MCMC Perspectives on Simulated Likelihood Estimation

Ivan Jeliazkov Esther Hee Lee
University of California, Irvine*

June 2010

Abstract

A major stumbling block in multivariate discrete data analysis is the problem of evaluating the outcome probabilities that enter the likelihood function. Calculation of these probabilities involves high-dimensional integration, making simulation methods indispensable in both Bayesian and frequentist estimation and model choice. We review several existing probability estimators and then show that a broader perspective on the simulation problem can be afforded by interpreting the outcome probabilities through Bayes' theorem, leading to the recognition that estimation can alternatively be handled by methods for marginal likelihood computation based on the output of Markov chain Monte Carlo (MCMC) algorithms. These techniques offer standalone approaches to simulated likelihood estimation, but can also be integrated with traditional estimators. Building on both branches in the literature, we develop new methods for estimating response probabilities and propose an adaptive sampler for producing high-quality draws from multivariate truncated normal distributions. A simulation study illustrates the practical benefits and costs associated with each approach. The methods are employed to estimate the likelihood function of a correlated random effects panel data model of women's labor force participation.

Keywords: Adaptive sampling; Discrete responses; Markov chain Monte Carlo (MCMC); Multivariate integration; Multivariate probit model; Random effects panel data model.

1 Introduction

Limited dependent variable models deal with binary, multivariate, multinomial, ordinal or censored outcomes that can arise in cross-sectional, time-series, or longitudinal (panel data) settings. To enable inference in this class of models, however, one must address a central problem in multivariate discrete data analysis, namely, evaluation of the outcome probability for each observation. Outcome probabilities are required in constructing the likelihood function and involve multivariate integration

*Department of Economics, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100. E-mail addresses: ivan@uci.edu and hwlee@uci.edu. We thank the editors and two anonymous referees for their detailed comments and helpful suggestions.

constrained to specific regions that correspond to the observed data. To illustrate the main ideas in some detail, consider the latent variable representation

$$\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}), \tag{1}$$

where, for $i = 1, \ldots, n$, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iJ})'$ is a vector of continuous latent variables underlying the discrete observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$, $\boldsymbol{X}_i$ is a $J \times k$ matrix of covariates with corresponding k-vector of parameters $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$ is a $J \times J$ covariance matrix in which the variances of any binary or ordinal variables y_{ij} are typically set to 1 for identification reasons. This latent variable framework is a general probabilistic construct in which different threshold-crossing mappings from $\boldsymbol{z}_i$ to the observed responses $\boldsymbol{y}_i$ can produce various classes of discrete data models such as the multivariate probit for binary and ordinal data, multinomial probit, panels of binary, ordinal, or censored (Tobit) outcomes, models with incidental truncation or endogenous treatment indicators, and Gaussian copula models. For example, the indicator function mapping $y_{ij} = 1\{z_{ij} > 0\}$ underlies binary data models, the relationship $y_{ij} = 1\{z_{ij} > 0\} z_{ij}$ leads to a Tobit model with censoring from below at 0, the discretization $y_{ij} = \sum_{s=1}^{S} 1\{z_{ij} > \gamma_{j,s}\}$ for some strictly increasing sequence of cutpoint parameters $\{\gamma_{j,s}\}_{s=1}^{S}$ arises in ordinal data modeling and copula models for count data, and so on. Variations on the distributional assumptions can be used to construct mixtures or scale mixtures of normals models including the Student's t-link ("robit") and logit models. In economics the latent $\boldsymbol{z}_i$ are interpreted as unobserved utility differences (relative to a baseline category), and discrete data models are often referred to as discrete choice models.

A representative example that will form the basis for discussion in the remainder of this paper is the multivariate probit model where the binary outcomes in $\boldsymbol{y}_i$ relate to the latent $\boldsymbol{z}_i$ in (1) through the indicator functions $y_{ij} = 1\{z_{ij} > 0\}$ for $j = 1, \ldots, J$. In this context the object of interest is the probability of observing $\boldsymbol{y}_i$, conditionally on $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$, which is given by

$$\begin{aligned} \Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) &= \int_{\mathcal{B}_{iJ}} \cdots \int_{\mathcal{B}_{i1}} f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})\, dz_{i1} \cdots dz_{iJ} \\ &= \int 1\{\boldsymbol{z}_i \in \mathcal{B}_i\} f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})\, d\boldsymbol{z}_i \end{aligned} \tag{2}$$

where $f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$ is the normal density with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Omega}$ (which is in correlation form), and the region of integration is given by $\mathcal{B}_i = \mathcal{B}_{i1} \times \mathcal{B}_{i2} \times \cdots \times \mathcal{B}_{iJ}$ with

$$\mathcal{B}_{ij} = \begin{cases} (-\infty, 0] & \text{if} \quad y_{ij} = 0 \\ (0, \infty) & \text{if} \quad y_{ij} = 1 \end{cases} .$$

The log-likelihood function is given by $\ln f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Omega}) = \sum_{i=1}^{n} \ln \Pr(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega})$, however, a major stumbling block in evaluating that function is that the multivariate integrals defining the likelihood contributions in (2) typically have no closed-form solution, but typically need to be evaluated at various values of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ for the purposes of estimation (e.g. in maximization algorithms) and model comparison (e.g. in evaluating likelihood ratio statistics, information criteria, Bayes factors and marginal likelihoods). Standard grid-based numerical approximations (e.g. Gauss-Legendre or quadrature methods) exist for univariate and bivariate problems, but the computational costs associated with these approaches rise exponentially with dimensionality, which makes them prohibitively expensive in higher dimensions. While in many instances the computational intensity of numerical integration can be moderated by sparse-grid approximations as in Heiss and Winschel (2008), the most widely-used approaches for obtaining (2) in discrete data analysis have been based on simulation. Such methods exploit a number of practical advantages that make them particularly appealing. For example, simulation methods typically rely on standard distributions which makes them conceptually and computationally straightforward and efficient, even in high dimensions. Moreover, simulation often resolves the problem of having to specify the location and size of a grid so that it corresponds to areas of high density. This is especially useful because knowledge of these features is often absent, especially in high-dimensional problems. For these reasons, simulation methods have become a fundamental tool in multivariate integration in general, and in simulated likelihood estimation in particular.

One popular approach for simulation-based evaluation of the outcome probabilities in discrete choice models is the Geweke, Hajivassiliou and Keane (GHK) method (Geweke, 1991; Börsch-Supan and Hajivassiliou, 1993; Keane, 1994; Hajivassiliou and McFadden, 1998). Another one is

studied by Stern (1992). These methods have risen to prominence because they are efficient and offer continuous and differentiable choice probabilities that are strictly bounded between 0 and 1, making them very suitable for maximum likelihood estimation and other problems that require gradient or Hessian evaluation. Other methods, such as the accept-reject (AR) simulator and its variants, are appealing because of their transparency and simplicity. Many of these techniques, together with other useful alternatives, have been carefully reviewed in Hajivassiliou and Ruud (1994), Stern (1997), and Train (2003).

In this paper we pursue several objectives. Our first main goal is to show that the probability of the observed response, given the model parameters, can be estimated consistently and very efficiently by a set of alternative techniques that have been applied in a very different context. In particular, the calculation of integrals which have no closed-form solution has been a central issue in Bayesian model comparison. The marginal likelihood, which is given by the integral of the likelihood function with respect to the prior distribution of the model parameters, is an important ingredient in producing Bayes factors and posterior odds of competing models. A large number of Markov chain Monte Carlo (MCMC) methods have been introduced to calculate marginal likelihoods, Bayes factors, and posterior odds (e.g. Ritter and Tanner, 1992; Newton and Raftery, 1994; Gelfand and Dey, 1994; Chib, 1995; Meng and Wong, 1996; DiCiccio et al., 1997; Geweke, 1999; Chib and Jeliazkov, 2001, 2005), but these methods have not yet been employed to estimate response probabilities and construct likelihood functions for discrete data models even though MCMC data augmentation techniques have been routinely used to obtain parameter estimates without computing those probabilities (see, e.g., Koop, 2003, Greenberg, 2008, and the references therein). A recent comparison of Bayesian and classical inferences in probit models is offered in Griffiths et al. (2006). Given the specifics of the current context, in this article we focus on MCMC estimation techniques that embody desirable characteristics such as continuity and differentiability, but mention that the other approaches can be very useful as well. Second, we design several new estimation methods by integrating two branches of the literature and combining features of the

classical and Bayesian methods. This allows for several enhancements in the resulting "hybrid" approaches that tend to improve the quality of the simulated latent data sample, the efficiency of the resulting estimates, and retain simplicity without sacrificing continuity and differentiability. Our third goal is to provide a comparison and document the performance of the alternative methods in a detailed simulation study that highlights the practical costs and benefits associated with each approach. Finally, we present an application to the problem of estimating the likelihood ordinate for a correlated random effects panel data model of women's labor force participation, which illustrates the applicability of the proposed techniques.

The rest of this paper is organized as follows. In Section 2, we review several traditional simulation methods that have been used to estimate the response probabilities in simulated likelihood estimation. A number of alternative MCMC approaches are discussed in Section 3. Building on existing work, we introduce new approaches for estimating outcome probabilities that are obtained by integrating features of the Bayesian and traditional techniques. Section 4 provides evidence on the relative performance of these simulation methods, while Section 5 applies the techniques to evaluate the likelihood function of a correlated random effects panel data model using data on women's labor force participation. Concluding remarks are presented in Section 6.

2 Existing Methods

We begin with a brief review of the basic idea behind the accept-reject (AR), or frequency, method which is perhaps the most straightforward approach for estimating the probability in (2). The AR method draws independent identically distributed (iid) random variables $\boldsymbol{z}_i^{(g)} \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$ for $g = 1, \ldots, G$. Draws that satisfy $\boldsymbol{z}_i^{(g)} \in \mathcal{B}_i$ are accepted, whereas those that do not are rejected. The probability in equation (2) is then calculated as the proportion of accepted draws

$$\widehat{\Pr}(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) = G^{-1} \sum_{g=1}^{G} 1\{\boldsymbol{z}_i^{(g)} \in \mathcal{B}_i\}. \tag{3}$$

The AR approach is very simple and intuitive, and is easy to implement with a variety of distributions for the random terms. This estimator has been applied in discrete choice problems by Lerman

and Manski (1981); additional discussion and applications of AR methods are offered in Devroye (1986) and Ripley (1987).

However, for a given finite number of draws, the AR approach has a number of pitfalls especially when used in the context of likelihood estimation. One is that the estimated probability is not strictly bounded between 0 and 1 and there is a positive probability of obtaining an estimate on the boundary, which can cause numerical problems when taking the logarithm of the estimated probability. The more important problem with the AR method, however, is the lack of differentiability of the estimated probability with respect to the parameter vector. Because the AR probability in equation (3) has the form of a step function with respect to the parameters, the simulated probability is either constant or jumps by a discrete amount with respect to a small change in the parameter values. These features of the estimator impede its use in numerical optimization and complicate the asymptotics of estimators that rely on it.

The difficulties of the AR method can be circumvented by replacing the indicator function $1\{\boldsymbol{z}_i \in \mathcal{B}_i\}$ in equation (2) with a smooth and strictly positive function. One strategy, suggested in McFadden (1989), is to approximate the orthant probability as

$$\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right) \approx \int K\left(\frac{\boldsymbol{z}_i}{b}\right) f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i\,, \tag{4}$$

where $K(\cdot)$ is a smooth kernel function, for example the logistic cdf, and $b \neq 0$ is a scale factor that determines the degree of smoothing. It can easily be seen that the function $K(\boldsymbol{z}_i/b)$ approaches $\mathbf{1}\{\boldsymbol{z}_i \in \mathcal{B}_i\}$ as $b \to 0$. This approach avoids the problem of non-differentiability, however, it comes at the cost of introducing a bias in the estimate of the probability (Hajivassiliou et al., 1996). Whereas the bias can be reduced by picking a value of b that is very close to 0, doing so can potentially revive the problem of non-differentiability if $K(\cdot)$ begins to approximate the indicator function $1\{\cdot\}$ too closely. In practice, therefore, the choice of b is not straightforward and must be done very carefully. Other complications, e.g. appropriate kernel selection, could arise with this approach when $\mathcal{B}_i$ is bounded from both below and above as in ordinal probit and copula models.

