} \Gamma\left(\frac{\rho + 1 - i}{2} \right) \right)^{-1}$$

is the normalizing constant and $\mathbf{R}$ is a hyperparameter matrix (*Roberts* [2001]). Then

$$\mathbf{D}^{-1} | \{\mathbf{b}_i\}, \mathbf{y}, \mathbf{\Omega}_i, \{\lambda_i\}, \{\eta_i\} = \mathbf{D}^{-1} | \{\mathbf{b}_i\}, \eta_i \sim \mathcal{W}_q \left(\rho_0 + N, \mathbf{R} \right) \tag{12}$$

where

$$\mathbf{R} = \left(\mathbf{R}_0^{-1} + \sum_{i=1}^{N} \eta_i \mathbf{b}_i \mathbf{b}_i' \right)^{-1} . \tag{13}$$

Result 4: Suppose that

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{W}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i | \sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \lambda_i^{-1} \mathbf{\Omega}_i) \, , \ i \leq N$$
$$\sigma^2 \sim \mathcal{IG}\left(\frac{v_0}{2}, \frac{\delta_0}{2} \right)$$

where $\mathcal{IG}(a, b)$ is the inverse-gamma distribution with density $\pi(\sigma^2 | a, b) \propto \left(\sigma^2\right)^{-a+1} \exp\left(-b/\sigma^2\right)$. Then

$$\sigma^2 | \mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \{\lambda_i\} \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \delta}{2} \right) \tag{14}$$

where

$$\delta = \sum_{i=1}^{N} \lambda_i \mathbf{e}_i' \mathbf{\Omega}_i^{-1} \mathbf{e}_i \tag{15}$$

and

$$\mathbf{e}_i = (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i \mathbf{b}_i)$$

1.4 Basic Variate Generators

In the application of MCMC methods it often occurs that the simulation of the given target distribution is reduced to a sequence of simulations from standard and familiar univariate and multivariate distributions. With that in mind, we present simulation routines for the distributions that are encountered in the sequel.

Gamma Variate: To obtain ψ from $\mathcal{G}(\alpha, \beta)$ with density proportional to $\psi^{\alpha-1} \exp(-\beta\psi)$, we draw θ from $\mathcal{G}(\alpha, 1)$ and set $\psi = \theta/\beta$. A draw of

a chi-squared variate χ^2_v with ν degrees of freedom is obtained by drawing from a $\mathcal{G}(\alpha/2, 1/2)$ distribution.

Inverse-Gamma Variate: A random variable that follows the inverse gamma distribution $\mathcal{IG}(\alpha, \beta)$ is equal in distribution to the inverse of random variable that follows the $\mathcal{G}(\alpha, \beta)$ distribution. Therefore, an inverse-gamma variate is obtained by drawing θ from $\mathcal{G}(\alpha, \beta)$ and setting $\psi = 1/\theta$.

Truncated normal Variate: A variate from

$$\psi \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2),$$

a univariate normal distribution truncated to the interval (a, b), is obtained by the inverse-cdf method. The distribution function of the truncated normal random variable is

$$F(t) = \begin{cases} 0 & \text{if} \quad \psi < a \\ \frac{1}{p_2 - p_1}\left(\Phi(\frac{t-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right) & \text{if} \quad a < \psi < b \\ 1 & \text{if} \quad b < \psi \end{cases} \tag{16}$$

where

$$p_1 = \Phi\left(\frac{a-\mu}{\sigma}\right) \; ; \; p_2 = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Therefore, if U is uniform on $(0, 1)$, then

$$\psi = \mu + \sigma\Phi^{-1}\left(p_1 + U(p_2 - p_1)\right) \tag{17}$$

is the required draw. Here Φ^{-1} is the inverse cdf of the standard normal distribution and can be computed by the method of *Page* [1977].

Multivariate Normal vector: To obtain a random vector $\boldsymbol{\psi}$ from $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Omega})$, we draw $\boldsymbol{\theta}$ from $\mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$ and set $\boldsymbol{\psi} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\theta}$ where $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}'$.

Wishart matrix: To obtain a random positive-definite matrix W from $\mathcal{W}_T(v, \mathbf{R})$, one first generates the random lower triangular matrix $\mathbf{T} = (t_{ij})$, such that

$$t_{ii} \sim \sqrt{\chi^2_{v-i+1}} \text{ and } t_{ij} \sim \mathcal{N}(0, 1)$$

Then the quantity

$$\mathbf{W} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}'$$

where $\mathbf{R} = \mathbf{L}\mathbf{L}'$ is the required draw.

2 Continuous Responses

As discussed in Section 1.1, Bayesian hierarchical modeling of subject-specific coefficients leads to the canonical model for unbalanced continuous outcomes

$$\begin{aligned}\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i, \boldsymbol{\Omega}_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \\ \mathbf{b}_i|\eta_i, \mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0}, \eta_i^{-1}\mathbf{D}\right) \\ \lambda_i \sim G \ , \ \eta_i &\sim F \\ (\boldsymbol{\beta}, \mathbf{D}, \sigma^2) &\sim \pi\end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ is the data on the ith individual over the n_i time periods, $\mathbf{W}_i$ is a set of variables whose effect $\mathbf{b}_i$ is assumed to heterogenous, $\mathbf{X}_i$ is a set of raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i} \ \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$.

There are many ways to proceed from this point. If G and F are degenerate at one, we get the Gaussian-Gaussian model. If we assume that

$$G = \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right)$$

and

$$F = \mathcal{G}\left(\frac{\nu_F}{2}, \frac{\nu_F}{2}\right)$$

then the distributions of $\boldsymbol{\varepsilon}_i$ and $\mathbf{b}_i$ marginalized over λ_i and η_i are multivariate student-t with ν_G and ν_F degrees of freedom, respectively. This model may be called the Student-Student model. Other models are obtained by making specific assumptions about the form of $\boldsymbol{\Omega}_i$. For example, if ε_i is assumed to to be serially correlated according to say an ARMA process, then $\boldsymbol{\Omega}_i$ is the covariance matrix of the assumed ARMA process. The distribution π is typically specified in the same way, regardless of the distributions adopted in other stages of the model. Specifically, it is common to assume that the parameters $(\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$ are apriori mutually independent with

$$\boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, B_0) \, ; \sigma^2 \sim \mathcal{IG}\left(\frac{v_0}{2}, \frac{\delta_0}{2}\right) \, ; \mathbf{D}^{-1} \sim \mathcal{W}_p(\rho_0, \mathbf{R}_0)$$

2.1 Gaussian-Gaussian model

To see how the analysis may proceed, consider the model in which the distributions of the error and the random-effects are both Gaussian. In particular,

suppose that

$$\begin{aligned}\boldsymbol{\varepsilon}_i|\sigma^2 &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\boldsymbol{\Omega}_i),\\ \mathbf{b}_i|\mathbf{D} &\sim \mathcal{N}_q(\mathbf{0}, \mathbf{D})\ , i \leq N\end{aligned}$$

where the matrix $\boldsymbol{\Omega}_i$ is assumed to be known. Under these assumptions the joint posterior of all the unknowns, including the random effects $\{\mathbf{b}_i\}$, is given by

$$\pi(\boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}^{-1}, \sigma^2|\mathbf{y}) = \pi(\boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}^{-1}, \sigma^2) \prod_{i=1}^{N} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \sigma^2) p(\mathbf{b}_i|\mathbf{D}). \tag{18}$$

Wakefield et.al [1994] propose a Gibbs MCMC approach for sampling the joint posterior distribution based on full blocking (ie., sampling each block of parameters from their full conditional distribution). This blocking scheme is not very desirable because the random effects and the fixed effects $\boldsymbol{\beta}$ tend to be highly correlated and treating them as separate blocks creates problems with mixing (*Gelfand, Sahu and Carlin* [1995]).

To deal with this problem, *Chib and Carlin* [1999] suggest a number of reduced blocking schemes. One of the simplest proceeds by noting that $\boldsymbol{\beta}$ and $\{\mathbf{b}_i\}$ can be sampled in one block by the method of composition: first sampling $\boldsymbol{\beta}$ marginalized over $\{\mathbf{b}_i\}$ and then sampling $\{\mathbf{b}_i\}$ conditioned on $\boldsymbol{\beta}$. What makes reduced blocking possible is the fact that the conditional distribution of the outcomes marginalized over $\mathbf{b}_i$ is normal which can be combined with the assumed normal prior on $\boldsymbol{\beta}$ in the usual way. In particular,

$$\begin{aligned} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}, \sigma^2) &= \int f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \sigma^2) g(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i \\ &\propto |\mathbf{V}_i|^{-1/2} \exp\{(-1/2)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\}\ , \end{aligned}$$

where $\mathbf{V}_i = \sigma^2\boldsymbol{\Omega}_i + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$, which, from Result 1, leads to the conditional posterior of $\boldsymbol{\beta}$ (marginalized over $\{\mathbf{b}_i\}$).

