MCMC Methods for Fitting and Comparing Multinomial Response Models

Siddhartha Chib
John M. Olin School of Business
Washington University
One Brookings Drive
St. Louis, MO 63130

Edward Greenberg
Department of Economics
Washington University
One Brookings Drive
St. Louis, MO 63130

Yuxin Chen
John M. Olin School of Business
Washington University
One Brookings Drive
St. Louis, MO 63130

January 20, 1998

Abstract

This paper is concerned with statistical inference in multinomial probit, multinomial-t and multinomial logit models. New Markov chain Monte Carlo (MCMC) algorithms for fitting these models are introduced and compared with existing MCMC methods. The question of parameter identification in the multinomial probit model is readdressed. Model comparison issues are also discussed and the method of Chib (1995) is utilized to find Bayes factors for competing multinomial probit and multinomial logit models. The methods and ideas are illustrated in detail with an example.

Keywords: Bayes factor; Gibbs sampling; Monte Carlo EM algorithm; Marginal likelihood; Metropolis-Hastings algorithm; Multinomial logit; Multinomial probit; Multinomial-t; Model comparison.

1 Introduction

The fitting of multinomial probit models has been viewed as a challenge for over twenty five years. One major difficulty is the problem of evaluating the likelihood function while another, somewhat neglected one, is the problem of estimating covariance parameters of the model given that only outcome per subject is observed. As a result of this missingness, which is inherent in multinomial data, it is possible that different combinations of regression and covariance parameters can produce virtually identical outcome probabilities.

Recently, developments in simulation-based Bayesian and classical methods have given rise to reasonably effective methods for estimating this model [McFadden (1989), Albert and Chib (1993), McCulloch, Polson, and Rossi (1994) and Stern (1997)]. Despite these developments, further improvements in the fitting of the model are possible, based on Markov chain Monte Carlo methods [Gelfand and Smith (1990), Chib and Greenberg (1996)].

In general terms, Markov chain simulation methods provide a rather attractive framework for dealing with the MNP and related multinomial models. The use of these methods in the context of probit models was initiated by Albert and Chib (1993). A central reason for studying these methods is that they are easy to implement and can be applied from both a classical and Bayesian perspective. One version of these methods can be used to sample the posterior distribution of the parameters, while another can be used to search for the maximum-likelihood estimate. As a bonus, these methods can be extended for the fitting of more general multinomial models than the MNP. One such model that is introduced in this paper, the multinomial-t, relies on a multivariate-t assumption for the latent data. It turns out that the basic algorithms have to be modified only slightly to apply to this model.

In addition to tackling the question of fitting the MNP model, another purpose of this paper is to develop a framework within which alternative multinomial models can be compared. This framework is important because there is a paucity of discussion in the literature on the practical benefits of the MNP model over the much simpler multinomial logit (MNL) model. Although it is well known that the MNL model suffers from a weakness not shared with the MNP model—that the ratio of probabilities of any two outcomes does not depend on the presence or absence of other outcomes—it appears that the importance of this weakness has not been assessed in empirical settings. One reason for this may be that the comparison of these non-nested models is difficult from a classical perspective. From a Bayesian viewpoint, however, such comparisons can be handled more conveniently. For a specified set of priors, a method due to Chib (1995) can be used to calculate the marginal likelihood of the model and the Bayes factor, which is used in the Bayesian context to compare models. We apply this technique to a data set and find that the support for the MNL model over both the MNP and MNT models is decisive. Moreover, the Bayes factor supports the MNL model in another example, which we do not report on, in which the data

are artificially generated from the MNP model. This result is possibly an artifact of the covariates and our design, but it nonetheless emphasizes the important point that support for the MNP model over the MNL model is not guaranteed once model complexity is taken into account.

The rest of the paper is organized as follows. In Section 2 the various multinomial models are described, and in Section 3 two new MCMC algorithms for fitting the MNP and related models are presented. This section also discusses the issues related to identification and points out why the parameters of the MNP model are likely to be weakly identified. Section 4 explains the computation of the marginal likelihood and Bayes factors and considers an application that involves the comparison of MNP, MNT, and MNL models. Results from a real data set are introduced at various places in the text to illustrate the methods. Concluding remarks are contained in Section 5.

2 Multinomial response models

Let $y_1, y_2, \ldots, y_n$ denote a set of unordered multinomial responses on n randomly selected subjects. Assume that each response takes the possible values $\{1, \ldots, J+1\}$ and that each subject and response is associated with a set of covariates $(\mathbf{v}_{ij}, \mathbf{w}_i)$, where $\mathbf{v}_{ij}$ is a covariate vector that varies across both subjects and responses and $\mathbf{w}_i$ contains characteristics of subject i. To formulate the probability model for the $J+1$ responses, let $\mathbf{Z}_i^* = (z_{i1}^*, \ldots, z_{iJ+1}^*)$ denote a continuous latent vector with multivariate distribution F and let $y_i = j$ if z_{ij}^* is the maximum of $\mathbf{Z}_i^*$.