Another strategy for overcoming the difficulties of estimating (2) was developed by Stern (1992) and relies on a particular decomposition of the correlation structure in (1). The basic idea underlying the Stern method is to decompose the error component $\boldsymbol{\varepsilon}_i$ in (1) into the sum of two terms – one that is correlated and another one that contains orthogonal errors. In particular, the Stern simulator is based on re-writing the model as

$$\boldsymbol{z}_i = \boldsymbol{v}_i + \boldsymbol{w}_i,$$

where $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega} - \boldsymbol{\Lambda})$ and $\boldsymbol{w}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \lambda \boldsymbol{I}$. Note that the mean $\boldsymbol{X}_i\boldsymbol{\beta}$ can be incorporated either in $\boldsymbol{v}_i$ or $\boldsymbol{w}_i$, or in the limits of integration (these representations are equivalent). Moreover, as a matter of simulation efficiency, Stern (1992) suggests that λ should be chosen as large as possible subject to leaving $(\boldsymbol{\Omega} - \boldsymbol{\Lambda})$ positive definite. This is done by setting λ close to the smallest eigenvalue of $\boldsymbol{\Omega}$.

With this decomposition, the likelihood contribution in (2) can be rewritten as

$$\begin{aligned}
\Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) &= \int_{\mathcal{B}_i} f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})\, d\boldsymbol{z}_i \\
&= \int_{\mathcal{C}_i} \int_{-\infty}^{+\infty} f_N(\boldsymbol{w}_i|\mathbf{0}, \boldsymbol{\Lambda})\, f_N(\boldsymbol{v}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega} - \boldsymbol{\Lambda})\, d\boldsymbol{v}_i d\boldsymbol{w}_i
\end{aligned}$$

where the change of variable implies that $\mathcal{C}_i = \mathcal{C}_{i1} \times \cdots \times \mathcal{C}_{iJ}$ with $\mathcal{C}_{ij} = (-\infty, -v_{ij})$ if $y_{ij} = 0$ and $\mathcal{C}_{ij} = [-v_{ij}, \infty)$ if $y_{ij} = 1$. Because the independent elements of $\boldsymbol{w}_i$ have a Gaussian density, which is symmetric, this probability can be expressed as

$$\Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) = \int \left[\prod_{j=1}^{J} \Phi\left(\frac{(-1)^{1-y_{ij}} v_{ij}}{\sqrt{\lambda}} \right) \right] f_N(\boldsymbol{v}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega} - \boldsymbol{\Lambda})\, d\boldsymbol{v}_i$$

where $\Phi(\cdot)$ denotes the standard normal cdf. Estimation of this integral then proceeds by

$$\widehat{\Pr}(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) = \frac{1}{G} \sum_{g=1}^{G} \left\{ \prod_{j=1}^{J} \Phi\left(\frac{(-1)^{1-y_{ij}} v_{ij}^{(g)}}{\sqrt{\lambda}} \right) \right\}$$

where $\boldsymbol{v}_i^{(g)} \sim \mathcal{N}(\boldsymbol{v}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega} - \boldsymbol{\Lambda})$ for $g = 1, \ldots, G$.

Another popular method is the GHK algorithm which builds upon simulation techniques for multivariate truncated normal distributions that were pioneered by Geweke (1991) and has been

successfully implemented in a variety of problems in cross-sectional, time-series, and panel data settings. The GHK algorithm has been extensively studied in Börsch-Supan and Hajivassiliou (1993), Hajivassiliou and Ruud (1994), Keane (1994), Hajivassiliou et al. (1996), and Hajivassiliou and McFadden (1998), and has been carefully reviewed in Train (2003).

The insight behind the GHK algorithm is that one can design a tractable importance density that could facilitate simulation-based estimation by writing the model as

$$\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{5}$$

where $\boldsymbol{L}$ is a lower triangular Cholesky factor of $\boldsymbol{\Omega}$ with elements l_{ij} such that $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Omega}$. Because the entries in $\boldsymbol{\eta}_i$ are independent and $\boldsymbol{L}$ is lower triangular, a recursive relation between the elements of $\boldsymbol{z}_i$ can be established to produce the importance density used in the GHK algorithm

$$\begin{aligned} h\left(\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) &= h\left(z_{i1}|y_{i1}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) h\left(z_{i2}|z_{i1}, y_{i2}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) \cdots h\left(z_{iJ}|z_{i1}, \ldots, z_{i,J-1}, y_{iJ}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) \\ &= \prod_{j=1}^{J} h\left(z_{ij}|\{z_{ik}\}_{k<j}, y_{ij}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) \end{aligned} \tag{6}$$

and the terms in the product are restricted to the set $\mathcal{B}_i$ by letting

$$\begin{aligned} h\left(z_{ij}|\{z_{ik}\}_{k<j}, y_{ij}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) &= f_{TN_{\mathcal{B}_{ij}}}\left(z_{ij}|\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{j-1} l_{jk}\eta_{ik}, l_{jj}^2\right) \\ &= 1\left\{z_{ij} \in \mathcal{B}_{ij}\right\} f_N\left(z_{ij}|\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{j-1} l_{jk}\eta_{ik}, l_{jj}^2\right)/c_{ij}, \end{aligned}$$

where $c_{ij} = \Phi\left((-1)^{(1-y_{ij})}\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{j-1} l_{jk}\eta_{ik}\right)/l_{jj}\right)$ is the normalizing constant of the truncated normal density $f_{TN_{\mathcal{B}_{ij}}}\left(z_{ij}|\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{j-1} l_{jk}\eta_{ik}, l_{jj}^2\right)$. As a result, taking the product in (6) produces

$$\begin{aligned} h\left(\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) &= \frac{\prod_{j=1}^{J} 1\left\{z_{ij} \in \mathcal{B}_{ij}\right\} f_N\left(z_{ij}|\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{j-1} l_{jk}\eta_{ik}, l_{jj}^2\right)}{\prod_{j=1}^{J} c_{ij}} \\ &= \frac{1\left\{\boldsymbol{z}_i \in \mathcal{B}_i\right\} f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}\right)}{\prod_{j=1}^{J} c_{ij}}, \end{aligned}$$

upon which one could write (1) as

$$\begin{aligned}
\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right) &= \int_{\mathcal{B}_i} f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i \\
&= \int_{\mathcal{B}_i} \frac{f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)}{h\left(\boldsymbol{z}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)} h\left(\boldsymbol{z}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i \\
&= \int_{\mathcal{B}_i} \frac{f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)}{f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)/\prod_{j=1}^{J} c_{ij}} h\left(\boldsymbol{z}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i \\
&= \int_{\mathcal{B}_i} \left\{\prod_{j=1}^{J} c_{ij}\right\} h\left(\boldsymbol{z}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i\,.
\end{aligned} \tag{7}$$

Therefore, $\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ can be estimated as

$$\widehat{\Pr}\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right) = \frac{1}{G}\sum_{g=1}^{G}\prod_{j=1}^{J} c_{ij}^{(g)}$$

with draws $\boldsymbol{z}_i^{(g)}$ obtained recursively as $z_{ij}^{(g)} \sim h\left(z_{ij}|\{z_{ik}^{(g)}\}_{k<j}, y_{ij}, \boldsymbol{\beta},\boldsymbol{\Omega}\right)$ for $j = 1,\ldots,J-1$, and $g = 1,\ldots,G$, using techniques such as the inverse cdf method (see, e.g., Devroye, 1986) or simulation-based techniques such as those proposed in Robert (1995).

Both the Stern and GHK methods provide continuous and differentiable multivariate probability estimates. They also typically produce smaller estimation variability than the AR method because the simulated probabilities are strictly bounded between $(0,1)$, whereas each draw in the AR method gives either 0 or 1. However, all three methods suffer from a common problem that can often produce difficulties. In particular, in all three approaches, the simulation draws come from proposal distributions that differ from the truncated normal distribution of interest, $\mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$. When this disparity is large, the efficiency of all methods can be adversely affected. For example, it is easy to recognize that the AR method provides a sample from the unrestricted normal distribution $\mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$, the Stern method generates draws from the normal distribution $\mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}-\boldsymbol{\Lambda})$, while GHK simulation relies on the recursive importance density in (6) in which draws depend only on the restrictions implied by y_{ij} but ignore the restrictions implied by subsequent $\{y_{ik}\}_{k>j}$. These mismatches between the proposal and target densities may adversely affect the efficiency of the AR, GHK, and Stern methods. We next introduce a class of simulated likelihood methods which are, in fact, based on draws from the truncated density of interest.

3 MCMC Methods

The calculation of multivariate integrals that generally have no analytical solution has been an important research area in Bayesian statistics. In particular, a key quantity of interest in Bayesian model comparison is the marginal likelihood, which is obtained by integrating the likelihood function with respect to the prior distribution of the parameters (for a discussion, see Kass and Raftery, 1995, and the references therein). It is one of the basic goals of this paper to link the simulated likelihood literature with that on Bayesian model choice in order to introduce MCMC methods as new and viable approaches to simulated likelihood estimation in discrete data analysis. Another goal is to develop new MCMC methods that are specifically tailored to simulated likelihood estimation. Our third goal is to provide an efficient simulation method for sampling $\boldsymbol{z}_i \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$, which is particularly important in this class of models but also has broad ramifications beyond simulated likelihood estimation. These goals are pursued in the remainder of this section.

3.1 The CRB Method

To see the common fundamentals between outcome probability estimation and Bayesian model choice, and to establish the framework for the estimation methods that will be discussed subsequently, we begin by rewriting the expression for $\Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega})$. In particular, note that we can write the probability in (2) as

$$\Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}) = \int 1\{\boldsymbol{z}_i \in \mathcal{B}_i\} f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}) d\boldsymbol{z}_i = \frac{1\{\boldsymbol{z}_i \in \mathcal{B}_i\} f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})}{f_{TN}(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})}, \tag{8}$$

which can be interpreted in terms of Bayes formula based on the recognition that the indicator function $1\{\boldsymbol{z}_i \in \boldsymbol{B}_i\}$ actually gives $\Pr(\boldsymbol{y}_i|\boldsymbol{z}_i)$ and hence can be treated as a "likelihood", $f_N(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$ can be treated as a "prior" because it does not respect the truncation implied by $\boldsymbol{y}_i$, and $f_{TN}(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$ can be viewed as a "posterior" that accounts for the truncation constraints reflected in $\boldsymbol{y}_i$. Thus, we can see that $\Pr(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Omega})$ can actually be viewed as a "marginal likelihood", i.e. the normalizing constant of the "posterior" $f_{TN}(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$. Even though the interpretation

of $\Pr(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega})$ as the normalizing constant of a truncated normal distribution is directly visible from (2), its reinterpretation in terms of the quantities in (8) is useful for developing empirical strategies for its estimation. In fact, the equivalent of equation (8) was used in Chib (1995) in developing his method for marginal likelihood estimation. This identity is particularly useful because, as discussed in Chib (1995), it holds for any value of $\boldsymbol{z}_i \in \mathcal{B}_i$ and therefore the calculation is reduced to finding the estimate of the ordinate $f_{TN}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$ at a single point $\boldsymbol{z}_i^* \in \mathcal{B}_i$. In our implementation, an estimate of the log-probability is obtained as

$$\ln \widehat{\Pr}\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right) = \ln f_N(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}) - \ln \widehat{f_{TN}}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}), \tag{9}$$

where we take $\boldsymbol{z}_i^*$ to be the sample mean of the MCMC draws $\boldsymbol{z}_i^{(g)} \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$, $g = 1,\ldots,G$, and make use of the fact that the numerator quantities $1\{\boldsymbol{z}_i^* \in \mathcal{B}_i\}$ and $f_N(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$ in (8) are directly available.

Draws $\boldsymbol{z}_i^{(g)} \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$ can be produced by employing the Gibbs sampling algorithm of Geweke (1991) in which a new value for $\boldsymbol{z}_i$ is generated by iteratively simulating each element z_{ij} from its full-conditional density $z_{ij} \sim f\left(z_{ij}|\{z_{ik}\}_{k\neq j}, y_{ij}, \boldsymbol{\beta},\boldsymbol{\Omega}\right) = \mathcal{TN}_{\mathcal{B}_{ij}}\left(\mu_{ij},\sigma_{ij}^2\right)$ for $j = 1,\ldots,J$, where μ_{ij} and σ_{ij}^2 are the conditional mean and variance of z_{ij} given $\{z_{ik}\}_{k\neq j}$, which are obtained by the usual updating formulas for a Gaussian density. Note that unlike the aforementioned importance sampling methods, a Gibbs sampler constructed in this way produces draws from the exact truncated normal distribution of interest and those draws will be used to estimate $f_{TN}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$, thereby leading to an estimate of $\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$.