The rest of the algorithm follows the steps of *Wakefield et. al* [1994]. In particular, the sampling of the random effects is from independent normal distributions that are derived by treating $(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ as the "data," $\mathbf{b}_i$ as the regression coefficient and $\mathbf{b}_i \sim \mathcal{N}_q(0, \mathbf{D})$ as the prior and applying Result 2. Next, conditioned on $\{\mathbf{b}_i\}$, the full conditional distribution of $\mathbf{D}^{-1}$ becomes independent of $\mathbf{y}$ and is obtained by combining the Wishart prior distribution of $\mathbf{D}^{-1}$ with the normal distribution of $\{\mathbf{b}_i\}$ given $\mathbf{D}^{-1}$. The resulting

distribution is Wishart with updated parameters obtained from Result 3. Finally, Result 4 yields the full-conditional distribution of σ^2. *In applying these results, λ_i and η_i are set equal to one in all the expressions.*

Algorithm: **Gaussian-Gaussian Panel (*Wakefield et. al* [1994] and *Chib and Carlin* [1999])**

1. Sample

   (a)
   $$\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

   (b)
   $$\mathbf{b}_i|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \leq N$$

2. Sample
   $$\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Sample
   $$\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D} \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \delta}{2}\right)$$

4. Goto 1

2.1.1 Example

As an illustration, we consider data from a clinical trial on the effectiveness of two antiretroviral drugs (didanosine or ddI and zalcitabine or ddC) in 467 persons with advanced HIV infection. The response variable y_{ij} for patient i at time j is the square root of the patient's CD4 count, a seriological measure of immune system health and prognostic factor for AIDS-related illness and mortality. The data set records patient CD4 counts at study entry and again at 2, 6, 12, and 18 months after entry, for the ddI and ddC groups, respectively.

The model is formulated as follows. If we let $\mathbf{y}_i$ denote a n_i vector of responses across time for the ith patient, then following the discussion in *Carlin and Louis* [2000], suppose

$$\begin{aligned} \mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \sigma^2 &\sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2\boldsymbol{\Omega}_i),\ \boldsymbol{\Omega}_i = \mathbf{I}_{n_i} \\ \mathbf{b}_i|\mathbf{D} &\sim \mathcal{N}_2(\mathbf{0}, \mathbf{D})\ , i \leq 467\,, \end{aligned} \tag{19}$$

where the j^{th} row of the patient i's design matrix $\mathbf{W}_i$ takes the form $\mathbf{w}_{ij} = (1, t_{ij})$, t_{ij} belongs to the set $\{0, 2, 6, 12, 18\}$ and the fixed design matrix $\mathbf{X}_i$ is obtained by horizontal concatenation of $\mathbf{W}_i$, $d_i\mathbf{W}_i$ and $a_i\mathbf{W}_i$, where d_i is a binary variable indicating whether patient i received ddI ($d_i = 1$) or ddC ($d_i = 0$), and a_i is a binary variable indicating if the patient was diagnosed as having AIDS at baseline ($a_i = 1$) or not ($a_i = 0$).

The prior distribution of $\boldsymbol{\beta} : 6 \times 1$ is assumed to be $\mathcal{N}_6(\boldsymbol{\beta}_0, \mathbf{B}_0)$ with

$$\begin{aligned}\boldsymbol{\beta}_0 &= (10, 0, 0, 0, -3, 0) \ , \quad \text{and} \\ \mathbf{B}_0 &= \text{Diag}(2^2, 1^2, (.1)^2, 1^2, 1^2, 1^2) \ ,\end{aligned}$$

while that on $\mathbf{D}^{-1}$ is taken to be Wishart $W(\mathbf{R}_0/\rho_0, 2, \rho_0)$ with $\rho_0 = 24$ and $\mathbf{R}_0 = \text{diag}(.25, 16)$. Finally, σ^2 is apriori assumed to follow the inverse-gamma distribution

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right) \ ,$$

with $\nu_0 = 6$ and $\delta_0 = 120$ (which imply a prior mean and standard deviation both equal to 30).

The MCMC simulation is run for 5000 cycles beyond a burn-in of a 100 cycles. The simulated values by iteration for each of the ten parameters are given in Figure 1.2. Except for the parameters that are approximately the same, the sampled paths of the parameters are clearly visible and display little correlation.

These draws from the posterior distribution are used to produce different summaries of the posterior distribution. In Figure 1.3 we report the marginal posterior distributions in the form of histogram plots. We see that three of the regression parameters are centered at zero, that D_{11} is large and D_{22} (which is the variance of the time-trend random effect) is small.

2.2 Robust modeling of $\mathbf{b}_i$: Student-Student and Student-mixture models

We now discuss models in which the error distribution of the observations in the ith cluster is multivariate-t and the distribution of $\mathbf{b}_i$ is modeled as multivariate-t or a mixture of normals. To begin, consider the student-

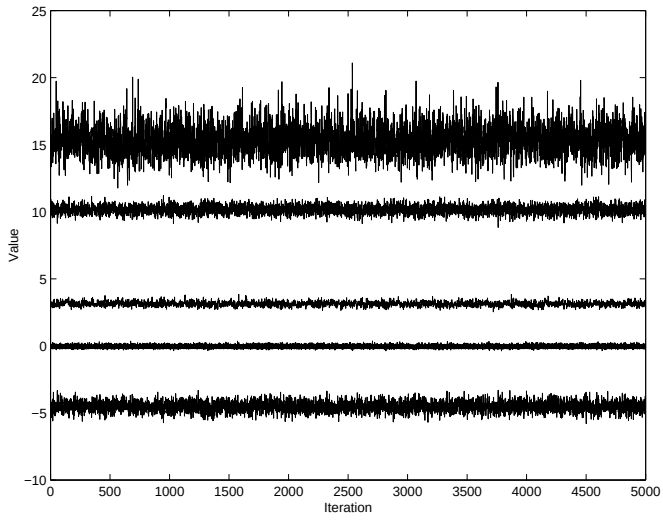

Figure 2: Aids clustered data: Simulated values by iteration for each of ten parameters.

student model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)$$
$$\mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D}); i \leq N$$
$$\lambda_i \sim \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right);\ \eta_i \sim \mathcal{G}\left(\frac{\nu_F}{2}, \frac{\nu_F}{2}\right)$$
$$\boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \mathbf{B}_0);\ \sigma^2 \sim \mathcal{IG}\left(\frac{v_0}{2}, \frac{\delta_0}{2}\right);\ \mathbf{D}^{-1} \sim \mathcal{W}_p(\rho_0, \mathbf{R}_0)$$

This model is easily analyzed by including λ_i and η_i, $i \leq N$, in the sampling. In that case, we follow the Gaussian-Gaussian MCMC algorithm, except that each step is implemented conditioned on $\{\lambda_i\}$ and $\{\eta_i\}$ and two new steps are added in which $\{\lambda_i\}$ and $\{\eta_i\}$ are sampled. The quantities that go into forming the various parameters in these updates are all obtained from the results of Section 1.3.

Algorithm: **Student-Student Panel**

1. Sample

   (a)
   $$\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \mathbf{D}, \{\lambda_i\}, \{\eta_i\} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

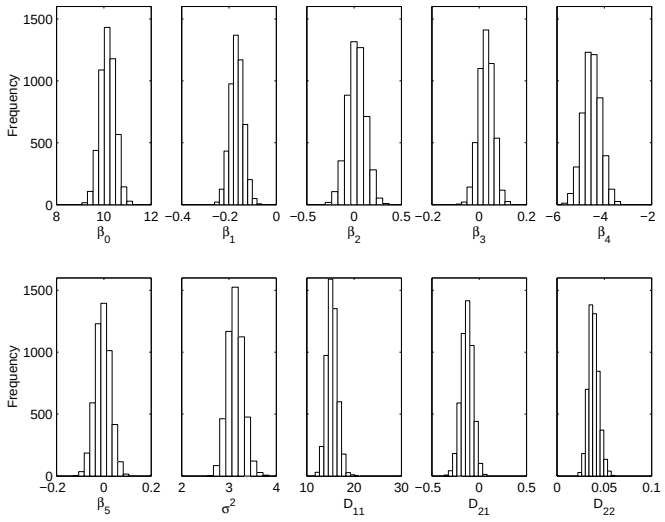

Figure 3: Aids clustered data: Marginal posterior distributions of parameters based on 5000 MCMC draws.

(b)

$$\mathbf{b}_i|\mathbf{y},\boldsymbol{\beta},\sigma^2,\mathbf{D},\lambda_i,\eta_i \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i,\mathbf{D}_i\right),$$

2. Sample

(a)

$$\lambda_i|\mathbf{y},\boldsymbol{\beta},\{\mathbf{b}_i\},\sigma^2 \sim \mathcal{G}\left(\frac{\nu_G+n_i}{2},\frac{\nu_G+\sigma^{-2}\mathbf{e}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{e}_i}{2}\right),$$

(b)

$$\eta_i|\mathbf{b}_i,\mathbf{D} \sim \mathcal{G}\left(\frac{\nu_F+q}{2},\frac{\nu_F+\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i}{2}\right), i \leq N$$

3. Sample

$$\mathbf{D}^{-1}|\mathbf{y},\boldsymbol{\beta},\{\mathbf{b}_i\},\sigma^2,\{\lambda_i\},\{\eta_i\} \sim \mathcal{W}_q\left\{\rho_0+N,\mathbf{R}\right\}$$

4. Sample

$$\sigma^2|\mathbf{y},\boldsymbol{\beta},\{\mathbf{b}_i\},\mathbf{D},\{\lambda_i\},\{\eta_i\} \sim \mathcal{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right)$$