2.1 Multinomial probit

If F is the multivariate normal distribution, the probabilities of the multinomial responses can be defined in terms of the vector $\mathbf{Z}_i = (z_{i1}, \ldots, z_{iJ})$, where $z_{ij} = z_{ij}^* - z_{iJ+1}^*$ are differences with respect to an arbitrarily chosen base, z_{iJ+1}^*. The probability model of z_{ij} is given by

$$\begin{aligned} z_{ij} &= (\mathbf{v}_{ij} - \mathbf{v}_{i,J+1})'\boldsymbol{\delta} + \mathbf{w}_i'\boldsymbol{\gamma}_j + \varepsilon_{ij} \\ &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \end{aligned}$$

or in vector notation as $\mathbf{Z}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\beta} = (\boldsymbol{\delta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)$,

$$\mathbf{X}_i = \begin{pmatrix} (\mathbf{v}_{i1} - \mathbf{v}_{i,J+1})' & \mathbf{w}_i' & \mathbf{0}' & \cdots & \mathbf{0}' \\ (\mathbf{v}_{i2} - \mathbf{v}_{i,J+1})' & \mathbf{0}' & \mathbf{w}_i' & \mathbf{0}' & \cdots \\ \vdots & \vdots & & \ddots & \vdots \\ (\mathbf{v}_{iJ} - \mathbf{v}_{i,J+1})' & \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{w}_i' \end{pmatrix},$$

and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})' \sim \mathcal{N}_J(\mathbf{0}, \boldsymbol{\Sigma})$. For identifiability reasons the $(1,1)$ element of $\boldsymbol{\Sigma}$, σ_{11}, is constrained to equal one and γ_{J+1} is normalized to zero. In terms of the latent values z_{ij}, the observed outcome is given by the conditions

$$Y_i = \begin{cases} j & \text{if } z_{ij} = \max\{\mathbf{Z}_i, 0\} \\ J+1 & \text{if } \max_l\{z_{il}\} \leq 0, \end{cases} \tag{1}$$

and the probability mass function of Y_i is

$$\Pr(Y_i = j|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{A_j} \phi_J(\mathbf{Z}_i|\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})\, d\mathbf{Z}_i, \quad j \leq J+1,$$

where ϕ_J is the density function of the J-variate normal distribution and

$$A_j = \begin{cases} \{\mathbf{Z}_i : z_{i1} < z_{ij}, \ldots, 0 < z_{ij},\ z_{i,j+1} < z_{ij}, \ldots, z_{iJ} < z_{ij}\}, & j \leq J \\ \{\mathbf{Z}_i : z_{i1} < 0, \ldots, z_{ij} < 0, \ldots z_{iJ} < 0\}, & j = J+1\,. \end{cases}$$

The multinomial probabilities thus require the computation of a complicated multivariate integral. One way to compute the integral is by the Monte Carlo importance sampling method developed by Geweke (1991), Hajivasilliou (1990), and Keane (1994), and known as the GHK method (see Appendix 1 for further details). For estimation purposes, it is not necessary to compute this probability, as is discussed below.

2.2 Multinomial-t

Now suppose that the distribution F on the underlying undifferenced latent values $\mathbf{Z}_i^*$ is multivariate-t with specified degrees of freedom ν. This gives rise to a model that we call the multinomial-t model. Albert and Chib (1993) extended the probit link to the t-link in the binary response case and provided a simple approach for estimating the resulting model. As in the MNP case, the MNT model can be expressed in terms of the differenced latent values $\mathbf{Z}_i$, where now $\mathbf{Z}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \text{MVT}_J(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$ with density

$$f(\mathbf{Z}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \left\{1 + \frac{1}{\nu}(\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}^{-(v+J)/2}.$$

As before, $\sigma_{11} = 1$ and the observed outcomes Y_i are defined by (1). Following Albert and Chib (1993), the model for the latent $\mathbf{Z}_i$ may be expressed as a scale mixture of normals by introducing a random variable $\lambda_i \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ and letting

$$\mathbf{Z}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \lambda_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Sigma}).$$

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

2.3 Multinomial logit

The assumption that F is the independent Weibull distribution leads to the multinomial logit model of McFadden (1975). The probabilities of the outcomes are given in closed form as

$$\Pr(Y_i = j|\boldsymbol{\beta}) = \frac{\exp(\mathbf{v}_{ij}'\boldsymbol{\delta} + \mathbf{w}_i'\boldsymbol{\gamma}_j)}{\sum_{j=1}^{J+1}\exp(\mathbf{v}_{ij}'\boldsymbol{\delta} + \mathbf{w}_i'\boldsymbol{\gamma}_j)}, \quad j = 1, \ldots, J+1\,,$$

with $\boldsymbol{\gamma}_{J+1}$ normalized to zero. Note that the covariates $\mathbf{v}_{ij}$ appear in their undifferenced form.