To estimate the ordinate $f\left(\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) = f_{TN}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$, the joint density is decomposed by the law of total probability as

$$f\left(\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) = \prod_{j=1}^{J} f\left(z_{ij}^*|\boldsymbol{y}_i,\{z_{ik}^*\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right).$$

In the context of Gibbs sampling when the full-conditional densities are fully known, Chib (1995) proposed finding the ordinates $f\left(z_{ij}^*|\boldsymbol{y}_i,\{z_{ik}^*\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ for $1 < j < J$ by Rao-Blackwellization

(Tanner and Wong, 1987; Gelfand and Smith, 1990) in which the terms in the decomposition are represented by

$$f\left(z_{ij}^{*}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)=\int f\left(z_{ij}^{*}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\{z_{ik}\}_{k>j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$$
$$\times f\left(\{z_{ik}\}_{k>j}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)d\{z_{ik}\}_{k>j}$$

and estimated as

$$\widehat{f}\left(z_{ij}^{*}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)=G^{-1}\sum_{g=1}^{G}f\left(z_{ij}^{*}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\{z_{ik}^{(g)}\}_{k>j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$$

where the draws $\{z_{ik}^{(g)}\}_{k>j}$ come from a reduced run in which the latent variables $\{z_{ik}^{*}\}_{k<j}$ are fixed and sampling is over $\{z_{ik}^{(g)}\}_{k\geq j}\sim f\left(\{z_{ik}\}_{k\geq j}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$. Excluding $z_{ij}^{(g)}$ from $\{z_{ik}^{(g)}\}_{k\geq j}$ yields draws $\{z_{ik}\}_{k>j}\sim f\left(\{z_{ik}\}_{k>j}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ that are required in the average. The ordinate $f\left(z_{i1}^{*}|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ is estimated with draws from the main MCMC run, while the ordinate $f\left(z_{iJ}^{*}|\boldsymbol{y}_i,\{z_{ik}^{*}\}_{k<J},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ is available directly, and hence the method requires $(J-2)$ reduced MCMC runs. An advantage of this approach is that it breaks a large-dimensional problem into a set of smaller and more manageable steps and, at the cost of additional MCMC simulation, typically leads to very efficient estimates in many practical problems.

In the remainder of this article, we will refer to Chib's method with Rao-Blackwellization as the CRB method. This method provides a direct application of existing MCMC techniques (Chib, 1995) to simulated likelihood estimation and forms an important benchmark case against which other MCMC methods can be compared. Moreover, the CRB method provides continuous and differentiable probability estimates in the context of estimating (2), which distinguishes it from the other MCMC methods referenced in the Introduction. It will also form a basis for the new estimators that will be developed in the remainder of this section.

3.2 The CRT Method

Our first extension aims to address a potential drawback of Rao-Blackwellization, namely the cost of the additional reduced MCMC runs that it requires. For this reason we examine a different

way of obtaining $\widehat{f_{TN}}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})$ that is required in (8) or (9). An approach to density estimation which is based on the Gibbs transition kernel and does not entail reduced runs is discussed in Ritter and Tanner (1992). In particular, the Gibbs transition kernel for moving from $\boldsymbol{z}_i$ to $\boldsymbol{z}_i^*$ is given by the product of well-known univariate truncated normal full-conditional densities

$$K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)=\prod_{j=1}^{J} f\left(z_{ij}^*|\boldsymbol{y}_i,\{z_{ik}^*\}_{k<j},\{z_{ik}^{(g)}\}_{k>j},\boldsymbol{\beta},\boldsymbol{\Omega}\right). \tag{10}$$

Because the full-conditional densities are the fundamental building blocks of the Gibbs sampler, the additional coding involved in evaluating (10) is minimized. By virtue of the fact that the Gibbs sampler satisfies Markov chain invariance (see, e.g., Tierney, 1994; Chib and Greenberg, 1996), we have that

$$f_{TN_{\mathcal{B}_i}}\left(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)=\int K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) f_{TN_{\mathcal{B}_i}}\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i, \tag{11}$$

which was exploited for density estimation in Ritter and Tanner (1992). Therefore, an estimate of the denominator in (8) can be obtained by invoking (11) and averaging the transition kernel $K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ with respect to draws from the truncated normal distribution $\boldsymbol{z}_i^{(g)} \sim \mathcal{TN}_{\mathcal{B}_i}\left(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)$, i.e.

$$\widehat{f_{TN_{\mathcal{B}_i}}}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega})=\frac{1}{G}\sum_{g=1}^{G} K\left(\boldsymbol{z}_i^{(g)},\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right). \tag{12}$$

As in the CRB method, the random draws $\boldsymbol{z}_i^{(g)}$ required in the average are generated by a Gibbs sampler that iteratively simulates each element z_{ij} from its full-conditional distribution $z_{ij} \sim f\left(z_{ij}|\{z_{ik}\}_{k\neq j},\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ for $j=1,\ldots,J$.

Because this method combines the marginal likelihood estimation approach of Chib (1995) with the density ordinate estimation approach of Ritter and Tanner (1992), it will be referred to as the CRT method in the remainder of this paper. Several remarks about the CRT method and its relationship to CRB can be made. First, because the CRT and CRB methods are continuous and differentiable, they are applicable in maximum likelihood estimation and other problems that require differentiation. Second, in contrast to CRB, CRT does not require reduced run simulation

as all ordinates are estimated with draws from the main MCMC run. However, CRT may require storage for the latent variables $\{\boldsymbol{z}_i^{(g)}\}$, because the point $\boldsymbol{z}_i^*$, typically taken to be the mean of $f_{TN_{\mathcal{B}_i}}(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$, may not be available during the main MCMC run, thus preventing concurrent evaluation of $K\left(\boldsymbol{z}_i^{(g)}, \boldsymbol{z}_i^*|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}\right)$. If storage is a problem, then estimation can involve some limited amount of pre-computation such as a short MCMC run to determine $\boldsymbol{z}_i^*$ for subsequent evaluation of the Gibbs kernel. Note, however, that such a problem rarely presents itself in Bayesian studies where $\boldsymbol{z}_i^*$ may be readily available from MCMC runs conducted during the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. Third, note that in bivariate problems CRB will be more efficient than CRT because it does not involve any reduced runs and only requires estimation of $f(z_{i1}^*|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega})$, whereas $f(z_{i2}^*|\boldsymbol{y}_i, z_{i1}^*, \boldsymbol{\beta}, \boldsymbol{\Omega})$ is directly available. Finally, the main ideas stemming from the CRB and CRT approaches—that response probability evaluation can be reduced to finding a density ordinate and that the Gibbs kernel can be employed in estimating this density ordinate—will form a foundation for the methods that we discuss next. The key distinction between the alternatives that we consider has to do with the way in which the sample of latent data $\{\boldsymbol{z}_i^{(g)}\}$ is generated.

3.3 The ARK Method

Our second extension deals with the AR estimator. As discussed in Section 2, the AR approach can be appealing because of its simplicity and ease of implementation, but can be problematic because of its non-differentiability and discontinuity and the potential for numerical instability when estimating probabilities near 0 or 1. In this section, we show that the integration of MCMC theory into AR sampling can produce a method that circumvents many of the drawbacks of standard AR estimation. An important advantage of the proposed method relative to the estimator in equation (4) is that continuity and differentiability are introduced without sacrificing simulation consistency or requiring additional tuning parameters. Because the approach combines the AR simulator with the kernel of the Gibbs sampler, we will refer to it as the ARK method.

The derivation of the ARK method is fairly uncomplicated. It proceeds by simply rewriting the

invariance condition in (11) as

$$\begin{aligned} f_{TN_{\mathcal{B}_i}}\left(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) &= \int K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) f_{TN_{\mathcal{B}_i}}\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i \\ &= \int K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right) 1\{\boldsymbol{z}_i\in\mathcal{B}_i\} f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i, \end{aligned} \tag{13}$$

which suggests a straightforward way of producing an estimate $\widehat{f_{TN_{\mathcal{B}_i}}}\left(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ that can be used to obtain $\widehat{\Pr}\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ by (8) or (9). Specifically, from equation (13) it follows that $f_{TN_{\mathcal{B}_i}}\left(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ can be estimated by drawing $\boldsymbol{z}_i\sim\mathcal{N}\left(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)$, accepting only draws that satisfy $\boldsymbol{z}_i\in\mathcal{B}_i$, and using those draws to average $K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ as in (12).

At this point, it may be helpful to review the main pros and cons of ARK estimation in some detail. First, the ARK method retains the simplicity of AR sampling, while simultaneously offering continuous, differentiable, and simulation consistent estimates of $\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ based on the Gibbs kernel (even though simulation of $\{\boldsymbol{z}_i^{(g)}\}$ does not involve Gibbs sampling as in CRB or CRT). Second, because ARK subsumes the traditional AR estimator, the AR estimate will also typically be available as a by-product of ARK estimation. Third, although both ARK and CRT average the kernel in (12) using latent data $\boldsymbol{z}_i\sim\mathcal{TN}_{\mathcal{B}_i}\left(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right)$, the fact that the latent data are obtained by either AR or Gibbs sampling can have important implications for the relative efficiency or ARK versus CRT. To see this, consider Figure 1. The figure shows that with low correlations, the Gibbs sampler can traverse the parameter space relatively quickly, without inducing much serial dependence in the sampled $\{\boldsymbol{z}_i^{(g)}\}$. When the elements of $\boldsymbol{z}_i$ are highly correlated, however, iterative sampling of the full-conditional distributions produces relatively small Gibbs steps that lead to slow mixing of the Markov chain. In contrast, ARK provides an independent sample of draws whose mixing is unaffected by the extent of correlation between the elements of $\boldsymbol{z}_i$.

One should keep in mind, however, that this advantage of the ARK approach comes at the cost of a well-known problem with AR samplers that too many rejections may occur if $\Pr\left(\boldsymbol{z}_i\in\mathcal{B}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ is relatively low, thereby adversely affecting simulation efficiency. In some cases, this problem may be remedied by estimating $\Pr\left(\boldsymbol{z}_i\in\mathcal{B}_i^c|\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ because when the probability of $\mathcal{B}_i$ is small, that of its

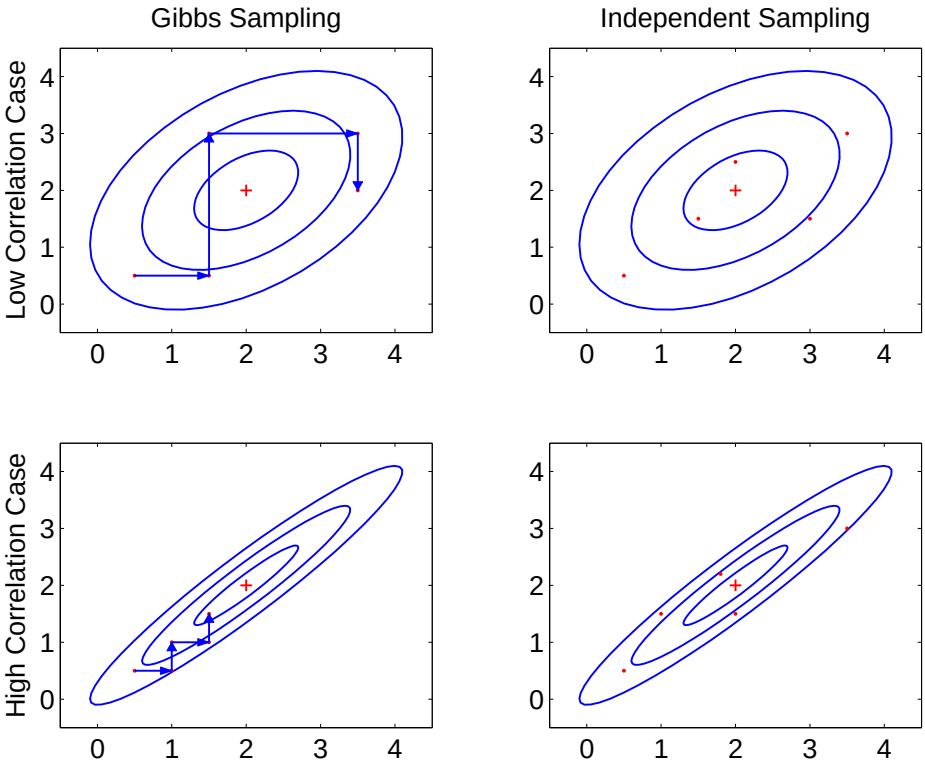

Figure 1: Gibbs versus independent sampling from distributions with varying degrees of correlation.

complement $\mathcal{B}_i^c$ must be relatively large. However, we caution that in doing so, one must be careful to ensure that a number of technical requirements are met. In particular, while the set $\mathcal{B}_i$ is convex, its complement $\mathcal{B}_i^c$ need not be. As a result, some choices of $\boldsymbol{z}_i^* \in \mathcal{B}_i^c$ may potentially introduce non-differentiability in the estimate of $\Pr(\boldsymbol{z}_i \in \mathcal{B}_i^c | \boldsymbol{\beta}, \boldsymbol{\Omega})$ because the kernel $K(\boldsymbol{z}_i, \boldsymbol{z}_i^* | \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega})$ may not be strictly positive for all $\{\boldsymbol{z}_i\}$. Even worse, for some settings of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ the non-convexity of $\mathcal{B}_i^c$ may lead to near reducibility of the Markov chain on $\mathcal{B}_i^c$, rendering convergence and kernel estimation altogether problematic. Therefore, ARK estimation of $\Pr(\boldsymbol{z}_i \in \mathcal{B}_i^c | \boldsymbol{\beta}, \boldsymbol{\Omega})$ should only be attempted after careful consideration of the aforementioned issues.