5. Goto 1

Another possibility is to assume that $\mathbf{b}_i$ is drawn from a finite mixture of Gaussian distributions. For example, one may assume that $\mathbf{b}_i \sim q_1\mathcal{N}(\mathbf{0}, \mathbf{D}_1) + q_2\mathcal{N}(\mathbf{0}, \mathbf{D}_2 = \eta\mathbf{D}_1)$ where $\eta > 1$ and q_j is the probability of drawing from the jth component of the mixture. Chen and Dunson (2003), for example, use a particular mixture prior in which one of the component random effects variances can be zero, which leads to a method for determing if the particular effect is random. Like any Bayesian analysis of a mixture model, analysis exploits the hierarchical representation of the mixture distribution:

$$\begin{aligned}\mathbf{b}_i|s_i = j &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}_j)\\ \Pr(s_i = j) &= q_j \ \ , j = 1, 2\end{aligned}$$

where $s_i = \{1, 2\}$ is a latent population indicator variable. The MCMC based fitting of this Gaussian-mixture model proceeds by sampling the posterior distribution

$$\begin{aligned}\pi(\boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}_1^{-1}, \sigma^2, \{\lambda_i\}, \eta, \{s_i\}, q|\mathbf{y}) &= \pi(\boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}_1^{-1}, \sigma^2,\\ &\quad \{\lambda_i\}, \eta, \{s_i\}) f(\mathbf{y}|\ \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2, \{\lambda_i\})\\ = \pi(\boldsymbol{\beta})\pi(\mathbf{D}^{-1})\pi(\sigma^2)\pi(\lambda)\pi(q) &\prod_{i=1}^{N} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \sigma^2, \lambda_i) p(\mathbf{b}_i|s_i, \mathbf{D}_{s_i}) p(s_i|q) p(\lambda_i)\end{aligned}$$

where the prior on η is (say) inverse-gamma and that of $q = (q_1, q_2)$ a Dirichlet with density proportional to $q_1^{m_{10}-1}q_2^{m_{20}-1}$ where the hyper-parameters m_{10} and m_{20} are known. This posterior density is sampled with some minor modifications of the Student-Student algorithm. Steps 1 and 2 are now conditioned on $\{s_i\}$; as a result $\mathbf{V}_i$ in the updates is replaced by $\mathbf{V}_{s_i} = \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i + \mathbf{W}_i\mathbf{D}_{s_i}\mathbf{W}_i'$ and $\mathbf{D}_i$ by $\mathbf{D}_{s_i}^* = (\mathbf{D}_{s_i}^{-1} + \sigma^{-2}\lambda_i\mathbf{W}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i)^{-1}$. Step 3 is now the sampling of $\mathbf{D}_1^{-1}$ where the sum over the outer-product of the $\mathbf{b}_i$'s is replaced by $\sum_{i:s_i=1}\mathbf{b}_i\mathbf{b}_i' + \eta^{-1}\sum_{i:s_i=2}\mathbf{b}_i\mathbf{b}_i'$. Steps 4 and 5 are unchanged. Finally, two new steps are inserted: Step 6 for sampling η and Step 6 for sampling q. Each of these steps is straightforward. In Step 6 we sample η from an updated inverse-gamma distribution based on those $\mathbf{b}_i$ that are associated with population 2; the update is therefore from the model $\mathbf{b}_i|\eta, \mathbf{D}_1 \sim \mathcal{N}(\mathbf{0}, \eta\mathbf{D}_1)$, $\eta \sim IG(a_0/2, b_0/2)$ which leads to an inverse-gamma distribution. The updated distribution of q in Step 7 is easily seen to be Dirichlet with parameters $m_{10}+ m_1$ and $m_{20}+m_2$, respectively, where m_j are the total number of observations ascribed to population j in that iteration of the MCMC sampling.

2.3 Heteroskedasticity

The methods described above are readily adapted to deal with heteroskedasticity in the observation error process by parameterizing the error covariance matrix $\sigma^2\mathbf{\Omega}_i$. Instead of assuming that $\boldsymbol{\varepsilon}_i|\sigma^2 \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{\Omega}_i)$, we assume

$$\boldsymbol{\varepsilon}_i|\sigma_i^2 \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_i^2\mathbf{I}_{n_i})$$

where σ_i^2 can be modeled hierarchically by assuming that

$$\sigma_i^2|\delta_0 \sim IG\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)$$
$$\delta_0 \sim G\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right)$$

a specification that appears in *Basu and Chib* [2003]. In the first stage of this prior specification, one assumes that conditioned on the scale of the inverse-gamma distribution, σ_i^2 is inverse-gamma and then the scale is in turn allowed to follow a gamma distribution. The fitting of this model is quite similar to the fitting of the Gaussian-Gaussian model except that $\sigma^2\mathbf{\Omega}_i$ is replaced by $\sigma_i^2\mathbf{I}_{n_i}$ in Steps 1 and 2, $\mathbf{\Omega}_i$ is replaced by $\mathbf{I}_{n_i}$ in Step 3, Step 4 is modified and a new Step 5 is inserted for the sampling of δ_0.

Algorithm: **Gaussian-Gaussian Hetroskedastic Panel (*Basu and Chib* [2003])**

1. Sample

   (a)
   $$\boldsymbol{\beta}|\mathbf{y}, \left\{\sigma_i^2\right\}, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

   (b)
   $$\mathbf{b}_i|\mathbf{y}, \boldsymbol{\beta}, \left\{\sigma_i^2\right\}, \mathbf{D} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \leq N$$

2. Sample
   $$\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \{b_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Sample
   $$\sigma_i^2|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}, \delta_0 \sim \mathcal{IG}\left(\frac{\nu_0 + n_i}{2}, \frac{\delta_0 + \parallel \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i \parallel^2}{2}\right)$$

4. Sample

$$\delta_0|\sigma_i^2 \sim \mathcal{G}\left(\frac{\nu_0+\nu_{00}}{2}, \frac{\sigma_i^{-2}+\delta_{00}}{2}\right)$$

5. Goto 1

2.4 Serial Correlation

To deal with the possibility of serial correlation in models with multivariate-t error and random effects distributions we now assume that

$$\boldsymbol{\varepsilon}_i|\lambda_i, \boldsymbol{\phi} \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)$$

where $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_i(\boldsymbol{\phi})$ is a $n_i \times n_i$ covariance matrix that depends on a set of p parameters $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)$. Typically, one will assume that the errors follow a low-dimensional stationary ARMA process and the matrix $\boldsymbol{\Omega}_i$ will then be the covariance matrix of the n_i errors. In that case, $\boldsymbol{\phi}$ represents the parameters of the assumed ARMA process. The fitting of this model by MCMC methods is quite straightforward. The one real new step is the sampling of $\boldsymbol{\phi}$ by the M-H algorithm along the lines of *Chib and Greenberg* [1994].

Algorithm: **Student-Student Correlated Error Panel**

1. Sample

   (a)

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \mathbf{D}, \{\lambda_i\}, \boldsymbol{\phi} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

   (b)

$$\mathbf{b}_i|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}, \{\lambda_i\}, \{\eta_i\}, \boldsymbol{\phi} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right) \ ,$$

2. Sample

   (a)

$$\lambda_i|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2, \boldsymbol{\phi} \sim \mathcal{G}\left(\frac{\nu_G+n_i}{2}, \frac{\nu_G+\sigma^{-2}\mathbf{e}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{e}_i}{2}\right),$$

   (b)

$$\eta_i|\mathbf{b}_i, \mathbf{D} \sim \mathcal{G}\left(\frac{\nu_F+q}{2}, \frac{\nu_F+\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i}{2}\right), i \leq N$$

3. Sample

$$\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2, \{\eta_i\}, \boldsymbol{\phi} \sim \mathcal{W}_q \{\rho_0 + N, \mathbf{R}\}$$

4. Sample

$$\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}, \boldsymbol{\phi} \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \delta}{2}\right)$$

5. Sample

$$\boldsymbol{\phi}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}, \sigma^2, \{\lambda_i\} \propto \pi(\boldsymbol{\phi}) \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i\right)$$

6. Goto 1

In the sampling of $\boldsymbol{\phi}$ in the above algorithm we use the tailored proposal density as suggested by *Chib and Greenberg* [1994]. Let

$$\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \ln \underbrace{\left\{\pi(\boldsymbol{\phi}) \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i\right)\right\}}_{g(\boldsymbol{\phi})}$$

be the conditional mode of the full conditional of $\boldsymbol{\phi}$ that is found by (say) a few steps of the Newton-Raphson algorithm, and let $\mathbf{V}$ be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of $\ln g(\boldsymbol{\phi})$ evaluated at $\hat{\boldsymbol{\phi}}$. Then, our proposal density is given by

$$q(\boldsymbol{\phi}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}, \sigma^2) = t_p(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}}, \mathbf{V}, \nu)$$

a multivariate-t density with mean $\hat{\boldsymbol{\phi}}$, dispersion matrix $\mathbf{V}$ and ν degrees of freedom. In this M-H step, given the current value $\boldsymbol{\phi}$, we now generate a proposal value $\boldsymbol{\phi}'$ from this multivariate-t density and accept or reject with probability of move

$$\alpha(\boldsymbol{\phi}, \boldsymbol{\phi}'|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D}, \sigma^2) = \min\left\{1, \frac{g(\boldsymbol{\phi}')}{g(\boldsymbol{\phi})} \frac{t_p(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}}, \mathbf{V}, \nu)}{t_p(\boldsymbol{\phi}'|\hat{\boldsymbol{\phi}}, \mathbf{V}, \nu)}\right\}$$

If the proposal value is rejected we stay at the current value $\boldsymbol{\phi}$ and move to Step 1 of the algorithm. As before, by setting λ_i and η_i to one we get the Gaussian-Gaussian version of the autoregressive model.

3 Binary Responses

Consider now the situation in which the response variable is binary $(0, 1)$ and the objective is to fit a panel model with random effects. The classical analysis of such models (under the probit link) was pioneered by *Chamberlain* [1980], *Heckman* [1981] and *Butler and Moffit* [1982].