3 MCMC fitting of MNP models

Markov chain Monte Carlo simulation methods provide a unified and coherent approach for estimating the multinomial response models in Section 2. We focus exclusively on these methods and refer the reader to McFadden (1989), Geweke, Keane, and Runkle (1994) and Stern (1997) for classical non-MCMC simulation-based methods. We focus our attention on the estimation of the MNP model because the estimation of the MVT model requires only a straightforward modification of the MNP algorithms and the estimation of the MNL model by MCMC presents no new difficulties. An algorithm for fitting it is described briefly.

3.1 Posterior sampling in the MNP model

One basic approach for fitting probit models by MCMC methods is due to Albert and Chib (1993). In this approach, the parameter space is augmented by the latent data $\{\mathbf{Z}_i\}$, and MCMC methods are used to sample the joint posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\sigma}, \{\mathbf{Z}_i\}|\mathbf{y})$, where $\boldsymbol{\sigma} = (\sigma_{12}, \sigma_{22}, \sigma_{31}, \ldots, \sigma_{JJ})$ denotes the unique elements of $\boldsymbol{\Sigma}$ and $\mathbf{y} = (y_1, \ldots, y_n)$ is the observed data. The MCMC algorithm is quite straightforward except that the constraint

on σ_{11} makes it difficult to sample $\boldsymbol{\sigma}$. To solve this problem, McCulloch and Rossi (1994) propose an algorithm that ignores the restriction on σ_{11} in the sampling. Their algorithm simulates the non-identified parameters of the model, obtaining draws of the identified parameters ex-post from the draws of the non-identified parameters. Nobile (1995) has pointed out that, as a consequence of sampling the non-identified parameters, this method is sensitive to the prior distribution.

To sample the identified parameters in a MCMC simulation with data augmentation, one iterates on the following steps a large number of times.

Basic algorithm for sampling the MNP posterior distribution

- Sample z_{ij} from $\pi(z_{ij}|y_i, \mathbf{Z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, for $j = 1, 2, \ldots, J$ and $i = 1, 2, \ldots, n$, where $\mathbf{Z}_{i(-j)}$ is the vector $\mathbf{Z}_i$ excluding the jth component;
- Sample $\boldsymbol{\beta}$ from $\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}, \boldsymbol{\Sigma})$; and
- Sample $\boldsymbol{\sigma}$ from $\pi(\boldsymbol{\sigma}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{Z})$.

We now explain how each of these distributions can be sampled. The distributions in the first step of this algorithm are the univariate normal distributions $f(z_{ij}|\mathbf{Z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ truncated to the region implied by the observed value of y_i:

$$\pi(z_{ij}|\mathbf{Z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = f(z_{ij}|\mathbf{Z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) I(z_{ij} \in R_{ij}), \tag{2}$$

where

$$R_{ij} = \begin{cases} (\max\{0, \max\{\mathbf{Z}_{i(-j)}\}\}, \infty) & \text{if} \quad y_i = j, \quad j = 1, \ldots, J \\ (-\infty, \max\{\mathbf{Z}_{i(-j)}\}) & \text{if} \quad y_i \neq j, \quad j = 1, \ldots, J \\ (-\infty, 0] & \text{if} \quad y_i = J + 1 \end{cases},$$

which follows from the set-valued inverse of the mapping in (1). The density $f(z_{ij}|\mathbf{Z}_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is obtained by the usual multivariate normal theory. Instead of sampling the z_{ij} in this manner, the entire vector $\mathbf{Z}_i$ can be sampled from $\pi(\mathbf{Z}_i|y_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ by the accept-reject method [Albert and Chib (1993)]. In this approach the vector $\mathbf{Z}_i$ is drawn from $N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and accepted as a valid draw if the vector falls in the region implied by y_i. The advantages of this method are that it requires little coding and that it tends to improve the serial correlation of the sampled output because the $\mathbf{Z}_i$ are drawn in one block. A disadvantage is

that several sampled vectors may have to be discarded before one is accepted. Nonetheless, because the accept-reject method is not a Markov chain sampler, the method is useful in initializing the Markov chain simulations for the latent data.

The next two distributions are proportional to the complete data density

$$\begin{aligned} f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\Sigma}) &= \prod_{i=1}^{n} f(\mathbf{Z}_i|\boldsymbol{\beta},\boldsymbol{\Sigma}) \\ &\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_i (\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})\right). \end{aligned} \tag{3}$$

If the prior density of $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta}_0, \mathbf{B}_0^{-1})$, then $\boldsymbol{\beta}|\mathbf{y},\mathbf{Z},\boldsymbol{\Sigma}$, which is independent of $\mathbf{y}$, is $N(\boldsymbol{\beta}_1, \mathbf{B}_1^{-1})$, where $\boldsymbol{\beta}_1 = \mathbf{B}_1^{-1}[\mathbf{B}_0\boldsymbol{\beta}_0 + \sum_i \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i]$ and $\mathbf{B}_1 = [\mathbf{B}_0 + \sum_i \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i]$. Because this distribution is in closed form, the conditional mean and variance can be used to estimate the posterior mean and variance of $\pi(\boldsymbol{\beta}|\mathbf{y})$ by averaging the conditional moments. For example, given that the simulation has been run for G iterations, the posterior mean of $\boldsymbol{\beta}$ can be estimated as $G^{-1}\sum_{g=1}^{G} \mathbf{B}_1^{-1}[\mathbf{B}_0\boldsymbol{\beta}_0 + \sum_i \mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_i^{(g)}]$, where $\mathbf{Z}_i^{(g)}$ denotes the sampled latent data in the gth iteration. The corresponding marginal variance can be estimated from the relationship between conditional and unconditional variance.