3.4 The ASK method

In this section we discuss an approach which aims to improve the quality of the sample of $\{\boldsymbol{z}_i\}$ that is used in estimation by addressing some of the simulation difficulties discussed in Section 3.3.

Another look at Figure 1 suggests that improving the mixing properties of the Gibbs sampler in problems with high correlation would be key to reducing the serial dependence in the MCMC sample $\boldsymbol{z}_i \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$ which, in turn, can reduce the sampling variability of the average in (12). Moreover, the discussion in Section 3.3 also indicates that Gibbs sampling has important advantages over AR sampling because every Gibbs draw satisfies $\boldsymbol{z}_i \in \mathcal{B}_i$, whereas meeting this requirement may lead to large rejection rates in AR simulation.

In developing the method, we link a variety of approaches and introduce a new adaptive MCMC algorithm for simulating $\boldsymbol{z}_i \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$ which improves the quality of the MCMC sample. We build upon Chib (1995) to relate estimation of $\Pr(\boldsymbol{z}_i \in \mathcal{B}_i|\boldsymbol{\beta}, \boldsymbol{\Omega})$ to that of $f_{TN_{\mathcal{B}_i}}(\boldsymbol{z}_i^*|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$, rely on ideas from Ritter and Tanner (1992) to obtain the latter quantity, and use full-conditional truncated normal sampling (see Geweke, 1991), but with the key difference that our proposed Gibbs algorithm improves mixing by adaptively sampling either the latent $\{\boldsymbol{z}_i\}$ or a particular transformation of those variables. Specifically, we use the Mahalanobis transformation to map $\{\boldsymbol{z}_i\}$ into *a priori* independent standard normal variables $\{\boldsymbol{\eta}_i\}$ such as those used in (5) to develop the recursive conditioning importance density of the GHK estimator. Due to the particular combination of inputs that appear in this method, in the remainder of this paper we shall refer to it as the adaptive sampling kernel (ASK) method.

The ASK approach proceeds along the following lines. We write the model as $\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{L}$ is a lower triangular Cholesky factor such that $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Omega}$. Then, solving for $\boldsymbol{\eta}_i$, we obtain $\boldsymbol{\eta}_i = \boldsymbol{L}^{-1}(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, which is the Mahalanobis transformation of $\boldsymbol{z}_i$. Even though the elements of $\boldsymbol{\eta}_i$ are *a priori* independent, it is important to note that conditioning on $\boldsymbol{y}_i$ introduces dependence through the constraints on each η_{ij} in the full-conditional distributions $f(\eta_{ij}|\{\eta_{ik}\}_{k\neq j}, \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega})$, $j = 1, \ldots, J$. To see this, note that the constraints on η_{ij} are obtained from those on $\boldsymbol{z}_i$ by solving the system $\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_i$, and that η_{ij} enters all equations for which the elements in the jth column of $\boldsymbol{L}$ are not zero ($\boldsymbol{L}$ is lower triangular by construction, but it can possibly contain zero elements below the main diagonal). Let $\mathcal{E}_{ijk}$ denote the feasible region

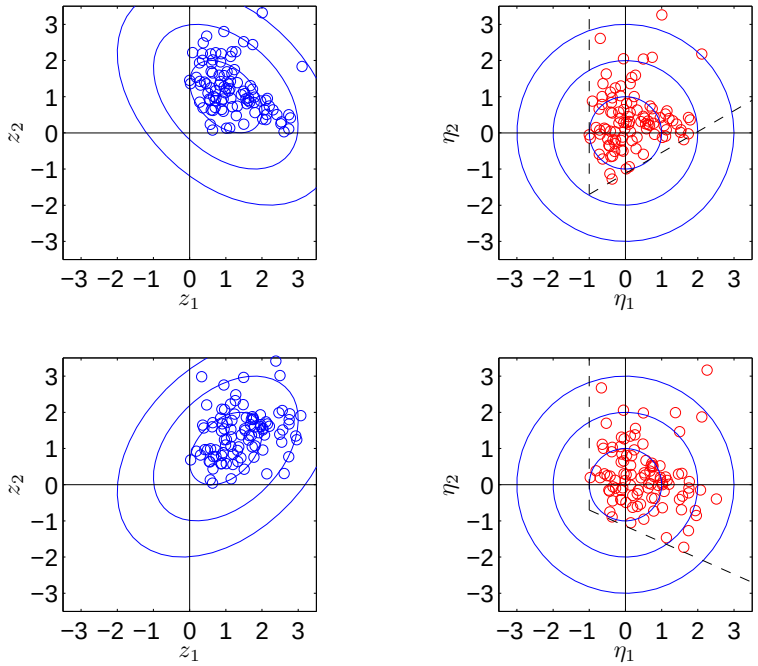

Figure 2: Correspondence between $\boldsymbol{z}_i \in \mathcal{B}_i$ and the Mahalanobis transform $\boldsymbol{\eta}_i \in \mathcal{E}_i$ for $\boldsymbol{y}_i = \mathbf{1}_2$. The Mahalanobis transform orthogonalizes and standardizes $\boldsymbol{z}_i$ to produce $\boldsymbol{\eta}_i$, but causes dependence to be reflected in the boundaries of $\mathcal{E}_i$.

for η_{ij} implied by the kth equation, and let $\mathcal{E}_{ij} = \cap_{k=j}^{J} \mathcal{E}_{ijk}$ and $\mathcal{E}_i = \{\boldsymbol{\eta}_i : \eta_{ij} \in \mathcal{E}_{ij}, \forall j\}$. Readers may recall that some constraints arising from $\boldsymbol{y}_i$ are ignored in the GHK method in order to obtain a tractable importance density. However, all constraints must be incorporated in the sequence of Gibbs steps

$$[\eta_{ij} | \{\eta_{ik}\}_{k \neq j}, \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}] \sim \mathcal{TN}_{\mathcal{E}_{ij}}(0, 1), \quad j = 1, \ldots, J,$$

leading to the Gibbs kernel

$$K\left(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i^{\dagger} | \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}\right) = \prod_{j=1}^{J} f\left(\eta_{ij}^{\dagger} | \boldsymbol{y}_i, \{\eta_{ik}^{\dagger}\}_{k<j}, \{\eta_{ik}^{(g)}\}_{k>j}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right), \tag{14}$$

so that MCMC simulation produces $\boldsymbol{\eta}_i \sim \mathcal{TN}_{\mathcal{E}_i}(\mathbf{0}, \boldsymbol{I})$ that correspond to $\boldsymbol{z}_i \sim \mathcal{TN}_{\mathcal{B}_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega})$.

Some intuition about the mechanics of the ASK approach can be gleaned from Figure 2, which relates the sets $\mathcal{B}_i$ and $\mathcal{E}_i$ implied by observing $\boldsymbol{y}_i = \mathbf{1}_2$. The Mahalanobis transformation demeans, orthogonalizes and re-scales the draws $\boldsymbol{z}_i$ to produce $\boldsymbol{\eta}_i$, but these operations also map $\mathcal{B}_i$ into $\mathcal{E}_i$ by shifting the vertex of the feasible set and rotating its boundaries (the axes) depending on the sign

of the covariance elements in $\boldsymbol{\Omega}$. Note that because η_{ij} enters the equations for $\{z_{ik}\}_{k\geq j}$, updating η_{ij} corresponds to simultaneously updating multiple elements of $\boldsymbol{z}_i$; conversely, updating z_{ij} affects all elements $\{\eta_{ik}\}_{k\leq j}$ that enter the jth equation. The key feature of the transformation that will be exploited here is that it offers a trade-off between correlation (in the case of $\boldsymbol{z}_i$) and dependence in the constraints (for the elements of $\boldsymbol{\eta}_i$) and implies that important benefits can be obtained by adaptively sampling the elements of $\boldsymbol{z}_i$ or those of the Mahalanobis transformation $\boldsymbol{\eta}_i$.

To understand the trade-offs between Gibbs simulation of $\eta_{ij} \sim f\left(\eta_{ij}|\{\eta_{ik}\}_{k\neq j}, \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Omega}\right)$ as a way of obtaining $\boldsymbol{\eta}_i \sim \mathcal{TN}_{\mathcal{E}_i}\left(\boldsymbol{0}, \boldsymbol{I}\right)$ and the resultant $\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_i$, and Gibbs sampling of $z_{ij} \sim f\left(z_{ij}|\{z_{ik}\}_{k\neq j}, \boldsymbol{\beta}, \boldsymbol{\Omega}\right)$ which yields $\boldsymbol{z}_i \sim \mathcal{TN}_{\mathcal{B}_i}\left(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}\right)$ directly, consider a setting where $\boldsymbol{\Omega}$ contains high correlations but the constraints implied by $\boldsymbol{y}_i$ are relatively mildly binding. In this case, it will be beneficial to simulate $\boldsymbol{\eta}_i$ because $f_{TN_{\mathcal{E}_i}}\left(\boldsymbol{\eta}_i|\boldsymbol{0}, \boldsymbol{I}\right) \to f_N\left(\boldsymbol{\eta}_i|\boldsymbol{0}, \boldsymbol{I}\right)$ as $\Pr\left(\boldsymbol{\eta}_i \in \mathcal{E}_i\right) \to 1$ and drawing $\boldsymbol{\eta}_i$ produces a sample that will be close to iid. In contrast, a traditional Gibbs sampler defined on the elements of $\boldsymbol{z}_i$ will exhibit high serial correlation between successive MCMC draws because such a sampler must traverse large portions of the support by taking small steps (recall the discussion of Figure 1). Note also that as the correlations in $\boldsymbol{\Omega}$ increase towards 1 for similar components of $\boldsymbol{y}_i$ or decrease towards -1 for dissimilar components of $\boldsymbol{y}_i$, the feasible sets tend to be binding on one η_{ij} but not the other, and the MCMC sampler is well behaved. In other cases, it may be better to sample $\boldsymbol{z}_i$ directly without transforming to $\boldsymbol{\eta}_i$, for example when the constraints $\mathcal{E}_{ij} = \cap_{k=j}^{J}\mathcal{E}_{ijk}$ on η_{ij} are such that they slow down the mixing of other $\{\eta_{ik}\}_{k\neq j}$. Some of these scenarios, together with measures of sampling inefficiency (to be discussed shortly) for each Gibbs kernel ($K_z(\cdot)$ and $K_\eta(\cdot)$) are presented in Figure 3. In yet other cases, for example when correlations are low or the probabilities to be estimated are small, the two sampling approaches will typically exhibit similar mixing and either approach will work well. However, in order to produce an MCMC sample that is as close to iid as possible, we have to be able to adaptively determine whether to simulate $\boldsymbol{\eta}_i$ (and convert to $\boldsymbol{z}_i$) or sample $\boldsymbol{z}_i$ directly. Our proposed approach for doing so is presented next.