Suppose that for the ith individual at time t, the probability of observing the outcome $y_{it} = 1$, conditioned on the random effect $\mathbf{b}_i$, is given by

$$\Pr(y_{it} = 1|\mathbf{b}_i) = \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i)\,,$$

where Φ is the cdf of the standard normal distribution, and $\mathbf{b}_i|\mathbf{D} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{D})$ independent of $\mathbf{x}_{it}$. Since the n_i observations in the ith cluster are correlated, the joint density of the observations $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ is

$$\begin{aligned}\Pr(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}) = \int \left\{ \prod_{t=1}^{T} [\Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i)]_{it}^{y} [1 - \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i)]^{1-y_{it}} \right\} \\ \times \mathcal{N}(\mathbf{b}_i|\mathbf{0}, \mathbf{D})\, d\,\mathbf{b}_i\end{aligned}$$

Under the assumption that the observations across individuals are independent, the likelihood function of the parameters $(\boldsymbol{\beta}, \mathbf{D})$ is the product of $\Pr(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D})$. Although methods are now available to evaluate this integral under some special circumstances, it turns out that it is possible to circumvent the calculation of the likelihood function. The method relies on the approach that was introduced by *Albert and Chib* [1993].

To understand the *Albert and Chib* algorithm, consider the cross-section binary probit model in which we are given n random observations such that $\Pr(y_i = 1) = \Phi(\mathbf{x}'_i\boldsymbol{\beta})$. An equivalent formulation of the model is in terms of latent variables $\mathbf{z} = (z_1, ..., z_n)$ where

$$z_i|\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}'_i\boldsymbol{\beta}, 1), \;\; y_i = I[z_i > 0]\,,$$

and I is the indicator function. *Albert and Chib* [1993] exploit this equivalence and propose that the latent variables $\{z_1, ..., z_n\}$, 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

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

where the term in braces in the probability of y_i given $(\boldsymbol{\beta}, z_i)$ and is one for $y_i = 1$ when $z_i > 0$ and is one for $y_i = 0$ when $z_i < 0$. The latter posterior density is sampled by a two-block Gibbs sampler composed of the full conditional distributions:

1. $\boldsymbol{\beta}|\mathbf{y}, \mathbf{z}$
2. $\mathbf{z}|\mathbf{y}$,$\boldsymbol{\beta}$.

Even though the parameter space has been enlarged, the introduction of the latent variables simplifies the problem considerably. The first conditional distribution, i.e., $\boldsymbol{\beta}|\mathbf{y}, \mathbf{z}$, is the same as the distribution $\boldsymbol{\beta}|\mathbf{z}$ since knowledge of $\mathbf{z}$ means that $\mathbf{y}$ has no additional information for $\boldsymbol{\beta}$. The distribution $\boldsymbol{\beta}|\mathbf{z}$ is easy to derive since the response variable is continuous. The second conditional distribution, i.e., $\mathbf{z}|\mathbf{y}, \boldsymbol{\beta}$, factors into n distributions $z_i|y_i, \boldsymbol{\beta}$ and is easily seen to be truncated normal given the value of y_i. Specifically, if $y_i = 1$, then

$$z_i \sim \mathcal{TN}_{(0,\infty)}(\mathbf{x}_i'\boldsymbol{\beta}, 1) \tag{20}$$

a truncated normal distribution with support $(0, \infty)$, whereas if $y_i = 0$, then

$$z_i \sim \mathcal{TN}_{(-\infty,0]}(\mathbf{x}_i'\boldsymbol{\beta}, 1) \tag{21}$$

a truncated normal distribution with support $(-\infty, 0)$. These truncated normal distributions are simulated by the method given in Section 1.4. For the case of (20), it reduces to

$$\mathbf{x}_i'\boldsymbol{\beta} + \Phi^{-1}\left[\Phi(-\mathbf{x}_i'\boldsymbol{\beta}) + U\left(1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta})\right)\right]$$

and for the case (21) to

$$\mathbf{x}_i'\boldsymbol{\beta} + \Phi^{-1}\left[U\Phi(-\mathbf{x}_i'\boldsymbol{\beta})\right],$$

where U is a uniform random variable on (0,1). Hence, the algorithm proceeds through the simulation of $\boldsymbol{\beta}$ given the latent data and the simulation of the latent data given $(\mathbf{y}, \boldsymbol{\beta})$.

Given this framework, the approach for the panel probit model becomes transparent. For the ith cluster, we define the vector of latent variable

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

and let

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

where $\mathbf{z}_i = (z_{i1}, ..., z_{in_i})'$, $\mathbf{W}_i$ is a set of variables whose effect $\mathbf{b}_i$ is assumed to heterogenous, $\mathbf{X}_i$ is a set of raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i} \;\; \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$. The MCMC implementation in this set-up proceeds by including the $\{z_{it}\}$ in the sampling. Given the $\{z_{it}\}$ the sampling resembles the steps of the Gaussian-Gaussian algorithm with z_{it} playing the role of y_{it} and $\sigma^2\lambda_i^{-1}\mathbf{\Omega}_i = \mathbf{I}_{n_i}$. The sampling of z_{it} is done marginalized over $\{\mathbf{b}_i\}$ from the conditional distribution of $z_{it}|\mathbf{z}_{i(-t)}, y_{it}, \boldsymbol{\beta}, \mathbf{D}$, where $\mathbf{z}_{i(-t)}$ is the vector $\mathbf{z}_i$ excluding z_{it}. It should be emphasized that the simulation of these distributions does not require the evaluation of the likelihood function.

Algorithm: **Gaussian-Gaussian Panel Probit (*Chib and Carlin* [1999])**

1. Sample

   (a)

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

   (b)

$$\boldsymbol{\beta}|\{z_{it}\}, \mathbf{D} \sim \mathcal{N}_k\left(\mathbf{B}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{z}_i), \mathbf{B}\right)$$
$$\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1};\;\; \mathbf{V}_i = \mathbf{I}_{n_i} + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$$

   (c)

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

2. Sample
$$\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\} \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Goto 1

Because of this connection with the continuous case, the analysis of binary panel data may be extended in ways that parallel the developments in the previous section. For example, we can analyze binary data under the assumption that ε_i is multivariate-t and/or the assumption that the random effects distribution is student-t or a mixture of normals. We present the algorithm for the student-student binary response panel model without comment.

Algorithm: **Student-Student Binary Panel**

1. Sample

   (a)

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

   (b)

$$\boldsymbol{\beta}|\{z_{it}\}, \mathbf{D}\left\{\lambda_i\right\}, \{\eta_i\} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

   (c)

$$\mathbf{b}_i|\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{D}, \lambda_i, \eta_i \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right)$$

2. Sample

   (a)

$$\lambda_i|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{G}\left(\frac{\nu_G+n_i}{2}, \frac{\nu_G+\sigma^{-2}\mathbf{e}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{e}_i}{2}\right), i \leq N$$

   (b)

$$\eta_i|\mathbf{b}_i, \mathbf{D} \sim \mathcal{G}\left(\frac{\nu_G+q}{2}, \frac{\nu_G+\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i}{2}\right), i \leq N$$

3. Sample

$$\mathbf{D}^{-1}|\left\{z_{it}\right\}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \{\lambda_i\}, \{\eta_i\} \sim \mathcal{W}_q\left\{\rho_0+N, \mathbf{R}\right\}$$

4. Goto 1

The fact that this, and other model variants for binary responses, are handled effortlessly is a testament to the flexibility and power of the Bayesian approach.

4 Other Outcome Types

4.1 Censored Outcomes

Given the discussion of the binary response models in the preceding section it should not be surprising that the Bayesian approach to censored data would proceed in much the same fashion. Consider then a Gaussian-Gaussian Tobit panel data model for the ith cluster:

$$\begin{aligned} z_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}) \\ \mathbf{b}_i &\sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \\ (\boldsymbol{\beta}, \sigma^2, \mathbf{D}) &\sim \pi \end{aligned}$$

where the observed outcomes are obtained as

$$y_{it} = \max\{z_{it}, 0\}$$

This model is fit along the lines of the Gaussian-Gaussian model by adopting the strategy of *Chib* [1992] wherein one simulates z_{it} for those observations that are censored from a truncated normal distribution, truncated to the interval $(-\infty, 0)$. In our description of the fitting method we let $\mathbf{y}_{i\mathbf{z}}$ be a $n_i \times 1$ vector with ith component y_{it} if that observation is not censored and z_{it} if it is censored. A new Step 1 is inserted in which the latent z_{it} are sampled conditioned on the remaining values of $\mathbf{y}_{i\mathbf{z}}$ in the ith cluster, which we denote by $y_{iz(-t)}$; then in Step 2 the only change is that instead of $\mathbf{y}_i$ we use $\mathbf{y}_{i\mathbf{z}}$; in Step 3 in the sampling of $\mathbf{b}_i$ we replace the vector $\mathbf{y}_i$ by the most current value of $\mathbf{y}_{i\mathbf{z}}$; Step 4 for the sampling of $\mathbf{D}^{-1}$ is unchanged; and in Step 5 dealing with the sampling of σ^2 we use $\mathbf{y}_{i\mathbf{z}}$ in place of $\mathbf{y}_i$ in the definition of δ.