Two ways of simulating the third distribution are considered next. The first is based on a Choleski decomposition, and the second utilizes properties of the Wishart distribution.

Choleski decomposition method

Our new approach to the simulation of $\boldsymbol{\Sigma}$ is based on the log-Choleski decomposition of a positive-definite matrix [Pinheiro and Bates (1996)]. For any symmetric or lower triangular matrix $\mathbf{A}$ with $a_{11} = 1$, let

$$\text{vech}^*(\mathbf{A}) = (a_{12}, a_{22}, a_{31}, \ldots, a_{J1}, \ldots, a_{JJ})'$$

denote the free elements of $\mathbf{A}$. Now let $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L} = \{l_{rs}\}$ is a $J \times J$ lower triangular matrix with $l_{11} = 1$, and let $\boldsymbol{\psi} = \text{vech}^*(\mathbf{L})$, a vector of dimension $p^* = (J+2)(J-1)/2$. The matrix $\mathbf{L}$ is made unique by restricting its diagonal elements to be positive; these appear in rows $(i+2)(i-1)/2$ of $\boldsymbol{\psi}$, $i \leq 2 \leq J$. Finally, define the parameter

$$\begin{aligned} \boldsymbol{\theta} &= (l_{21}, \log(l_{22}), l_{31}, l_{32}, \log(l_{33}), \ldots, l_{J1}, \ldots, \log(l_{JJ}))' \\ &\equiv (\theta_1, \ldots, \theta_{p^*})'. \end{aligned}$$

The mapping between $\mathbf{\Sigma}$ and $\boldsymbol{\theta}$ is one-to-one. This parameterization of $\mathbf{\Sigma}$ leaves the vector $\boldsymbol{\theta}$ entirely *unrestricted.* Any $\boldsymbol{\theta} \in R^{p^*}$ leads to a matrix $\mathbf{\Sigma}$ that is symmetric, positive definite, and has $\sigma_{11} = 1$.

To understand the nature of this parameterization, consider the case $J = 2$, where

$$\mathbf{L} = \begin{pmatrix} 1 & 0 \\ l_{21} & l_{22} \end{pmatrix}.$$

From $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}'$ it follows that $\sigma_{12} = l_{12}$ and $\sigma_{22} = l_{12}^2 + l_{22}^2$. These imply that $l_{22}^2 = \sigma_{22} - \sigma_{12}^2$, which is the determinant of $\mathbf{\Sigma}$ and is positive if $\mathbf{\Sigma}$ is positive definite. Thus, the parameterization $\boldsymbol{\theta} = (l_{12}, \log(l_{22}))$ imposes the required properties of positive definiteness along with the condition that $\sigma_{11} = 1$.

A major advantage of the $\boldsymbol{\theta}$ parameterization from a Bayesian perspective is that it permits a straightforward use of MCMC methods. Furthermore, a prior distribution on $\boldsymbol{\theta}$ can be assigned by specifying a prior distribution on each σ_{ij} and then using this prior distribution to infer the required distribution of $\boldsymbol{\theta}$. To illustrate this idea, suppose that our prior beliefs about $\text{vech}^*(\mathbf{\Sigma})$ are proportional to a normal distribution with mean vector $\mathbf{s}_0$ and covariance matrix $\mathbf{S}_0$, as in Chib and Greenberg (1995b). The required prior on $\boldsymbol{\theta}$ can be determined by the following Monte Carlo procedure:

1. Set $i = 1$

   (a) While i is less than I (a prespecified quantity), sample a vector $\text{vech}^*(\mathbf{\Sigma})^i \propto N(\mathbf{s}_0, \mathbf{S}_0)$ and form the matrix $\mathbf{\Sigma}^i = \mathbf{L}^i\mathbf{L}^{i'}$. From $\mathbf{L}^i$ compute and store the vector $\boldsymbol{\theta}^i$.

   (b) Increment i and go to (1a).

2. Compute $\mathbf{v}_0 = I^{-1}\sum_{i=1}^{I}\boldsymbol{\theta}^i$ and $\mathbf{G}_0 = I^{-1}\sum_{i=1}^{I}(\boldsymbol{\theta}^i - \mathbf{v}_0)(\boldsymbol{\theta}^i - \mathbf{v}_0)'$, the mean and covariance of $\{\boldsymbol{\theta}^i\}$. Let the prior distribution of $\boldsymbol{\theta}$ be $N(\mathbf{v}_0, \mathbf{G}_0)$.