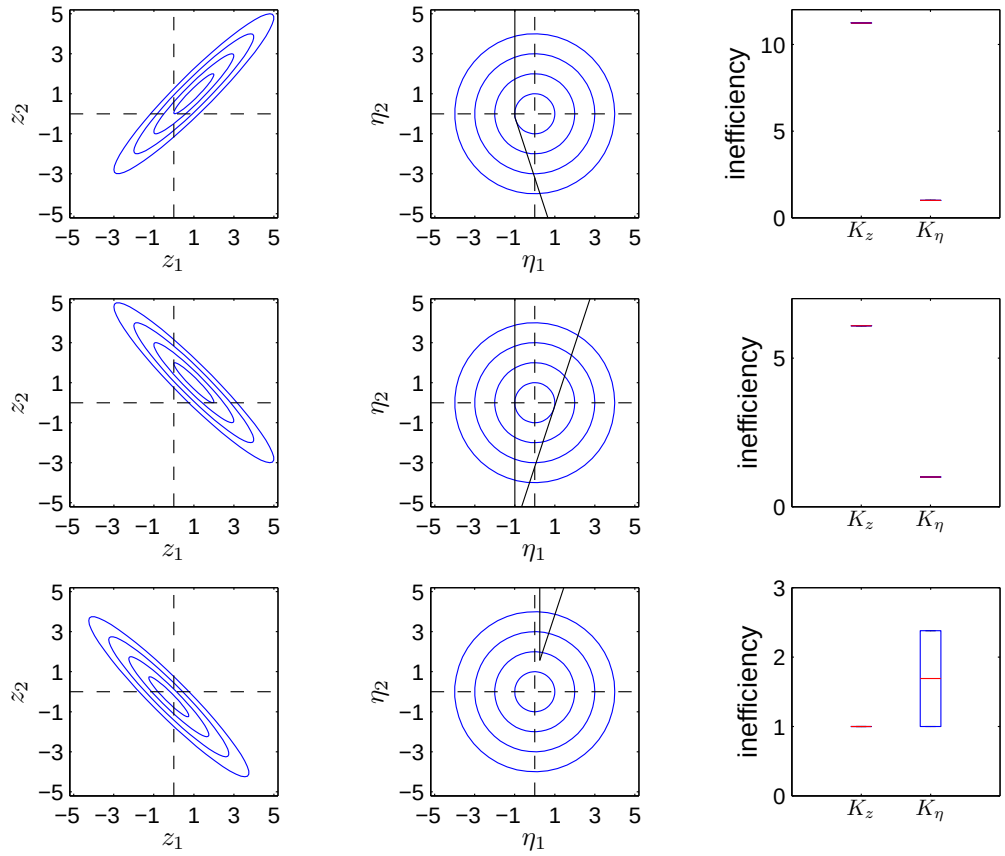

Figure 3: Performance of $K_z(\cdot)$ and $K_\eta(\cdot)$ in different settings. Higher inefficiencies indicate stronger autocorrelations between successive MCMC draws $\{\boldsymbol{z}_i\}$ for each kernel (1 indicates iid sampling).

Algorithm 1 ***Adaptive Gibbs Sampler for Multivariate Truncated Normal Simulation***

1. *Initialize* $p_{\boldsymbol{\eta}} \in (0,1)$ *and let* $p_{\boldsymbol{z}} = 1 - p_{\boldsymbol{\eta}}$*;*

2. *Given the current* $\boldsymbol{z}_i$ *and the corresponding* $\boldsymbol{\eta}_i$ *in the Markov chain, with probability* $p_{\boldsymbol{\eta}}$ *sample* $\boldsymbol{\eta}_i$ *using the Gibbs kernel* $K_\eta(\cdot)$ *in* (14) *and convert to* $\boldsymbol{z}_i$ *by* (5)*, or, with probability* $p_{\boldsymbol{z}}$ *sample* $\boldsymbol{z}_i$ *directly using the Gibbs kernel* $K_z(\cdot)$ *in* (10)*;*

3. *After a burn-in period, accumulate the sample* $\{\boldsymbol{z}_i\}$ *while keeping track of the draws obtained by* $K_\eta(\cdot)$ *and* $K_z(\cdot)$*;*

4. *Periodically update* $p_{\boldsymbol{\eta}}$ *using a rule* $\mathcal{P}_\eta : \Re^{(2J)} \to [0,1]$ *that maps the autocorrelations from the two kernels to the closed interval* $[0,1]$*;* $\mathcal{P}_\eta$ *is an increasing function in the autocorrelations*

of the draws produced by $K_z(\cdot)$ and a decreasing function of those produced by $K_\eta(\cdot)$.

We now discuss Algorithm 1 in greater detail. From a theoretical point of view, the algorithm is quite transparent: it is very simple to show that the mixture of kernels, each of which converges to the target distribution, also converges to that distribution. Specifically, one only has to observe that invariance is satisfied for each kernel (see, for example, Chib and Greenberg, 1996) and therefore for any weighted average (mixture) of those kernels. An interesting observation, based on our experience with step 1, is that good mixing in the initial stages of sampling does not require that the better mixing sampler be favored by the initial choice of $p_{\boldsymbol{\eta}}$ and $p_{\boldsymbol{z}}$. In fact, a "neutral" probability choice $p_{\boldsymbol{\eta}} = p_{\boldsymbol{z}} = 0.5$ typically leads to a mixed sampler whose performance is more than proportionately closer to that of the more efficient transition kernel. The goal of steps 3 and 4 is to ensure that if one kernel dominates the other on every margin (i.e. sampling is more efficient for every element of $\boldsymbol{z}_i$), the mixed chain settles on that more efficient kernel; otherwise, the aim is to produce an adaptive Markov chain that strikes a balance between the two kernels in a way that reduces overall inefficiency. There are many possible ways in which $p_{\boldsymbol{\eta}}$ and $p_{\boldsymbol{z}}$ can be determined depending on one's aversion (as captured by some loss function) to slow mixing for each element of $\boldsymbol{z}_i$. In our examples we considered the following $\mathcal{P}_\eta$:

$$p_{\boldsymbol{\eta}} = \begin{cases} 1 & \text{if } \boldsymbol{r}_z \gg \boldsymbol{r}_\eta \\ 0 & \text{if } \boldsymbol{r}_\eta \gg \boldsymbol{r}_z \\ \frac{\boldsymbol{w}'\boldsymbol{r}_z}{\boldsymbol{w}'\boldsymbol{r}_z + \boldsymbol{w}'\boldsymbol{r}_\eta} & \text{otherwise} \end{cases},$$

where $\boldsymbol{w}$ is a vector of (loss function) weights, $\boldsymbol{r}_z$ and $\boldsymbol{r}_\eta$ are J-vectors of inefficiency measures for each element of $\boldsymbol{z}_i$ under the two MCMC sampling kernels, and '$\gg$' denotes element-by-element inequality. Let $\rho_{jl} \equiv \text{corr}\left(z_{ij}^{(g)}, z_{ij}^{(g-l)}\right)$ be the lth autocorrelation for the sampled draws of z_{ij}. Note that in computing ρ_{jl} for each of the two kernels in Algorithm 1, one should ensure that the draws $z_{ij}^{(g)}$ come from the kernel of interest, even though the draws $z_{ij}^{(g-l)}$ could have been generated by either kernel. As a quick and inexpensive (but fairly accurate) measure of the inefficiency factors $1+\sum_{l=1}^{\infty}\rho_{jl}$ of the two samplers, we use the draws generated by $K_z(\cdot)$ to compute $\boldsymbol{r}_z[j] = (1-\rho_{j1})^{-1}$ and similarly base the computation of $\boldsymbol{r}_\eta[j] = (1-\rho_{j1})^{-1}$ on draws generated by $K_\eta(\cdot)$. These

expressions require calculation of only the first-order autocorrelations and minimize computational and bookkeeping requirements but approximate the inefficiency factors rather well. Note also that in determining the mixing probabilities, the vector $\boldsymbol{w}$ could contain equal weights if the goal is to improve overall MCMC mixing. However, the weights can easily be adjusted when it may be desirable to weigh the mixing of a particular subset of $\boldsymbol{z}_i$ more heavily, such as in problems when a subset of $\boldsymbol{z}_i$ must be integrated out. A final remark is that it will typically suffice to update $p_{\boldsymbol{\eta}}$ only a few times in the course of sampling and that the sampling probability tends to stabilize very rapidly, almost immediately in the case of algorithms that exhibit widely diverging MCMC mixing properties.

The definition of the ASK simulator is completed by noting that once a sample of draws $\{\boldsymbol{z}_i^{(g)}\}$ is available, then estimation proceeds by (12) and (9). We emphasize that while by construction it is true that

$$\begin{aligned}\Pr\left(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\Omega}\right) &= \int_{\mathcal{B}_i} f_N\left(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}\right) d\boldsymbol{z}_i \\ &= \int_{\mathcal{E}_i} f_N(\boldsymbol{\eta}_i|\boldsymbol{0},\boldsymbol{I}) d\boldsymbol{\eta}_i,\end{aligned}$$

we do not use the second representation in estimation (and only rely on it in simulation) because after the Mahalanobis transformation the dependence in the constraints, seen in Figure 2, implies that some values of $\boldsymbol{\eta}_i$ will possibly lead to $K\left(\boldsymbol{\eta}_i,\boldsymbol{\eta}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)=0$ which may lead to non-differentiability of the resulting probability estimate. This, however, is not a problem when the draws $\{\boldsymbol{\eta}_i^{(g)}\}$ are converted to $\{\boldsymbol{z}_i^{(g)}\}$ and the kernel $K\left(\boldsymbol{z}_i,\boldsymbol{z}_i^*|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Omega}\right)$ is used in estimation.

3.5 Summary and Additional Considerations

In this section we have presented a variety of MCMC methods for estimating response probabilities in discrete data models. We have shown that simulated likelihood estimation can proceed by adapting methods from the Bayesian literature on marginal likelihood estimation and have developed a set of new techniques designed to address features that are specific to simulated likelihood evaluation. The methods are applicable in binary, ordinal, censored, count, and other settings,

and can be easily extended to handle mixtures and scale-mixtures of normal distributions that include the Student's t-link and logit models, among others (see, e.g., Andrews and Mallows, 1974; Poirier, 1978; Albert and Chib, 1993; Geweke, 1993), and to models with heteroskedasticity (Gu et al., 2009). Moreover, even though for most of the approaches presented here we have discussed Gibbs kernel versions of estimating the posterior ordinate (as in Ritter and Tanner, 1992), we emphasize that it is possible to use Rao-Blackwellization as in Section 3.1, which can be desirable in high-dimensional problems or in settings where natural groupings of the latent variables may be present.

An important goal of this paper has been to consider approaches for obtaining MCMC samples $\{\boldsymbol{z}_i\}$ that result in better mixing of the Markov chain and improved efficiency of estimation. The improvements in simulation made possible by Algorithm 1 have ramifications not only for estimation of response probabilities, but also for problems in which high-quality samples from a truncated normal distribution are needed. For example, Chib's approach, which was discussed in Section 3.1, can be combined with the output of Algorithm 1 to further improve its efficiency. Many of the methods discussed here can also be combined with recently developed simulation techniques such as slice sampling (Neal, 2003; Damien and Walker, 2001) and antithetic draws such as those produced by reflection samplers and Halton sequences (see, e.g., Tierney, 1994; Train, 2003; Bhat, 2001, 2003). In this paper we focused on algorithms that provide continuous and differentiable probability estimates, but have also cited a number of important MCMC approaches that lead to non-differentiable estimates. It is useful to keep in mind that many of these latter methods can still be applied in optimization algorithms that do not require differentiation – for example in simulated annealing (Goffe et al., 1994) and particle swarming (Kennedy and Eberhart, 2001), although such algorithms involve computationally intensive stochastic search that typically requires numerous evaluations of the objective function.

Finally, we remark that although our discussion has centered on discrete data models, the techniques developed in this paper are directly applicable to the computation of p-values for multivariate

directional hypothesis tests.