Algorithm: **Gaussian-Gaussian Tobit Panel**

1. Sample

   (a)

$$\begin{aligned} z_{it}|y_{iz(-t)}, y_{it}, \boldsymbol{\beta}, \sigma^2, \mathbf{D} &\propto \mathcal{N}(\mu_{it}, v_{it})\, I(z_{it} < 0) \quad \text{if } y_{it} = 0 \\ \mu_{it} &= \mathrm{E}(z_{it}|y_{iz(-t)}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \\ v_{it} &= \mathrm{Var}(z_{it}|y_{iz(-t)}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \end{aligned}$$

   (b)

$$\boldsymbol{\beta}|\mathbf{y}_z, \sigma^2, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

(c)

$$\mathbf{b}_i|\mathbf{y}_z, \boldsymbol{\beta}, \sigma^2, \mathbf{D} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \leq N$$

2. Sample

$$\mathbf{D}^{-1}|\mathbf{y}_z, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Sample

$$\sigma^2|\mathbf{y}_z, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D} \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \delta}{2}\right)$$

4. Goto 1

Just as in the case of continuous and binary outcomes, this algorithm is easily modified to allow the random effects have a student-t or a mixture of normals distribution and to allow the observation errors be student-t. Analysis of any of these models is quite difficult from the frequentist perspective.

4.2 Count Responses

Bayesian methods are also effectively applied to panel data in which the responses are counts. A framework for fitting such models under the assumption that the distribution of the counts, given the random effects, is Poisson is developed by *Chib, Greenberg and Winkelmann* [1998]. To describe the set-up, for the ith cluster

$$\begin{aligned} y_{it}|\boldsymbol{\beta}, \mathbf{b}_i &\sim \text{Poisson}(\lambda_{it}) \\ \ln(\lambda_{it}) &= \ln \tau_{it} + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i \end{aligned}$$

where the covariate vectors $\mathbf{x}'_{it}$ and $\mathbf{w}'_{it}$ are the tth row of the matrices $\mathbf{X}_i$ and $\mathbf{W}_i$, respectively, and $\mathbf{X}_i$ are the raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i} \;\; \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$. The quantity τ_{it} which is one if each count is measured over the same interval of time. This specification of the model produces the likelihood function

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}) = \prod_{i=1}^{n} \int \mathcal{N}(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) d\mathbf{b}_i \tag{22}$$

where

$$p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) = \prod_{t=0}^{n_i} \frac{\lambda_{it}^{y_{it}} \exp\left(-\lambda_{it}\right)}{y_{it}!} \tag{23}$$

is the product of the Poisson mass function with mean λ_{it}.

The interesting aspect of the MCMC algorithm in this case is the sampling of both $\boldsymbol{\beta}$ and $\{\mathbf{b}_i\}$ by tailored M-H steps. This is because the full conditional distributions in this model do not belong to any known family of distributions. At each step of the algorithm, there are $n+1$ M-H steps. It may appear that the computational burden is high when n is large. This turns out not to be case.

Algorithm: **Panel Poisson (*Chib, Greenberg and Winkelmann* [1998])**

1. Calculate the parameters $(\mathbf{m}_0, \mathbf{V}_0)$ as the mode and inverse of the negative Hessian of

$$\log \mathcal{N}_k(\ \boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0) + \sum_{i=1}^{N} \log p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)$$

   propose $\boldsymbol{\beta}' \sim \mathcal{T}(\boldsymbol{\beta}|\mathbf{m}_0, \mathbf{V}_0, \nu)$ (the multivariate-t density) and move to $\boldsymbol{\beta}'$ with probability

$$\min\left\{\frac{\prod_{i=1}^{N} p(\mathbf{y}_i|\boldsymbol{\beta}', \mathbf{b}_i)\mathcal{N}_k(\boldsymbol{\beta}'|\mathbf{0}, \mathbf{B})}{\prod_{i=1}^{N} p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)\mathcal{N}_k(\boldsymbol{\beta}|\mathbf{0}, \mathbf{B})} \frac{\mathcal{T}(\boldsymbol{\beta}|\mathbf{m}_0, \mathbf{V}_0, \nu)}{\mathcal{T}(\boldsymbol{\beta}'|\mathbf{m}_0, \mathbf{V}_0, \nu)}, 1\right\}$$

2. Calculate the parameters $(\mathbf{m}_i, \mathbf{V}_i)$ as the mode and inverse of the negative Hessian of

$$\log \mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) + \log p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)$$

   propose $\mathbf{b}_i' \sim \mathcal{T}(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)$ and move to $\mathbf{b}_i'$ with probability

$$\min\left\{\frac{p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i')\mathcal{N}_q(\mathbf{b}_i'|\mathbf{0}, \mathbf{D})}{p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i)\mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D})} \frac{\mathcal{T}(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)}{\mathcal{T}(\mathbf{b}_i'|\mathbf{m}_i, \mathbf{V}_i, \nu)}, 1\right\}$$

3. Sample

$$\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

4.3 Multinomial responses

Multinomial panel responses arise in several different areas and the fitting of this model when the link function is assumed to be multinomial logit is exactly the same as the algorithm for count responses. The only difference is

that instead of the Poisson link function we now have the multinomial logit link. Let y_{it} be a multinomial random variable taking values $\{0, 1, ..., J\}$ and assume that

$$\Pr(y_{it} = j|\boldsymbol{\beta}, \mathbf{b}_i) = \frac{\exp\left(\alpha_j + \mathbf{x}'_{itj}\boldsymbol{\beta} + \mathbf{w}'_{itj}\mathbf{b}_i\right)}{\sum_{l=0}^{J}\exp\left(\alpha_l + \mathbf{x}'_{itl}\boldsymbol{\beta} + \mathbf{w}'_{itl}\mathbf{b}_i\right)}$$

where for identifiability α_0 is set equal to zero. The joint probability of the outcomes in the ith cluster, conditioned on the $\mathbf{b}_i$, is now given by

$$p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) = \prod_{t=1}^{n_i}\Pr(y_{it} = j_t|\boldsymbol{\beta}, \mathbf{b}_i) \tag{24}$$

where j_t is the observed outcome at time t. The structure of the problem is seen to be identical to that in the count case and the preceding algorithm applies directly to this problem by replacing the mass function in (23) with the mass function in (24).

Chiang, Chib and Narasimhan [1999] develop an interesting variant of this model in which the possible values that y_{it} can take is not the same across clusters. Such a situation arises when the multinomial outcomes are choices made by a subject (for example choice of transportation mode or choice of brand of a product) and where the assumption that the choice set is the same across subjects is too strong and must be relaxed. The model discussed in the paper only appears to be fittable by Bayesian methods. The paper includes a detailed example.

5 Binary Endogenous Regressor

In many applied studies, one is interested in the effect of a given (binary) covariate on the response but under the complication that the binary covariate is not sequentially randomly assigned. In other words, the assumption of sequential exogeneity is violated. This problem has not been extensively studied in the literature but interestingly it is possible to develop a Bayesian approach to inference that in many ways is quite straightforward. For concreteness, suppose that in the context of the model in (2) the last covariate in $\mathbf{x}_{1it}$ (namely x_{12it}) is the covariate of interest and the model is given by

$$y_{it} = \mathbf{x}'_{11it}\boldsymbol{\beta}_{11} + x_{12it}\beta_{12} + \mathbf{w}'_{it}\mathbf{c}_{2i} + e_{it}$$

where $\mathbf{x}_{1it} = (\mathbf{x}_{11it}, x_{12it})$ and $\mathbf{x}_{11it} : k_{11} \times 1$. Assume that the covariates $\mathbf{x}_{11it}$ and $\mathbf{w}_{it} : q \times 1$ satisfy the assumption of sequential exogeneity but that

x_{12it} does not. Now let $z_{it} : k_z \times 1$ be time-varying instruments and suppose that the model generating the endogenous covariate is

$$x_{12it} = I\left(\mathbf{x}'_{11it}\boldsymbol{\gamma} + \mathbf{w}'_{it}\mathbf{d}_{2i} + \mathbf{z}_{it}\boldsymbol{\delta} + u_{it} > 0\right)$$

where

$$\begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & 1 \end{pmatrix}\right)$$

and $\omega_{12} \neq 0$. Letting $x^*_{12it} = \mathbf{x}'_{11it}\boldsymbol{\gamma} + \mathbf{w}'_{it}\mathbf{d}_{3i} + \mathbf{z}_{it}\boldsymbol{\delta} + u_i$, the model is reexpressed as

$$\underbrace{\begin{pmatrix} y_{it} \\ x^*_{12it} \end{pmatrix}}_{\mathbf{y}^*_{it}} = \underbrace{\begin{pmatrix} \mathbf{x}'_{11it} & x_{12it} & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{0} & \mathbf{x}'_{11it} & z_{it} \end{pmatrix}}_{\mathbf{X}_{1it}} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_{11} \\ \beta_{12} \\ \boldsymbol{\gamma} \\ \boldsymbol{\delta} \end{pmatrix}}_{\boldsymbol{\beta}_1} + \underbrace{\begin{pmatrix} \mathbf{w}'_{it} & \mathbf{0}' \\ \mathbf{0}' & \mathbf{w}'_{it} \end{pmatrix}}_{\mathbf{W}_{it}} \underbrace{\begin{pmatrix} \mathbf{c}_{2i} \\ \mathbf{d}_{2i} \end{pmatrix}}_{\boldsymbol{\beta}_{2i}} + \underbrace{\begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}}_{\boldsymbol{\varepsilon}_{it}}$$

or as

$$\mathbf{y}^*_{it} = \mathbf{X}_{1it}\boldsymbol{\beta}_1 + \mathbf{W}_{it}\boldsymbol{\beta}_{2i} + \boldsymbol{\varepsilon}_{it}$$

where β_1 is $k_1 \times 1$ with $k_1 = 2k_{11} + 1 + k_z$ and β_{2i} is $2q \times 1$. If we assume that $\boldsymbol{\beta}_{2i}$ as before is modeled in terms of covariates $\mathbf{a}_i : r \times 1$ as