Note that the above prior on $\{\sigma_{ij}\}$ overcomes the well known limitation of the Wishart distribution wherein the spread of the distribution is controlled by a single scalar degrees of freedom parameter. A notable advantage of working in the $\boldsymbol{\theta}$ parameterization is that it

leads to a unrestricted posterior density. In contrast, the posterior density of vech$^*(\boldsymbol{\Sigma})$ is restricted to the region that produces a positive-definite matrix.

Now consider the sampling of $\boldsymbol{\theta}$ (equivalently the sampling of $\boldsymbol{\Sigma}$) from the density $\pi(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\beta})$. By definition the full conditional density is

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\beta}) &\propto \pi(\boldsymbol{\theta})\prod_{i=1}^{n}\phi(\mathbf{Z}_i|\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}) \\ &\propto \pi(\boldsymbol{\theta})f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\theta}), \quad \boldsymbol{\theta}\in\mathcal{R}^{p^*}, \end{aligned} \tag{4}$$

where $\pi(\boldsymbol{\theta})$ is the unnormalized Gaussian prior density for $\boldsymbol{\theta}$ and the value of the normalizing constant is not required. This posterior density can be sampled by the MH algorithm with a tailored proposal density. Tailoring is achieved by finding the mode and curvature of $\log f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\theta})$ from a few Newton-Raphson steps. The mode and curvature are then used to create a multivariate-t proposal density, $f_T(\boldsymbol{\theta}|\boldsymbol{\mu},\tau\mathbf{V},\nu)$, where μ is the mode, $\mathbf{V}$ is the inverse of the negative Hessian at the mode, and τ and ν are adjustable parameters. With $\boldsymbol{\theta}$ denoting the current point in the iterations, the MCMC algorithm proceeds by iterating on the following steps.

Algorithm MNP 1

- Sample $\mathbf{Z}$ as in the basic algorithm for sampling the MNP posterior distribution;
- Sample $\boldsymbol{\beta}$ as in the basic algorithm for sampling the MNP posterior distribution;
- Sample $\boldsymbol{\theta}^t$ from $f_T(.|\boldsymbol{\mu},\tau\mathbf{V},\nu)$ and compute

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^t)=\min\left\{1,\frac{f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\theta}^t)}{f(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\theta})}\frac{f_T(\boldsymbol{\theta}|\boldsymbol{\mu},\tau\mathbf{V},\nu)}{f_T(\boldsymbol{\theta}^t|\boldsymbol{\mu},\tau\mathbf{V},\nu)}\right\}.$$

  Move to $\boldsymbol{\theta}^t$ with probability $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^t)$ and stay at $\boldsymbol{\theta}$ with probability $1-\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^t)$.

It should be noted that this algorithm is easily modified if the covariance matrix has more constraints than $\sigma_{11}=1$. In that case one can operate directly on the unique elements of $\boldsymbol{\Sigma}$, as in Chib and Greenberg (1995b) in a different but related context. This point is illustrated in one of the examples considered below.

Posterior sampling without augmentation

Algorithm 1 exploits the simplification that arises from data augmentation. One question is whether it is possible to sample the posterior distribution without augmentation. The main problem (one that is avoided by data augmentation) is that it is necessary to compute the likelihood function at least once during each point in the iterations. This can be a prohibitive computational burden if the sample size and the number of alternatives are large. In the case of smaller models, however, one may proceed as follows.

Let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ denote the parameters of the model and consider sampling $\boldsymbol{\psi}$ in one block with the MH algorithm. To find the proposal density for $\boldsymbol{\psi}$ one can utilize the output of Algorithm 1. Specifically, one can run Algorithm 1 for $G = 5000$ iterations (say) to find the mean vector $\boldsymbol{\mu} = G^{-1} \sum_{g=1}^{G} \boldsymbol{\psi}^{(g)}$ and the sample covariance matrix $\mathbf{V} = G^{-1} \sum_{g=1}^{G} (\boldsymbol{\psi}^{(g)} - \boldsymbol{\mu})(\boldsymbol{\psi}^{(g)} - \boldsymbol{\mu})'$. Based on these quantities, the proposal density can be specified as $f_T(\boldsymbol{\psi}|\boldsymbol{\mu}, \tau\mathbf{V}, \nu)$, where f_T is the multivariate-t density with ν degrees of freedom. A sample of draws from the posterior distribution can then be obtained by repeating the following step.

Algorithm MNP 2

- Sample $(\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)$ from $f_T(\boldsymbol{\psi}|\boldsymbol{\mu}, \tau\mathbf{V}, \nu)$ and let

$$\alpha[(\boldsymbol{\beta}, \boldsymbol{\theta}), (\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)] = \min \left\{ 1, \frac{p(\mathbf{y}|\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)}{p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})} \frac{f_T(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{\mu}, \tau\mathbf{V}, \nu)}{f_T(\boldsymbol{\beta}^t, \boldsymbol{\theta}^t|\boldsymbol{\mu}, \tau\mathbf{V}, \nu)} \right\}$$

  denote the probability of move. Then move to $(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ with probability $\alpha[(\boldsymbol{\beta}, \boldsymbol{\theta}), (\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)]$ and stay at $(\boldsymbol{\beta}, \boldsymbol{\theta})$ with probability $1 - \alpha[(\boldsymbol{\beta}, \boldsymbol{\theta}), (\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)]$.