4 Comparison of Simulators

We carried out a simulation study to examine the performance of the techniques proposed in Section 3 and compare them to the methods discussed in Section 2. In particular, we report estimates of the probability

$$\Pr\left(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\Omega}\right)=\int_{\mathcal{B}}f_N\left(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Omega}\right)d\boldsymbol{z}$$

under several settings of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Because the problem of estimating any orthant probability can always be represented as an equivalent problem of estimating the probability of another orthant by simple rotation of the space, without loss of generality we let $\boldsymbol{y}$ be a J-dimensional vector of ones, and hence $\mathcal{B}$ is the positive orthant. We vary the dimension of integration from $J=3$ to $J=12$ in increments of 3. In each case, we consider three settings of $\boldsymbol{\mu}$ and four settings of $\boldsymbol{\Omega}$. Specifically, when $J=3$, we let $\boldsymbol{\mu}_A=(0,0.5,1)'$ be the value of $\boldsymbol{\mu}$ that makes $\boldsymbol{y}$ "likely", $\boldsymbol{\mu}_B=(-0.5,0,0.5)'$ as the "intermediate" value of $\boldsymbol{\mu}$, and $\boldsymbol{\mu}_C=(-1,-0.5,0)'$ as the "least likely" value. For $J=6$ the "likely", "intermediate", and "least likely" values are obtained by setting $\boldsymbol{\mu}=(\boldsymbol{\mu}_A',\boldsymbol{\mu}_A')'$ or $\boldsymbol{\mu}=(\boldsymbol{\mu}_B',\boldsymbol{\mu}_B')'$ or $\boldsymbol{\mu}=(\boldsymbol{\mu}_C',\boldsymbol{\mu}_C')'$, respectively. The means are similarly constructed for higher values of J. We use a covariance matrix $\boldsymbol{\Omega}$ of the type $\boldsymbol{\Omega}\left[k,j\right]=\rho^{|k-j|}$, i.e.,

$$\boldsymbol{\Omega}=\begin{pmatrix}1 & \rho & \rho^2 & \cdots & & \rho^{J-1}\\ \rho & 1 & \rho & \cdots & & \rho^{J-2}\\ \rho^2 & \rho & 1 & & & \vdots\\ \vdots & & \ddots & \ddots & & \rho^2\\ & & & & 1 & \rho\\ \rho^{J-1} & \rho^{J-2} & \cdots & \rho^2 & \rho & 1\end{pmatrix},$$

where $\rho\in\{-0.7,-0.3,0.3,0.7\}$, which allows for high and low positive and negative correlations in the examples. Finally, the reported results for all simulators are based on simulation runs of length 10 000; for the three simulators requiring MCMC draws (CRB, CRT, and ASK) the main run is preceded by a burn-in of 1000 cycles.

Tables 1–4 present results for different settings of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ for $J = 3$, $J = 6$, $J = 9$, $J = 12$, respectively. We find that for low values of J and settings of $\boldsymbol{\mu}$ that make $\boldsymbol{y}$ "likely", all methods produce point estimates that agree closely. However, the variability differs widely across estimators and different settings of J, ρ, and $\boldsymbol{\mu}$ (note that the entries in parentheses have to be divided by 100 to obtain the actual numerical standard errors of the estimates). Among the traditional estimators, we see that GHK outperforms AR and Stern, regardless of the values of J, ρ, and $\boldsymbol{\mu}$. AR performs worst and can also fail in high-dimensional problems or in other settings where the outcome is "unlikely" and no draws are accepted. These findings for the traditional estimators are consistent with earlier studies (e.g. Börsch-Supan and Hajivassiliou, 1993; Hajivassiliou et al., 1996).

Table 1: Log-probability estimates ($J = 3$) with numerical standard errors ($\times 10^{-2}$) in parentheses.

	ρ	AR	STERN	GHK	CRB	CRT	ASK	ARK
$\boldsymbol{\mu} = \boldsymbol{\mu}_A$	-0.7	-1.574	-1.557	-1.556	-1.557	-1.557	-1.558	-1.560
		(1.956)	(0.896)	(0.386)	(0.086)	(0.082)	(0.085)	(0.171)
	-0.3	-1.396	-1.393	-1.392	-1.393	-1.393	-1.393	-1.394
		(1.743)	(0.414)	(0.123)	(0.018)	(0.017)	(0.017)	(0.033)
	0.3	-1.069	-1.059	-1.067	-1.065	-1.066	-1.065	-1.066
		(1.382)	(0.729)	(0.080)	(0.033)	(0.033)	(0.033)	(0.051)
	0.7	-0.835	-0.840	-0.836	-0.838	-0.836	-0.841	-0.834
		(1.143)	(0.895)	(0.094)	(0.265)	(0.318)	(0.239)	(0.395)
$\boldsymbol{\mu} = \boldsymbol{\mu}_B$	-0.7	-3.523	-3.498	-3.498	-3.502	-3.502	-3.502	-3.503
		(5.736)	(1.212)	(0.543)	(0.029)	(0.026)	(0.026)	(0.176)
	-0.3	-2.658	-2.664	-2.663	-2.665	-2.665	-2.665	-2.666
		(3.642)	(0.518)	(0.170)	(0.009)	(0.009)	(0.009)	(0.040)
	0.3	-1.862	-1.853	-1.867	-1.865	-1.865	-1.865	-1.865
		(2.332)	(1.151)	(0.113)	(0.023)	(0.024)	(0.023)	(0.057)
	0.7	-1.427	-1.406	-1.424	-1.423	-1.423	-1.424	-1.423
		(1.780)	(1.341)	(0.134)	(0.203)	(0.253)	(0.200)	(0.427)
$\boldsymbol{\mu} = \boldsymbol{\mu}_C$	-0.7	-7.824	-7.208	-7.208	-7.213	-7.213	-7.213	-7.220
		(49.990)	(1.463)	(0.676)	(0.008)	(0.008)	(0.008)	(1.881)
	-0.3	-4.656	-4.645	-4.645	-4.648	-4.648	-4.648	-4.648
		(10.211)	(0.607)	(0.212)	(0.004)	(0.004)	(0.004)	(0.097)
	0.3	-3.006	-2.979	-3.002	-3.000	-3.000	-3.000	-3.000
		(4.382)	(1.851)	(0.148)	(0.017)	(0.017)	(0.017)	(0.065)
	0.7	-2.248	-2.213	-2.237	-2.234	-2.235	-2.233	-2.230
		(2.910)	(2.105)	(0.179)	(0.175)	(0.199)	(0.172)	(0.532)

Table 2: Log-probability estimates ($J = 6$) with numerical standard errors ($\times 10^{-2}$) in parentheses.

	ρ	AR	STERN	GHK	CRB	CRT	ASK	ARK
$\boldsymbol{\mu} = \mathbf{1}_2 \otimes \boldsymbol{\mu}_A$	-0.7	-3.032	-3.097	-3.066	-3.074	-3.074	-3.076	-3.071
		(4.444)	(1.700)	(0.643)	(0.112)	(0.121)	(0.129)	(0.576)
	-0.3	-2.859	-2.831	-2.823	-2.828	-2.828	-2.828	-2.829
		(4.056)	(0.729)	(0.235)	(0.026)	(0.027)	(0.027)	(0.128)
	0.3	-2.039	-2.030	-2.041	-2.037	-2.037	-2.038	-2.036
		(2.586)	(1.237)	(0.221)	(0.049)	(0.055)	(0.053)	(0.146)
	0.7	-1.350	-1.355	-1.376	-1.371	-1.386	-1.372	-1.360
		(1.691)	(1.315)	(0.433)	(0.408)	(0.539)	(0.449)	(0.818)
$\boldsymbol{\mu} = \mathbf{1}_2 \otimes \boldsymbol{\mu}_B$	-0.7	-7.131	-7.211	-7.164	-7.174	-7.175	-7.175	-7.195
		(35.341)	(2.846)	(0.912)	(0.035)	(0.037)	(0.036)	(2.917)
	-0.3	-5.473	-5.480	-5.469	-5.475	-5.475	-5.475	-5.473
		(15.398)	(0.969)	(0.311)	(0.013)	(0.014)	(0.014)	(0.453)
	0.3	-3.527	-3.518	-3.534	-3.527	-3.528	-3.528	-3.529
		(5.746)	(2.238)	(0.297)	(0.037)	(0.041)	(0.040)	(0.183)
	0.7	-2.294	-2.246	-2.283	-2.277	-2.280	-2.278	-2.271
		(2.985)	(2.178)	(0.555)	(0.351)	(0.464)	(0.379)	(1.090)
$\boldsymbol{\mu} = \mathbf{1}_2 \otimes \boldsymbol{\mu}_C$	-0.7	.	-15.476	-15.444	-15.458	-15.458	-15.458	.
		.	(4.408)	(1.140)	(0.010)	(0.010)	(0.010)	.
	-0.3	.	-9.669	-9.656	-9.663	-9.663	-9.663	.
		.	(1.227)	(0.374)	(0.006)	(0.006)	(0.007)	.
	0.3	-5.497	-5.620	-5.629	-5.621	-5.621	-5.621	-5.624
		(15.585)	(4.486)	(0.374)	(0.027)	(0.030)	(0.028)	(0.450)
	0.7	-3.537	-3.518	-3.514	-3.498	-3.502	-3.503	-3.501
		(5.776)	(3.925)	(0.730)	(0.288)	(0.366)	(0.342)	(1.745)

The interesting finding from Tables 1–4 in this study is that the MCMC-based estimation methods perform very well. In fact, CRB, CRT and ASK outperform GHK most of the time, although the methods are roughly on par in "likely" cases characterized by high values of ρ. However, in cases with unlikely outcomes, the MCMC-based methods typically produce numerical standard errors that are much lower than those of GHK. Although it could be argued that some of the efficiency of CRB comes at the cost of additional reduced runs, neither CRT nor ASK require reduced runs and are still typically more efficient than GHK. These results present a strong case in favor of the proposed MCMC-based approaches. Even ARK, which similarly to AR could also fail when no draws are accepted, seem to provide very efficient estimates that are close to those of the other

Table 3: Log-probability estimates ($J = 9$) with numerical standard errors ($\times 10^{-2}$) in parentheses.

	ρ	AR	STERN	GHK	CRB	CRT	ASK	ARK
$\boldsymbol{\mu} = \mathbf{1}_3 \otimes \boldsymbol{\mu}_A$	-0.7	-4.585	-4.599	-4.610	-4.590	-4.588	-4.589	-4.566
		(9.851)	(2.578)	(0.864)	(0.155)	(0.142)	(0.156)	(1.719)
	-0.3	-4.200	-4.258	-4.269	-4.263	-4.263	-4.263	-4.260
		(8.103)	(0.957)	(0.318)	(0.035)	(0.034)	(0.032)	(0.350)
	0.3	-2.966	-3.001	-3.005	-3.008	-3.008	-3.009	-3.005
		(4.292)	(1.789)	(0.326)	(0.062)	(0.071)	(0.069)	(0.318)
	0.7	-1.885	-1.880	-1.890	-1.893	-1.901	-1.905	-1.906
		(2.363)	(1.784)	(0.611)	(0.563)	(0.738)	(0.519)	(1.394)
$\boldsymbol{\mu} = \mathbf{1}_3 \otimes \boldsymbol{\mu}_B$	-0.7	.	-10.877	-10.878	-10.846	-10.845	-10.846	.
		.	(5.093)	(1.277)	(0.046)	(0.043)	(0.040)	.
	-0.3	-7.824	-8.281	-8.293	-8.285	-8.285	-8.285	-8.449
		(49.990)	(1.312)	(0.421)	(0.017)	(0.017)	(0.016)	(4.721)
	0.3	-5.116	-5.157	-5.186	-5.190	-5.190	-5.190	-5.188
		(12.871)	(3.919)	(0.440)	(0.047)	(0.053)	(0.050)	(0.657)
	0.7	-3.128	-3.100	-3.107	-3.107	-3.113	-3.113	-3.090
		(4.672)	(3.294)	(0.910)	(0.457)	(0.609)	(0.520)	(2.229)
$\boldsymbol{\mu} = \mathbf{1}_3 \otimes \boldsymbol{\mu}_C$	-0.7	.	-23.764	-23.743	-23.702	-23.702	-23.702	.
		.	(8.661)	(1.615)	(0.012)	(0.012)	(0.011)	.
	-0.3	.	-14.675	-14.689	-14.678	-14.678	-14.678	.
		.	(1.725)	(0.505)	(0.008)	(0.008)	(0.008)	.
	0.3	-8.112	-8.141	-8.236	-8.241	-8.242	-8.242	-8.100
		(57.726)	(9.785)	(0.557)	(0.035)	(0.039)	(0.037)	(15.765)
	0.7	-4.804	-4.743	-4.733	-4.741	-4.743	-4.738	-4.740
		(10.998)	(6.980)	(1.264)	(0.375)	(0.518)	(0.473)	(4.517)

estimators (provided at least some draws are accepted).

In comparing the MCMC approaches to each other, we see that the ASK estimates, as expected, are at least as efficient as those from CRT, but that in many settings all three methods (ASK, CRT and CRB) perform similarly. This suggests that ASK (which nests CRT as a special case) may be preferable to CRB in those cases because of its lower computational demands. The advantages of adaptive sampling by Algorithm 1 become more pronounced the higher the correlation ρ.