$$\begin{pmatrix} \mathbf{c}_{2i} \\ \mathbf{d}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_q \otimes \mathbf{a}'_i & \mathbf{0}' \\ \mathbf{0}' & \mathbf{I}_q \otimes \mathbf{a}'_i \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{21} \\ \boldsymbol{\beta}_{22} \end{pmatrix} + \begin{pmatrix} \mathbf{b}_{1i} \\ \mathbf{b}_{2i} \end{pmatrix}$$

or compactly as

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

where $\boldsymbol{\beta}_2 : k_2 \times 1$ and $k_2 = 2qr$, then we can rewrite the outcome vector for subject i at time t as

$$\mathbf{y}^*_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{W}_{it}\mathbf{b}_i + \boldsymbol{\varepsilon}_{it}$$

where

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

$\beta = (\beta_1, \beta_2) : k \times 1$, and $k = k_1 + k_2$. This is similar to the models that we have dealt with except that this is a system of two equations for each (i, t) with the second component of the outcome being latent. For the ith cluster

the preceding model (in conjunction with the standard assumptions about $\mathbf{b}_i$) is written as

$$\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \ , \ \boldsymbol{\varepsilon}_i|\lambda_i, \boldsymbol{\Omega} \sim \mathcal{N}_{2n_i}(\mathbf{0}, \lambda_i^{-1}\{I_{n_i} \otimes \boldsymbol{\Omega}\})$$
$$\mathbf{b}_i|\mathbf{D} \sim \mathcal{N}_{2q}(\mathbf{0}, \mathbf{D})$$

This model is fit along the lines of the binary panel by simulating $\{x_{12it}^*\}_{t=1}^{n_i}$ in $\mathbf{y}_i^*$ (these appear in rows 2, 4, 6, etc in the vector y_i^*) from appropriate truncated normal distributions, according to the device of Albert and Chib (1993), marginalized over $\mathbf{b}_i$. In our description of the fitting method given below it is to be understood that $\mathbf{y}_i^*$ contains the most recently simulated values of $\{x_{12it}^*\}_{t=1}^{n_i}$. A new step is the sampling of $(\omega_{11}, \omega_{12})$. The best way of working with these parameters is to reparameterize them to (σ^2, ω_{12}) where $\sigma^2 = \omega_{11} - \omega_{12}^2$ and then assuming that prior information on the transformed parameters is represented by the conditionally conjugate distribution

$$\pi(\sigma^2, \omega_{12}) = \mathcal{IG}\left(\sigma^2|\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)\mathcal{N}\left(\omega_{12}|m_0, \sigma^2 M_0\right) \tag{25}$$

Now conditioned on $\{x_{12it}^*\}_{i,t}$ and $\{\mathbf{b}_i\}$ it follows that

$$\widetilde{y}_{it} = \omega_{12}u_{it} + v_{it} \tag{26}$$

where

$$\widetilde{y}_{it} = y_{it} - \mathbf{x}_{11it}'\beta_{11} - x_{12it}\beta_{12} - \mathbf{w}_{it}'\mathbf{c}_{2i} \ ,$$
$$u_{it} = x_{12it}^* - \mathbf{x}_{11it}'\gamma + \mathbf{w}_{it}'\mathbf{d}_{2i} + z_{it}\delta$$

and

$$v_{it} \sim N(0, \sigma^2)$$

The prior in (25) and the sampling model in (26) when combined by Bayes theorem produce an updated distribution of (σ^2, ω_{12}) that is sampled in one block. To see the details, we express the model in (26) for all $M = \sum_{i=1}^N T_i$ observations as

$$\widetilde{\mathbf{y}} = \omega_{12}\mathbf{u} + \mathbf{v}$$

where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_M)$. By simple calculations it is seen that the updated distribution of σ^2 marginalized over ω_{12} is

$$\mathcal{IG}\left(\sigma^2|\frac{\nu_0 + M}{2}, \frac{\delta_0 + (\widetilde{\mathbf{y}} - \mathbf{u}m_0)'\left(I_M + \mathbf{u}M_0\mathbf{u}'\right)^{-1}(\widetilde{\mathbf{y}} - \mathbf{u}m_0)}{2}\right)$$

while that of ω_{12} conditioned on σ^2 is

$$\mathcal{N}\left(\omega_{12}|, W\sigma^{-2}\left(M_0 m_0 + \mathbf{u}'\widetilde{\mathbf{y}}\right), W = \sigma^2\left(M_0 + \mathbf{u}'\mathbf{u}\right)^{-1}\right)$$

Algorithm: **Gaussian-Gaussian Binary Endogenous Panel**

1. Sample

   (a)

$$x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Omega} \sim\propto \mathcal{N}\left(\mu_{it}, v_{it}\right)\left\{I\left(x^*_{12it} < 0\right)^{1-x_{12it}} + I\left(x_{12it} > 0\right)^{x_{12it}}\right\}$$
$$\mu_{it} = \mathrm{E}(x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Omega})$$
$$v_{it} = \mathrm{Var}(x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Omega})$$

   (b)

$$\boldsymbol{\beta}|\mathbf{y}^*, \boldsymbol{\Omega}, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\boldsymbol{\beta}}, \mathbf{B}\right)$$

   (c)

$$\mathbf{b}_i|\mathbf{y}^*, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Omega} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \leq N$$

2. Sample

$$\mathbf{D}^{-1}|\mathbf{y}^*, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Sample

   (a)

$$\sigma^2|\mathbf{y}^*, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \mathbf{D} \sim$$
$$\mathcal{IG}\left(\sigma^2|\frac{\nu_0 + M}{2}, \frac{\delta_0 + \left(\widetilde{\mathbf{y}} - \mathbf{u}m_0\right)'\left(I_M + \mathbf{u}M_0\mathbf{u}'\right)^{-1}\left(\widetilde{\mathbf{y}} - \mathbf{u}m_0\right)}{2}\right)$$

   (b)

$$\omega_{12}|\mathbf{y}^*, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \sigma^2, \mathbf{D} \sim$$
$$\mathcal{N}\left(\omega_{12}|, W\sigma^{-2}\left(M_0 m_0 + \mathbf{u}'\widetilde{\mathbf{y}}\right), W = \sigma^2\left(M_0 + \mathbf{u}'\mathbf{u}\right)^{-1}\right)$$

4. Goto 1

6 Informative Missingness

It is possible to develop a range of panel data models in which the outcome on a given subject at time t is potentially missing. Each individual at time t supplies two observations: c_{it} and y_{it}. The variable c_{it} is binary and takes the value 1 in which case y_{it} is observed or the value 0 in which case the observation y_{it} is missing. The two random variables are correlated due to the presence of common unobserved random variables. The missingness mechanism is thus non-ignorable. To describe the basic components of such a model, suppose y_{it} is the outcome (which could be continuous, discrete, or censored) and c_{it} is an indicator variable of non-missingness. As an example suppose that the variable c_{it} is one if the individual is working and 0 otherwise and y_{it} is a continuous variable indicating the person's wage. Thus, the variable y_{it} is observed when c_{it} is one; otherwise the variable y_{it} is missing. Let c_{it}^* denote a continuous random variable that is marginally generated as

$$c_{it}^* = \mathbf{x}_{it}'\boldsymbol{\gamma}_i + z_{it}\delta_i + u_i$$

and let

$$c_{it} = I\left(c_{it}^* > 0\right)$$

where γ_i and δ_i are subject-specific coefficients and z_{it} is an additional covariate (the instrument). For simplicity we are assuming that the effect of each covariate is subject-specific although this can be relaxed, much as we have done in the models discussed previously. Also suppose that the outcome y_{it} (under the assumption that it is continuous) is marginally generated as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\alpha}_i + \varepsilon_{it}$$

where

$$\begin{pmatrix} \varepsilon_{it} \\ u_{it} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & 1 \end{pmatrix}\right)$$

To complete the model, we specify the distribution of the heterogenous coefficients with a hierarchical prior. Let $\boldsymbol{\beta}_i = (\alpha_i', \gamma_i', \delta_i)'$ and assume that

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}, \mathbf{D} \sim \mathcal{N}\left(\mathbf{A}_i\boldsymbol{\beta}, \mathbf{D}\right)$$

where $\mathbf{D}$ is a full matrix. Under this latter specification, the two components of the model are tied together not just by correlation amongst the errors but also by the dependence between α_i and (γ_i, δ_i) as measured by the off-diagonal blocks D_{12} and D_{13} of D. It is also assumed that the covariates x_{it} and z_{it} are observable even when y_{it} is missing (ie., when $c_{it} = 0$).

We mention that a variant of this model is considered by *Chib, Seethuraman and Strijnev* [2003]. In that model y_{it} is multinomial indicating choice amongst a given set of brands in a particular category (say cola) and c_{it} is a variable that indicates whether purchase into the category occurs at shopping visit t; if the individual does not purchase in the category then the brand-choice outcome is missing. They describe the Bayesian MCMC fitting of the model and apply the model and the algorithm to a scanner panel data set.

7 Prediction

In some problems one is interested in predicting one or more post-sample observations on a given individual. Specifically, for an individual in the sample, we are interested in making inferences about the set of observations

$$\mathbf{y}_{if} = (y_{in_i+1}, ..., y_{in_i+s})$$

given sample data and a particular hierarchical Bayesian model. In the Bayesian context, the problem of prediction is solved by the calculation of the predictive density

$$f(\mathbf{y}_{if}|\mathbf{y}) = \int p(\mathbf{y}_{if}|\mathbf{y}, \boldsymbol{\delta}_i, \boldsymbol{\theta})\pi(\boldsymbol{\delta}_i, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\delta}_i d\boldsymbol{\theta}$$

where $\boldsymbol{\delta}_i$ denotes the set of cluster-specific unobserved random-variables (such as $\mathbf{z}_i$ in binary and censored response models and the random effects $\mathbf{b}_i$) and $\boldsymbol{\theta}$ denote the entire set of parameters. The predictive density is the density of $\mathbf{y}_{if}$ marginalized over $(\boldsymbol{\delta}_i, \boldsymbol{\theta})$ with respect to the posterior distribution of $(\boldsymbol{\delta}_i, \boldsymbol{\theta})$.