It should be noted that if J is large it may be necessary to sample $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in two blocks. In that case, however, the likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ must be evaluated twice within each iteration and the proposal density for each block must also be defined in a different way. At this point, therefore, it does not seem feasible to implement this algorithm in general without incurring an enormous computational cost.

Starting values for algorithms

It is often useful to initialize posterior sampling algorithms in regions that have high mass under the posterior distribution. This seems to be particularly important in the fitting of MNP models. One way to compute a high density point is by the Monte Carlo EM (MCEM) algorithm, which also relies on data augmentation and delivers the approximate maximum likelihood estimate [Natarajan, Kiefer, and McCulloch (1995)]. Let $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ denote the current value of the parameters and $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$ the estimates obtained at convergence. The algorithm is implemented by iterating on the following steps.

Algorithm MCEM

- Sample $\mathbf{Z}^{(j)}$ as in the basic algorithm for sampling posterior distribution. Repeat this step N times.
- Update $\boldsymbol{\beta}$ through the expression $\boldsymbol{\beta}^{(t+1)} = (\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}^{-1} \mathbf{X}_i)^{-1} (\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}^{-1} \mathbf{Z}_i)$, where $\mathbf{Z}_i = N^{-1} \sum_{j=1}^{N} \mathbf{Z}_i^{(j)}$ is the average of $\mathbf{Z}_i$ over the N draws.
- Update $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}^{(t+1)}$ by maximizing the function $\sum_{j=1}^{N} \log f(\mathbf{Z}^{(j)} | \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\Sigma})$, where $f(.)$ is the complete data density.

In implementing this algorithm N is initially chosen to be a small number, and its value is steadily increased as the maximizer is approached. In the examples below, N is set equal to ten for the first twenty iterations and is increased to four hundred close to convergence.

A well known problem with the EM algorithm is that it does not automatically provide an estimate of the observed information matrix at convergence. This is not a problem if one is using the MCEM algorithm to supply starting values for the full posterior sampling algorithms. If standard errors are required, then one can compute the observed information matrix using the Louis (1982) formula $-E\{\Delta^2 \log f(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\} - \mathrm{Var}\{\Delta \log f(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\}$, where the expectation and variance are with respect to the distribution $\mathbf{Z}|y, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}$ and Δ denotes differentiation w.r.t. the parameters. Each of these terms can be estimated by taking M additional draws $\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}\}$ from $\mathbf{Z}|y, \boldsymbol{\beta}, \hat{\Sigma}$ and computing the expectation and variance as corresponding sample averages.

3.2 MCMC sampling of the MNT and MNL models

Consider now the fitting of the MNT model by MCMC methods. In this case, Algorithm 1 is easily modified because of the fundamental connection between the multivariate-t and multivariate normal distributions. The general idea is to conduct the sampling with λ_i $(i \leq n)$ as additional parameters of the model. Then, conditional on λ_i, the latent data $\mathbf{Z}_i$ follow the distribution

$$\mathbf{Z}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Sigma}).$$

Accordingly, the full conditional distributions of z_{ij} and $\boldsymbol{\beta}$ are obtained by replacing $\boldsymbol{\Sigma}$ by $\lambda_i^{-1}\boldsymbol{\Sigma}$ in the expressions presented above. To sample $\boldsymbol{\sigma}$, the MH approach given in the context of Algorithm 1 can again be applied by noting that $\mathbf{Z}_i\lambda_i^{1/2}$ is distributed as normal with mean $\mathbf{X}_i\boldsymbol{\beta}\lambda_i^{1/2}$ and variance $\boldsymbol{\Sigma}$. Finally, the mixing variable λ_i $(i \leq n)$ is sampled from the gamma distribution

$$\lambda_i|\mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \text{Gamma}\left(\frac{\nu + J}{2}, \frac{\nu + (\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right), \; i \leq n.$$

Algorithm 2 can also be modified by making use of the GHK algorithm to evaluate $\Pr(y_i = j|\boldsymbol{\beta}, \boldsymbol{\Sigma})$, but now under the assumption that the distribution of the latent data is multivariate-t. The GHK algorithm in this case requires simulation from univariate student-t distributions as discussed in Appendix 1.