An important point to note, in light of the results presented in this section and in anticipation of the application in Section 5, is that precise estimation of the log-likelihood is essential for inference. For example, it is crucial for computing likelihood ratio statistics, information criteria, marginal

Table 4: Log-probability estimates ($J = 12$) with numerical standard errors ($\times 10^{-2}$) in parentheses.

	ρ	AR	STERN	GHK	CRB	CRT	ASK	ARK
$\boldsymbol{\mu} = \mathbf{1}_4 \otimes \boldsymbol{\mu}_A$	-0.7	-5.914	-6.096	-6.084	-6.102	-6.101	-6.103	-6.129
		(19.219)	(3.775)	(1.207)	(0.170)	(0.162)	(0.180)	(3.428)
	-0.3	-5.599	-5.699	-5.696	-5.697	-5.697	-5.697	-5.685
		(16.409)	(1.220)	(0.412)	(0.040)	(0.037)	(0.038)	(1.898)
	0.3	-3.868	-3.961	-3.979	-3.979	-3.979	-3.978	-3.987
		(6.844)	(2.332)	(0.389)	(0.074)	(0.078)	(0.083)	(0.481)
	0.7	-2.429	-2.397	-2.410	-2.410	-2.417	-2.404	-2.365
		(3.217)	(2.326)	(0.836)	(0.628)	(0.909)	(0.747)	(2.305)
$\boldsymbol{\mu} = \mathbf{1}_4 \otimes \boldsymbol{\mu}_B$	-0.7	.	-14.504	-14.484	-14.516	-14.516	-14.516	.
		.	(8.462)	(1.864)	(0.049)	(0.046)	(0.047)	.
	-0.3	.	-11.091	-11.092	-11.094	-11.094	-11.094	.
		.	(1.734)	(0.547)	(0.020)	(0.019)	(0.019)	.
	0.3	-6.725	-6.858	-6.850	-6.852	-6.851	-6.851	-6.818
		(28.850)	(5.393)	(0.524)	(0.055)	(0.058)	(0.062)	(5.829)
	0.7	-3.821	-3.923	-3.914	-3.933	-3.944	-3.937	-3.943
		(6.683)	(4.651)	(1.213)	(0.540)	(0.709)	(0.762)	(3.553)
$\boldsymbol{\mu} = \mathbf{1}_4 \otimes \boldsymbol{\mu}_C$	-0.7	.	-31.980	-31.901	-31.945	-31.945	-31.945	.
		.	(13.751)	(2.411)	(0.013)	(0.012)	(0.013)	.
	-0.3	.	-19.684	-19.690	-19.693	-19.693	-19.693	.
		.	(2.327)	(0.656)	(0.009)	(0.009)	(0.009)	.
	0.3	.	-11.090	-10.860	-10.862	-10.862	-10.861	.
		.	(12.964)	(0.667)	(0.041)	(0.043)	(0.046)	.
	0.7	-5.776	-6.044	-5.959	-5.972	-5.974	-5.961	-6.003
		(17.933)	(11.969)	(1.718)	(0.428)	(0.642)	(0.588)	(7.573)

likelihoods and Bayes factors for model comparisons and hypothesis testing. Estimation efficiency is also key to mitigating simulation biases (due to Jensen's inequality and the non-linearity of the logarithmic transformation) in the maximum simulated likelihood estimation of parameters, standard errors, and confidence intervals (see, for example, McFadden and Train, 2000, Section 3, and Train, 2003, Chapter 10).

To summarize, the results suggest that the MCMC simulated likelihood estimation methods perform very well and dominate other estimation methods over a large set of possible scenarios. Their performance improves with the ability of the Markov chain to mix well, making Algoirthm 1 an important component of the estimation process.

4.1 Computational Caveats and Directions for Further Study

In this article we have compared a number of new and existing estimators for a fixed Monte Carlo simulation size. Such comparisons are easy to perform and interpret in practically any simulation study. However, an important goal for future research would be to optimize the code, perform formal operation counts, and study the computational intensity of each estimation algorithm. This would enable comparisons based on a fixed computational budget (running time), which are less straightforward and more difficult to generalize because they depend on various nuances of the specific application. In this section we highlight some of the subtleties that must be kept in mind.

For instance, although AR and ARK are simple and fast, the computational cost to achieve a certain estimation precision is entirely dependent on the context. Importance sampling and MCMC simulators such as GHK, CRT, CRB, and ASK, on the other hand, involve more coding and more costly iterations, but they are also more reliable and statistically efficient, especially in estimating small orthant probabilities. Based on rough operation counts, GHK, CRT and ASK involve comparable computations and simulations, while the efficiency of CRB depends on the number of reduced runs that is required.

The complications of optimizing these methods for speed, while retaining their transparency and reliability, go well beyond simply removing redundant computations (e.g. inversions, multiplications, conditional moment calculations) and making efficient use of storage. Although these steps are essential in producing efficient algorithms, another difficulty arises because random number generators may have to rely on a mix of techniques in order to be reliable and general. For example, to simulate truncated normal draws close to the mean, one can use the inverse cdf method. However, it is well known that the inverse cdf method can fail in the tails. Fortunately, in those circumstances the algorithms proposed in Robert (1995) are very efficient and reliable. Because in a given application the estimation algorithms may use a different mix of these simulation approaches, the computational times across algorithms may not be directly comparable.

Another caveat arises due to the specifics of algorithms that rely on MCMC samplers and has

to do with determining an appropriate way to account for the dual use (a benefit) and the need for burn-in sampling (a cost) of MCMC simulation. Specifically, in many instances MCMC draws have dual use in addition to evaluation of the likelihood function (e.g. for computing marginal effects, point elasticities, etc.) or are already available from an earlier step in the MCMC estimation (so likelihood estimation requires no simulation but only the computation and averaging of certain conditional densities and transition kernels). Similarly, the costs of burn-in simulation would typically not be an issue in Bayesian studies where a Markov chain would have already been running during the estimation stage, but could be an additional cost in hill-climbing algorithms. Of course, for well-mixing Markov chains convergence and burn-in costs are trivial, but should otherwise be properly accounted into the cost of MCMC simulation.

These special considerations are compounded by having to examine the estimators in the context of different dimensionality, mean, and covariance matrix combinations, making a thorough computational examination of the methods an important and necessary next step in this area of research. Gauss programs for the new estimators are available on the authors' websites.

5 Application to a Model for Binary Panel Data

This section offers an application of the techniques to the problem of likelihood ordinate estimation in models for binary panel data. In particular, we consider data from Chib and Jeliazkov (2006) that deals with the intertemporal labor force participation decisions of 1545 married women in the age range of 17-66. The data set, obtained from the Panel Study of Income Dynamics, contains a panel of women's working status indicators (1 = working during the year, 0 = not working) over a 7 year period (1979-1985), together with a set of 7 covariates given in Table 5. The sample consists of continuously married couples where the husband is a labor force participant (reporting both positive earnings and hours worked) in each of the sample years. Similar data have been analyzed by Chib and Greenberg (1998), Avery et al. (1983), and Hyslop (1999) using a variety of techniques.

We considered two competing specifications that differ in their dynamics. For $i = 1, \ldots, n$ and

Table 5: Variables in the women's labor force participation application.

Variable	Explanation	Mean	SD
WORK	wife's labor force status (1=working, 0=not working)	0.7097	0.4539
INT	an intercept term (a column of ones)		
AGE	the woman's age in years	36.0262	9.7737
RACE	1 if black, 0 otherwise	0.1974	0.3981
EDU	attained education (in years) at time of survey	12.4858	2.1105
CH2	number of children aged 0-2 in that year	0.2655	0.4981
CH5	number of children aged 3-5 in that year	0.3120	0.5329
INC	total annual labor income of head of household[a]	31.7931	22.6417

[a]Measured as nominal earnings (in thousands) adjusted by the consumer price index (base year 1987).

$t = 1, \ldots, T$, the first specification, model $\mathcal{M}_1$, is given by

$$y_{it} = 1\left\{\tilde{\boldsymbol{x}}_{it}'\boldsymbol{\delta} + \boldsymbol{w}_{it}'\boldsymbol{\beta}_i + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it} > 0\right\}, \qquad \varepsilon_{it} \sim \mathcal{N}(0,1),$$

and captures state dependence through two lags of the dependent variable but does not involve serial correlation in the errors. The second specification, model $\mathcal{M}_2$, involves only a single lag of the dependent variable, but allows for AR(1) serial correlation in the errors:

$$y_{it} = 1\left\{\tilde{\boldsymbol{x}}_{it}'\boldsymbol{\delta} + \boldsymbol{w}_{it}'\boldsymbol{\beta}_i + g(s_{it}) + \phi_1 y_{i,t-1} + \varepsilon_{it} > 0\right\}, \qquad \varepsilon_{it} = \rho\varepsilon_{i,t-1} + \nu_{it}, \qquad \nu_{it} \sim \mathcal{N}(0,1).$$

Both $\mathcal{M}_1$ and $\mathcal{M}_2$ include mutually exclusive sets of covariates $\tilde{\boldsymbol{x}}_{it}$ and $\boldsymbol{w}_{it}$, where the effects of the former, $\boldsymbol{\delta}$, are modeled as common across women, and the effects of the latter, $\boldsymbol{\beta}_i$, are individual-specific (random); the models also include a covariate $\boldsymbol{s}_{it}$ whose effect is modeled through an unknown function $g(\cdot)$ which is estimated nonparametrically. In both specifications $y_{it} = WORK_{it}$, $\tilde{\boldsymbol{x}}_{it}' = (RACE_i, EDU_{it}, \ln(INC_{it}))$, $s_{it} = AGE_{it}$, $\boldsymbol{w}_{it}' = (1, CH2_{it}, CH5_{it})$, and heterogeneity is modeled in a correlated random effects framework which allows $\boldsymbol{\beta}_i$ to be correlated with observables through

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\gamma} + \boldsymbol{b}_i, \qquad \boldsymbol{b}_i \sim N_3(0, \boldsymbol{D}). \tag{15}$$

We let all three random effects depend on the initial conditions and the effects of CH2 and CH5

also depend on average husbands' earnings through

$$\boldsymbol{A}_i = \begin{pmatrix} \bar{y}_{i0} & & & & & \\ & 1 & \bar{y}_{i0} & \overline{\ln(INC_i)} & & \\ & & & & 1 & \bar{y}_{i0} & \overline{\ln(INC_i)} \end{pmatrix},$$

where neither $\tilde{\boldsymbol{x}}_{it}$ nor the first row of $\boldsymbol{A}_i$ involves a constant term because the unknown function $g(\cdot)$ is unrestricted and absorbs the overall intercept.

Upon substituting (15) into each of the two models, stacking the observations over time for each woman, and grouping the common and individual-specific terms, we can write models $\mathcal{M}_1$ and $\mathcal{M}_2$, similarly to (1), in the latent variable form

$$\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{g}_i + \boldsymbol{\varepsilon}_i \tag{16}$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iT})'$, $\boldsymbol{X}_i = \left(\tilde{\boldsymbol{X}}_i : \boldsymbol{W}_i\boldsymbol{A}_i : \boldsymbol{L}_i\right)$, $\tilde{\boldsymbol{X}}_i = (\tilde{\boldsymbol{x}}_{i1}, \ldots, \tilde{\boldsymbol{x}}_{iT})'$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT})'$, $\boldsymbol{L}_i$ contains the requisite lags of y_{it}, $\boldsymbol{\beta} = \left(\boldsymbol{\delta}', \boldsymbol{\gamma}', \boldsymbol{\phi}'\right)'$, and $\boldsymbol{g}_i = \left(g\left(s_{i1}\right), \ldots, g\left(s_{iT_i}\right)\right)'$. The errors $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \ldots, \varepsilon_{iT}\right)'$ follow the distribution $\boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}_i\right)$, where $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma} + \boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'$ and $\boldsymbol{\Sigma}$ is the Toeplitz matrix implied by the autoregressive process, i.e., $\boldsymbol{\Sigma} = \boldsymbol{I}_T$ for model $\mathcal{M}_1$ and $\boldsymbol{\Sigma}\left[j,k\right] = \rho^{|j-k|}/\left(1-\rho^2\right)$ for model $\mathcal{M}_2$. Because $\mathcal{M}_1$ requires two lags of the dependent variable, both models are estimated conditionally on the initial two observations in the data.