This predictive density is summarized in the same way that we summarized the posterior density of the parameters - by sampling it. Sampling of the predictive density is conducted by the *method of composition*. According to the method of composition, if $f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$, and $\mathbf{x}^{(g)}$ is a draw from $\pi(\mathbf{x})$, then $\mathbf{y}^{(g)}$ drawn from $f(\mathbf{y}|\mathbf{x}^{(g)})$ is a draw from $f(\mathbf{y})$. Thus, a draw from the marginal is obtained simply by sampling the conditional density $f(\mathbf{y}|\mathbf{x})$ for each value drawn from $\pi(\mathbf{x})$.

The method of composition leads to an easily implementable procedure for calculating the predictive density in every panel data model that we have considered. For example in the Gaussian-Gaussian model, given ($\boldsymbol{\beta}^{(g)}$, $\sigma^{2(g)}, \mathbf{b}_i^{(g)}$), the gth MCMC draw on ($\boldsymbol{\beta}$, $\sigma^2, \mathbf{b}_i$), the gth draw from the

predictive density is obtained by drawing

$$\varepsilon_{it}^{(g)} \sim \mathcal{N}(0, \sigma^{2(g)}) \ , \ t = n_i + 1, ..., n_i + s$$

and setting

$$y_{it}^{(g)} = \mathbf{x}_{it}'\boldsymbol{\beta}^{(g)} + \mathbf{w}_{it}'\mathbf{b}_i^{(g)} + \varepsilon_{it}^{(g)} \ , \ t = n_i + 1, ..., n_i + s$$

The resulting sample of draws are summarized in terms of moments, quantiles and density plots.

8 Residual Analysis

One approach to Bayesian residual analysis relies on the idea of "realized errors" introduced by *Zellner* [1975] and studied more recently by *Chaloner and Brant* [1988] and *Albert and Chib* [1995]. The idea is to compute the posterior distribution of the error and define a residual to be outlying if the posterior distribution is concentrated on large values.

Consider for simplicity the Gaussian-Gaussian model for continuous responses. In that case, the error conditioned on y_{it} is given by

$$\varepsilon_{it} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \mathbf{w}_{it}'\mathbf{b}_i$$

and, therefore, the posterior distribution of ε_{it} is determined by the posterior distribution of $\boldsymbol{\beta}$ and $\mathbf{b}_i$. To obtain this posterior distribution, at each iteration of the sampling, we compute the value

$$\varepsilon_{it}^{(g)} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}^{(g)} - \mathbf{w}_{it}'\mathbf{b}_i^{(g)}$$

where $\{\boldsymbol{\beta}^{(g)}, \mathbf{b}_i^{(g)}\}$ are the gth sampled values. Then, the collection of values $\{\varepsilon_{it}^{(g)}\}$ constitutes a sample from the posterior distribution $\pi(\varepsilon_{it}|\mathbf{y})$. There are various ways to summarize this posterior distribution in order to find outlying observations. One possibility is to compute the posterior probability

$$\Pr\left(\left|\frac{\varepsilon_{it}}{\sigma}\right| > k|\mathbf{y}\right)$$

where k is 2 or 3, and compare the posterior probability (computed from the simulated draws $\varepsilon_{it}^{(g)}/\sigma^{(g)}$) with the prior probability that the standardized residual is bigger that k in absolute value. The observation is classified as on outlier if the ratio of the posterior probability to the prior probability is large. Interestingly, similar ideas are used in panel probit models as discussed by *Albert and Chib* [1995].

9 Model Comparisons

Posterior simulation by MCMC methods does not require knowledge of the normalizing constant of the posterior density. Nonetheless, if we are interested in comparing alternative models, then knowledge of the normalizing constant is essential. This is because the standard and 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

$$m(\mathbf{y}|\mathcal{M}) = \int p(\mathbf{y}|\mathcal{M}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}\,, \tag{27}$$

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

$$\mathrm{B}_{kl} = \frac{m(\mathbf{y}|\mathcal{M}_k)}{m(\mathbf{y}|\mathcal{M}_l)}\,. \tag{28}$$

Computation of the marginal likelihood is, therefore, of some importance in Bayesian statistics (*DiCicio, Kass, Raftery and Wasserman* [1997], *Chen and Shao* [1997], *Roberts* [2001]). 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 $m(\mathbf{y})$ by virtue of being the normalizing constant of the posterior density can be expressed as

$$m(\mathbf{y}|\mathcal{M}) = \frac{p(\mathbf{y}|\mathcal{M}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathcal{M})}{\pi(\boldsymbol{\theta}^*|\mathcal{M}, \mathbf{y})}\,, \tag{29}$$

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

$$\log m(\mathbf{y}|\mathcal{M}) = \log p(\mathbf{y}|\mathcal{M}, \boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*|\mathcal{M}) - \log \hat{\pi}(\boldsymbol{\theta}^*|\mathcal{M}, \mathbf{y})\,. \tag{30}$$

In the context of both single and multiple block M-H chains, good estimates of the posterior ordinate are available. For example, when the MCMC simulation is run with B blocks, to estimate the posterior ordinate we employ

the marginal-conditional decomposition

$$\pi(\boldsymbol{\theta}^*|\mathcal{M},\mathbf{y}) = \pi(\boldsymbol{\theta}_1^*|\mathcal{M},\mathbf{y}) \times ... \times \pi(\boldsymbol{\theta}_i^*|\mathcal{M},\mathbf{y},\boldsymbol{\psi}_{i-1}^*) \times ... \times \pi(\boldsymbol{\theta}_B^*|\mathcal{M},\mathbf{y},\boldsymbol{\psi}_{B-1}^*)\ , \tag{31}$$

where on letting $\boldsymbol{\psi}_i = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_i)$ and $\boldsymbol{\psi}^i = (\boldsymbol{\theta}_i, ..., \boldsymbol{\theta}_B)$ denote the list of blocks upto i and the set of blocks from i to B, respectively, and $\mathbf{z}$ denoting the latent data, and dropping the model index for notational convenience, the typical term is of the form

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y},\boldsymbol{\psi}_{i-1}^*) = \int \pi(\boldsymbol{\theta}_i^*|\mathbf{y},\ \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1}, \mathbf{z})\, \pi(\boldsymbol{\psi}^{i+1}, \mathbf{z}|\ \mathbf{y}, \boldsymbol{\psi}_{i-1}^*) d\boldsymbol{\psi}^{i+1} d\mathbf{z}$$

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 $(\boldsymbol{\psi}^{i+1}, \mathbf{z})$ and that the integrating measure is conditioned on $\boldsymbol{\psi}_{i-1}^*$.

Consider first the case where the normalizing constant of each full conditional density is known. Then, the first term of (31) is estimated by the Rao-Blackwell method. To estimate the typical reduced conditional ordinate, one conducts a MCMC run consisting of the full conditional distributions

$$\left\{\pi(\boldsymbol{\theta}_i|\mathbf{y},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1},\mathbf{z})\,;\ ...;\ \boldsymbol{\pi}(\boldsymbol{\theta}_B|\mathbf{y},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\theta}_i,...,\boldsymbol{\theta}_{B-1},\mathbf{z})\ ;\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^i)\right\} \tag{32}$$

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

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

over the simulated values of $\boldsymbol{\psi}^{i+1}$ and $\mathbf{z}$ from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (31) 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

$$\log \hat{m}(\mathbf{y}|\mathcal{M}) = \log p(\mathbf{y}|\mathcal{M},\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*|\mathcal{M}) - \sum_{i=1}^{B} \log \hat{\pi}(\ \boldsymbol{\theta}_i^*|\mathcal{M},\mathbf{y},\boldsymbol{\psi}_{i-1}^*) \tag{33}$$

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

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

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

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ and the denominator expectation E_2 is with respect to the proposal density of $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\theta}^*$, $q(\boldsymbol{\theta}^*,\ \boldsymbol{\theta}|\mathbf{y})$, and $\alpha(\boldsymbol{\theta},\ \boldsymbol{\theta}^*|\mathbf{y})$ is the probability of move in the M-H step. This leads to the simulation consistent estimate

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^*|\mathbf{y})q(\ \boldsymbol{\theta}^{(g)}, \boldsymbol{\theta}^*|\mathbf{y})}{J^{-1}\sum_{j=1}^{M}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}|\mathbf{y})}\ , \tag{34}$$

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

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

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{i-1}^*) = \frac{E_1\left\{\alpha(\ \boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})q_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\right\}} \tag{35}$$

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

9.1 Gaussian-Gaussian model

As an example of the calculation of the marginal likelihood consider the calculation of the posterior ordinate for the Gaussian-Gaussian continuous response model. The ordinate is written as

$$\pi(\mathbf{D}^{-1*}, \sigma^{2*}, \boldsymbol{\beta}^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\sigma^{2*}|\mathbf{y}, \mathbf{D}^*)\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*, \sigma^{2*})\ ,$$

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

densities

$$\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{D}^*, \sigma^2)\,;\, \pi(\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}^*, \{\mathbf{b}_i\})\,;\pi(\{\mathbf{b}_i\}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}^*, \sigma^2)\ ,$$

where each conditional utilizes the fixed value of $\mathbf{D}$. The second ordinate is now estimated by averaging the inverse gamma full conditional density of σ^2 at σ^{2*} over the draws on $(\boldsymbol{\beta}, \{\mathbf{b}_i\})$ from this reduced run. The third ordinate is multivariate normal as given above and available directly.