To conduct MCMC sampling of $\boldsymbol{\beta}$ in the MNL model we note that the posterior density of $\boldsymbol{\beta}$ is proportional to

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i,j}\left(\frac{\exp(\mathbf{v}_{ij}'\boldsymbol{\delta} + \mathbf{w}_i'\boldsymbol{\gamma}_j)}{\sum_j^{J+1}\exp(\mathbf{v}_{ij}'\boldsymbol{\delta} + \mathbf{w}_i'\boldsymbol{\gamma}_j)}\right)^{d_{ij}}\pi(\boldsymbol{\beta}),$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases}$$

and $\pi(\boldsymbol{\beta})$ is the prior distribution for $\boldsymbol{\beta}$, assumed to be multivariate normal with known mean vector and covariance matrix. This density can be sampled by the MH algorithm in which the proposal density $q(\boldsymbol{\beta})$ is taken to be multivariate-t with mean vector equal to the mode of the posterior distribution and scale matrix equal to the curvature at the mode of the posterior distribution. The algorithm is then implemented by iterating on the following steps.

Algorithm MNL

- Let $\boldsymbol{\beta}$ be the current value and choose $\boldsymbol{\beta}^\dagger$ from $q(\boldsymbol{\beta})$.
- Accept $\boldsymbol{\beta}^\dagger$ as the next value in the sample with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^\dagger) = \min\left\{1, \frac{\pi(\boldsymbol{\beta}^\dagger|\mathbf{y})q(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta}|\mathbf{y})q(\boldsymbol{\beta}^\dagger)}\right\}.$$

- Accept $\boldsymbol{\beta}$ as the next value in the sample with probability $1 - \alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^\dagger)$.

3.3 Comparison of algorithms for the MNP model

The algorithms for the MNP model are now compared with data on four multinomial choices. The results are similar for the MNT model and are suppressed. The data consist of 210 observations on highway and transit usage between Sydney, Melbourne, and New South Wales, Australia, that were collected by David Hensher and are contained in the LIMDEP computer package. The choices are whether to travel by air (A), train (T), bus (B), or car (C), with car treated as the base choice. The covariates are terminal waiting time (TTME), in-vehicle time (INVT), in-vehicle cost (INVC), a generalized cost measure (GC), indicator variables for the first three choices (IND1, IND2, IND3), household income times A (HA), and traveling party size times A (PA). Data for the first two observations are presented in Table 1. The covariates are in their undifferenced form ($\mathbf{v}_{ij}$).

One model that is useful for these data consists of the seven covariates TTIME, INVT, IND1, IND2, IND3, HA and PA. On the assumption that the prior information on the parameters is represented by the distributions

$$\boldsymbol{\beta} \sim N(0, 10I_k) \quad \text{and} \quad \boldsymbol{\sigma} \propto N(\mathbf{s}_0, \mathbf{S}_0),$$

where $\mathbf{s}_0 = (0, 1, 0, 0, .75)$ and $\mathbf{S}_0 = \text{diag}(1, 0.51, 1, 1, 0.51)$, we find (via the simulation method described in Section 2) that the prior mean of $\boldsymbol{\theta}$ is $(-0.01, -0.057, 0.006, 0.006, -0.383)$ and that the prior variance is approximately 0.28 for each component of $\boldsymbol{\theta}$. The prior on $\boldsymbol{\theta}$ is taken to be Gaussian with these moments. Algorithms 1–2 are run for 10,000 cycles, and the MH parameters in Algorithms 1 and 2 are set at $\tau = 1$ and $\nu = 20$. Each of the posterior sampling algorithms is initialized by the point estimate from the MCEM algorithm.

y_i	TTIME	INVT	INVC	GC	IND1	IND2	IND3	HA	PA
4	69	59	100	70	1	0	0	35	1
4	34	31	372	71	0	1	0	0	0
4	35	25	417	70	0	0	1	0	0
4	0	10	180	30	0	0	0	0	0
4	64	58	68	68	1	0	0	30	2
4	44	31	354	84	0	1	0	0	0
4	53	25	399	85	0	0	1	0	0
4	0	11	255	50	0	0	0	0	0

Table 1: Data for the first two of 210 subjects.

Results are summarized in Tables 2 and 3. Point estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$—MLE and posterior means—are fairly close across the various algorithms. Some of the differences may be attributable to identification problems inherent in this model that are discussed below. Differences between the MLE and the posterior means may also reflect asymmetries in the posterior distribution with a resulting difference between modes and means. It is also interesting to compare the serial correlation of the sampled output from Algorithms 1-2. Figures 1 and 2 reproduce the output of $\boldsymbol{\sigma}$. It is seen that the serial correlation of the output from Algorithm 2 dissipates quickly relative to Algorithm 1. However, the point estimates from the two algorithms are very close (as are the predicted probabilities computed below) and, therefore, one may conclude that the benefits that accrue from adopting Algorithm 2 are outweighed by the computational burden.