Our focus in this section is on the problem of estimating the log-likelihood ordinate conditionally on the model parameters. For details on the estimation of the parameters in the two models, interested readers are referred to Chib and Jeliazkov (2006). As the model construction shows, both the cluster means and covariance matrices depend on cluster-specific characteristics, and hence the panel data setup with multidimensional heterogeneity is quite useful for examining the performance of the estimators in a variety of possible circumstances that occur in the data.

Estimates of the log-likelihood function obtained by various methods are presented in Table 6. The log-likelihood and numerical standard error (NSE) estimates were obtained from runs of length 10 000 draws for each likelihood contribution (there are $n = 1545$ likelihood contributions, one for each woman in the sample). The NSEs of the log-likelihood contribution estimates are presented in

Table 6: Log-likelihood estimates in the women's labor force participation application.

	Model $\mathcal{M}_1$		Model $\mathcal{M}_2$	
Estimator	Log-likelihood	NSE	Log-likelihood	NSE
Traditional Estimators				
Stern	-2501.435	(0.291)	-2537.926	(0.573)
GHK	-2501.434	(0.100)	-2537.631	(0.137)
AR	-2502.005	(2.355)	-2540.702	(2.440)
MCMC-Based Estimators				
CRB	-2501.425	(0.027)	-2537.593	(0.061)
CRT	-2501.403	(0.039)	-2537.572	(0.081)
ASK	-2501.411	(0.036)	-2537.563	(0.073)
ARK	-2501.498	(0.090)	-2537.898	(0.202)

Figures 4 and 5. The results in Table 6 and Figures 4 and 5 show that in this application, the new MCMC methods are more efficient than existing approaches. While the argument can be made that the higher efficiency of CRB is due to its reliance on additional reduced runs, the results also reveal that the remaining MCMC methods are also generally more efficient even though they do not require any reduced run simulations. We can also see that the improvement in MCMC sampling due to Algorithm 1 used in the ASK method leads to lower standard errors relative to CRT. A much more striking improvement in efficiency, however, can be seen in a comparison between the AR and ARK methods. What makes the comparison impressive is that both methods are based on the same simulated draws (with the AR estimate being obtained as a by-product of ARK estimation), yet ARK is orders of magnitude more efficient.

Comparison of the estimates for models $\mathcal{M}_1$ and $\mathcal{M}_2$ shows that allowing for autocorrelated errors (the estimated value of ρ is -0.29), at the cost of excluding a second lag of y_{it} from the mean, has a detrimental effect on the efficiency of all estimators. While the relative efficiency rankings of estimators are largely preserved as we move from $\mathcal{M}_1$ to $\mathcal{M}_2$ (with the exception of GHK and ARK), traditional methods appear to exhibit more high-variability outliers, whereas MCMC-based methods show a general increase in variability of estimation across all clusters (the plots for ARK, similarly to those of AR, shows both features).

This section has considered the application of several simulated likelihood estimation techniques

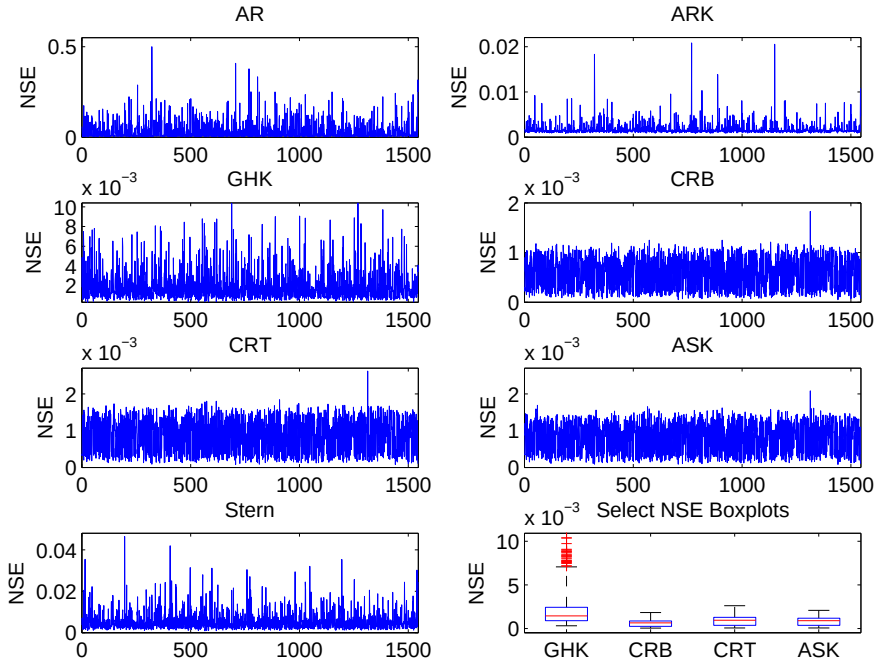

Figure 4: Numerical standard error (NSE) estimates for Model $\mathcal{M}_1$.

to a hierarchical semiparametric models for binary panel data with state dependence, serially correlated errors, and multidimensional heterogeneity correlated with the covariates and initial conditions. Using data on women's labor force participation, we have illustrated that the proposed MCMC-based estimation methods are practical and can lead to improved efficiency of estimation in a variety of environments occurring in a real-world data set. Comparisons of these and other simulated likelihood estimators in other model settings is an important item for future research.

6 Conclusions

This article has considered the problem of evaluating the likelihood functions in a broad class of multivariate discrete data models. We have reviewed traditional simulation methods that produce continuous and differentiable estimates of the response probability and can be used in hill-climbing algorithms in maximum likelihood estimation. We have also shown that the problem can be handled by MCMC-based methods designed for marginal likelihood computation in Bayesian econometrics. New computationally efficient and conceptually straightforward MCMC algorithms have been de-

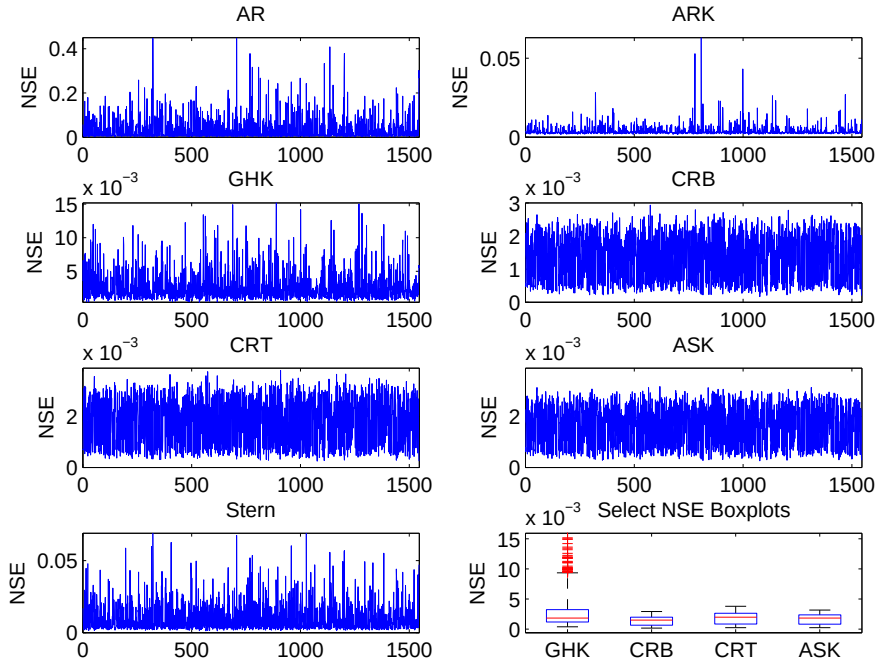

Figure 5: Numerical standard error (NSE) estimates for Model $\mathcal{M}_2$.

veloped for (i) estimating response probabilities and likelihood functions and (ii) simulating draws from multivariate truncated normal distributions. The former of these contributions aims to provide simple, efficient, and sound solutions from Markov chain theory to outstanding problems in simulated likelihood estimation; the latter is motivated by the need to provide high-quality samples from the target multivariate truncated normal density. A simulation study has shown that the methods perform well, while an application to a correlated random effects panel data model of women's labor force participation shows that they are practical and easy to implement.

In addition to their simplicity and efficiency, an important advantage of the methods considered here is that they are modular and can be mixed and matched as components of composite estimation algorithms in a variety of multivariate discrete and censored data settings. Important topics for future work in this area would be to examine the effectiveness of the estimators in practical applications, to explore extensions and develop additional hybrid approaches, and to perform detailed computational efficiency studies in a range of contexts.

References

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

Andrews, D. F. and Mallows, C. L. (1974), "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society – Series B*, 36, 99–102.

Avery, R., Hansen, L., and Hotz, V. (1983), "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24, 21–35.

Bhat, C. R. (2001), "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," *Transportation Research Part B*, 35, 677–693.

Bhat, C. R. (2003), "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences," *Transportation Research Part B*, 37, 837–855.

Börsch-Supan, A. and Hajivassiliou, V. A. (1993), "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics*, 58, 347–368.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S. and Greenberg, E. (1996), "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 12, 409–431.

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

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

Chib, S. and Jeliazkov, I. (2005), "Accept-Reject Metropolis-Hastings Sampling and Marginal Likelihood Estimation," *Statistica Neerlandica*, 59, 30–44.

Chib, S. and Jeliazkov, I. (2006), "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data," *Journal of the American Statistical Association*, 101, 685–700.

Damien, P. and Walker, S. G. (2001), "Sampling Truncated Normal, Beta, and Gamma Densities," *Journal of Computational and Graphical Statistics*, 10, 206–215.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.

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 Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society – Series B*, 56, 501–514.

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.

Geweke, J. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," in *Computing Science and Statistics*, ed. E. M. Keramidas, Proceedings of the Twenty-Third Symposium on the Interface, pp. 571–578, Fairfax: Interface Foundation of North America, Inc.

Geweke, J. (1993), "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8, S19–S40.

Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication," *Econometric Reviews*, 18, 1–73.

Goffe, W. L., Ferrier, G. D., and Rogers, J. (1994), "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65–99.

Greenberg, E. (2008), *Introduction to Bayesian Econometrics*, Cambridge University Press, New York.

Griffiths, W. E., Hill, R. C., and O'Donnell, C. J. (2006), "A Comparison of Bayesian and Sampling Theory Inferences in a Probit Model," in *Essays in Honor of Stanley R. Johnson*, eds. M. Holt and J.-P. Chavas, article 12, available at http://www.bepress.com/sjohnson/.

Gu, Y., Fiebig, D. G., Cripps, E., and Kohn, R. (2009), "Bayesian Estimation of a Random Effects Heteroscedastic Probit Model," *Econometrics Journal*, 12, 324–339.

Hajivassiliou, V. A. and McFadden, D. (1998), "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica*, 66, 863–896.

Hajivassiliou, V. A. and Ruud, P. (1994), "Classical Estimation Methods for LDV Models Using Simulation," *Handbook of Econometrics*, 4, 2383–2441.

Hajivassiliou, V. A., McFadden, D. L., and Ruud, P. (1996), "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives: Theoretical and Computational Results," *Journal of Econometrics*, 72, 85–134.

Heiss, F. and Winschel, V. (2008), "Likelihood Approximation by Numerical Integration on Sparse Grids," *Journal of Econometrics*, 144, 62–80.

Hyslop, D. (1999), "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294.

Kass, R. and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Keane, M. (1994), "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, pp. 95–116.

Kennedy, J. and Eberhart, R. C. (2001), *Swarm Intelligence*, Morgan Kaufmann, San Francisco, California.

Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons, New York.

Lerman, S. and Manski, C. (1981), "On the use of Simulated Frequencies to Approximate Choice Probabilities," *Structural Analysis of Discrete Data with Econometric Applications*, pp. 305–319.

McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometica*, 57, 995–1026.

McFadden, D. and Train, K. (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 447–470.

Meng, X.-L. and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Neal, R. M. (2003), "Slice Sampling," *The Annals of Statistics*, 31, 705–767.

Newton, M. A. and Raftery, A. E. (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society – Series B*, 56, 3–48.

Poirier, D. J. (1978), "A Curious Relationship between Probit and Logit Models," *Southern Economic Journal*, 44, 640–641.

Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.

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

Stern, S. (1992), "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models," *Econometica*, 60, 943–952.

Stern, S. (1997), "Simulation-based Estimation," *Journal of Economic Literature*, 35, 2006–2039.

Tanner, M. A. and Wong, W. H. (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," *Annals of Statistics*, 22, 1701–1761.

Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.