9.2 Gaussian-Gaussian Tobit model

As another example, consider the Gaussian-Gaussian Tobit censored regression model. The likelihood ordinate is not available directly but can be estimated by a simulation-based approach. For the posterior ordinate we again utilize the decomposition

$$\pi(\mathbf{D}^{-1*}, \sigma^{2*}, \boldsymbol{\beta}^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\sigma^{2*}|\mathbf{y}, \mathbf{D}^*)\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*, \sigma^{2*})\ ,$$

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

$$\begin{aligned}&\pi(\boldsymbol{\beta}|\mathbf{y}_z, \mathbf{D}^*, \sigma^2)\,;\ \pi(\{\mathbf{z}_i\}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}^*, \sigma^2);\\ &\pi(\sigma^2|\mathbf{y}_z, \boldsymbol{\beta}, \mathbf{D}^*, \{\mathbf{b}_i\})\,;\ \pi(\{\mathbf{b}_i\}|\mathbf{y}_z, \boldsymbol{\beta}, \mathbf{D}^*, \sigma^2)\ ,\end{aligned}$$

and estimate the second ordinate by averaging the inverse gamma full conditional density of σ^2 at σ^{2*} over the draws on $(\boldsymbol{\beta}, \{\mathbf{z}_i\}, \{\mathbf{b}_i\})$ from this run. Finally, to estimate the last ordinate we also fix σ^2 at σ^{2*} and continue the reduced runs with the full-conditional densities

$$\pi(\boldsymbol{\beta}|\mathbf{y}_z, \mathbf{D}^*, \sigma^{2*})\,;\ \ \pi(\{\mathbf{z}_i\}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}^*, \sigma^{2*});\ \pi(\{\mathbf{b}_i\}|\mathbf{y}_z, \boldsymbol{\beta}, \mathbf{D}^*, \sigma^{2*})\ ,$$

and average the multivariate normal density given in Step 1 of the MCMC algorithm at the point $\boldsymbol{\beta}^*$.

9.3 Panel Poisson model

As a last example of the calculation of the marginal likelihood, consider the panel poisson model in which the full conditional of $\boldsymbol{\beta}$ is not of known form. Now the posterior ordinate given the sampling scheme in the Panel count algorithm is decomposed as

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

where the first ordinate is found by averaging the Wishart density over draws on $\{\mathbf{b}_i\}$ from the full run. The second ordinate is found by the method of *Chib and Jeliazkov* [2001] as

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*) = \frac{M^{-1}\sum_{g=1}^{M} \alpha(\boldsymbol{\beta}^{(g)}, \boldsymbol{\beta}^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\}) q(\boldsymbol{\beta}^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\})}{J^{-1}\sum_{j=1}^{J} \alpha(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(j)}|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})}$$

where the draws in the numerator are from a reduced run comprising the full conditional distributions of $\boldsymbol{\beta}$ and $\{\mathbf{b}_i\}$, conditioned on $\mathbf{D}^*$ whereas the draws in the denominator are from a second reduced run comprising the full conditional distributions of $\{\mathbf{b}_i\}$, conditioned on $(\mathbf{D}^*, \boldsymbol{\beta}^*)$ with an appended step in which $\boldsymbol{\beta}^{(j)}$ is drawn from $q(\boldsymbol{\beta}|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})$. The log of the likelihood ordinate $p(\mathbf{y}|\boldsymbol{\beta}^*, \mathbf{D}^*)$ is found by importance sampling.

10 Conclusion

In this chapter we have illustrated how Bayesian methods provide a complete inferential tool-kit for a variety of panel data models. The methods are based on a combination of hierarchical prior modeling and MCMC simulation methods. Interestingly, the approaches are able to tackle estimation and model comparison questions in situations that are quite challenging by other means. We discussed applications to models for continuous, binary, censored, count, multinomial response models under various realistic and robust distributional and modeling assumptions. The methods are quite practical and straightforward, even in complex models settings such as those with binary and count responses, and enable the calculation of the entire posterior distribution of the unknowns in the models. The algorithm for fitting panel probit models with random effects is particularly interesting in that it highlights the value of augmentation in simplifying the simulations and in circumventing the calculation of the likelihood function. Procedures for dealing with missing data, predicting future outcomes and for detecting outliers have also been discussed.

The methods discussed in this chapter, which have arisen in the course of a revolutionary growth in Bayesian statistics in the last decade, offer a unified approach for analyzing a whole array of panel models. The pace of growth of Bayesian methods for longitudinal data continues unimpeded as the Bayesian approach attracts greater interest and adherents.

11 References

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

Albert, J. and S. Chib (1995), Bayesian residual analysis for binary response models, *Biometrika*, 82, 747-759.

Basu, S. and Chib, S. (2003), Marginal likelihood and Bayes Factors for Dirichlet process mixture models, *Journal of the American Statistical Association*, (2003), 98, 224-235.

Beasley, J. D. and S. G. Springer (1977), Algorithm 111, *Applied Statistics*, 26, 118–121.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society*, B, 36, 192-236.

Best, D. J. (1978), Letter to the Editor, *Applied Statistics*, 29, 181.

Butler, J.S. and R. Moffitt (1982), A computationally efficient quadrature procedure for the one factor multinomial probit model, *Econometrica*, 50, 761-764.

Carlin, B. P. and T. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, New York.

Chaloner, K. and Brant, R. (1988), A Bayesian approach to outlier detection and residual analysis, *Biometrika*, 75, 651-659.

Chamberlain, G. (1980), Analysis of covariance with qualitative data, *Review of Economic Studies*, 47, 225-238.

Chen Z. and Dunson, D.B. (2003), Random effects selection in linear mixed models, Biometrics, 59, 762-769.

Chen, M-H and Shao, Q-M (1998), On Monte Carlo methods for estimating ratios of normalizing constants, *Annals of Statistics*, 25, 1563-1594.

Chiang, J., Chib, S. and Narasimhan, C. (1999), Markov Chain Monte Carlo and models of consideration set and parameter heterogeneity, *Journal of Econometrics*, 89, 223-248.

Chib, S. (1992), Bayes regression for the Tobit censored regression model, *Journal of Econometrics*, 51, 79–99.

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

Chib, S. (2001), Markov Chain Monte Carlo methods: Computation and inference, in *Handbook of Econometrics* volume 5 (eds J.J. Heckman and E. Leamer), North Holland, Amsterdam, 3569-3649.

Chib, S. and E. Greenberg (1994), Bayes inference for regression models with ARMA(p, q) errors, *Journal of Econometrics*, 64, 183-206.

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

Chib, S. and E. Greenberg (1996), Markov chain Monte Carlo simulation methods in econometrics, *Econometric Theory*, 12, (1996), 409-431.

Chib, S. and Greenberg, E. and Winklemnann , R. (1998), Posterior simulation and Bayes factors in panel count data models, *Journal of Econometrics*, 86, 33-54.

Chib, S. and B.P. Carlin (1999), On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing*, 9, 17-26.

Chib, S. and I. Jeliazkov (2001), Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association*, 96, 270-281.

Chib, S. and Hamilton, B. (2002), Semiparametric Bayes analysis of longitudinal data treatment models, *Journal of Econometrics*, 110, 67-89.

Chib, S., Seetharaman, P.B. and Strijnev, A., (2003) Model of brand choice with a no-purchase option calibrated to scanner panel data, *Journal of Marketing Research*, in press.

Congdon, P. (2001) *Bayesian Statistical Modelling*, John Wiley & Sons, Chichester.

DiCiccio, T.J., Kass, R.E., Raftery, A.E., and Wasserman, L. (1997), Computing Bayes factors by combining simulation and asymptotic approximations, *Journal of the American Statistical Association*, 92, 903-915.

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

Gelfand, A. E., Sahu, S.K. and Carlin, B.P. (1995), Efficient parameterizations for normal linear mixed models, Biometrika, 82, 479–488.

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

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics,* (eds., J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), New York: Oxford University Press, 169–193.

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

Heckman, J.J. (1981), Statistical models for discrete panel data, in *Structural Analysis of Discrete Data with Econometric Applications*, ed C. F. Manski and D. McFadden, pp 114-178, Cambridge: MIT Press.

Lindley, D.V. and Smith, A.F.M. (1972), Bayes estimates for the linear model, *Journal of the Royal Statistical Society* Ser. B, 34, 1-41.

Liu, J. S., W. W. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika* 81, 27–40.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, 21, 1087–1092.

Page, E (1977), Approximations to the cumulative normal function and its inverse for use on a pocket calculator, *Applied Statistics*, 26, 75-76.

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

Ripley, B. (1987), *Stochastic simulation*, New York: John Wiley & Sons.

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

STOUT, W. F. (1974), *Almost Sure Convergence*, New York, Academic Press.

TANNER, M. A. and W. H. WONG (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association,* 82, 528–549.

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

WAKEFIELD, J. C., A. F. M. SMITH, A. Racine POON, and A. E. GELFAND (1994), Bayesian analysis of linear and non-linear population models by using the Gibbs sampler, *Applied Statistics*, 43, 201-221.

ZELLNER, A (1975), Bayesian analysis of regression error terms, *Journal of the American Statistical Association*, 70, 138-144.