Finally, we compare the posterior predicted probability of the observed choice of each individual from each of the three algorithms. This probability is computed from the posterior sample of the parameters generated by each of the algorithms as

$$\Pr(Y_i = j_i|a) = G^{-1} \sum_g \Pr(Y_i = j_i|\boldsymbol{\beta}_a^{(g)}, \boldsymbol{\Sigma}_a^{(g)}), \tag{5}$$

where j_i is the choice made by the ith subject, $(\boldsymbol{\beta}_a^{(g)}, \boldsymbol{\Sigma}_a^{(g)})$ are draws from the posterior distribution, and $a = 1, 2$ indexes, respectively, the MNP Algorithms 1 and 2. Figure 3

	MCEM		Algorithm 1		Algorithm 2	
Variable	MLE	Std Error	Mean	Std Dev	Mean	Std Dev
TTME	-0.030	0.007	-0.040	0.007	-0.039	0.007
GC	-0.011	0.002	-0.012	0.002	-0.012	0.002
IND1	2.096	0.743	2.807	0.601	2.666	0.601
IND2	1.474	0.317	1.786	0.271	1.714	0.273
IND3	1.272	0.316	1.511	0.269	1.477	0.266
HA	0.013	0.005	0.013	0.006	0.014	0.006
PA	-0.471	0.125	-0.523	0.125	-0.512	0.122

Table 2: Posterior results for β.

	MCEM		Algorithm 1		Algorithm 2	
	MLE	Std Error	Mean	Std Dev	Mean	Std Dev
σ_{21}	0.350	0.150	0.266	0.209	0.254	0.204
σ_{22}	0.493	0.250	0.928	0.347	0.879	0.345
σ_{31}	0.203	0.143	0.076	0.222	0.080	0.197
σ_{32}	0.215	0.133	0.334	0.189	0.295	0.188
σ_{33}	0.226	0.157	0.474	0.188	0.413	0.182

Table 3: Posterior results for Σ

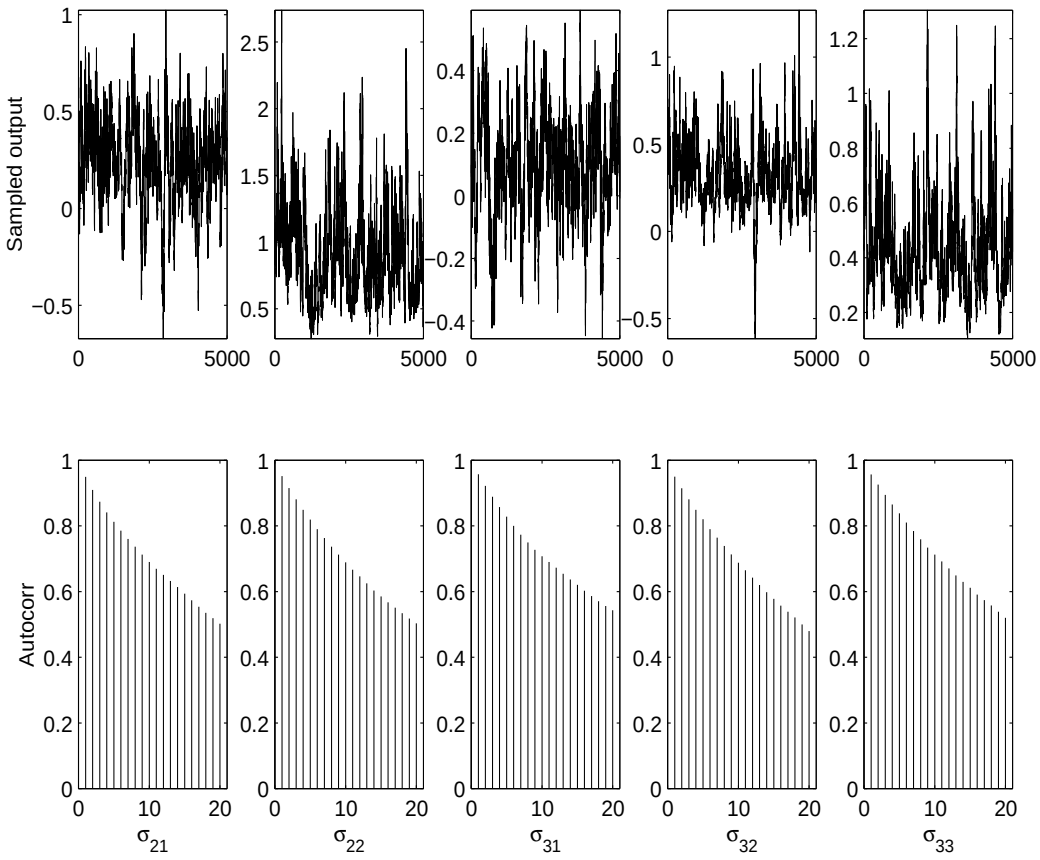

Figure 1: Sampled output and autocorrelation plots of σ from Algorithm 1.

displays the scatter plot of the probabilities for two pairs of the algorithms. It will be seen that the points lie on or very close to the 45° line. The correlations between the predicted probabilities are over 0.999, indicating that the results from the algorithms are indistinguishable.

3.4 Identification issues

In fitting MNP models it is important to keep in mind the issue of parameter identification. Keane (1992) points out that the parameters of the MNP model are weakly identified and attributes this problem to the lack of exclusion restrictions in $\boldsymbol{\beta}$. He argues that "movements in the regressor coefficients can effectively mimic the effects of changes in the covariance parameters," thus leading to a flat likelihood surface. We attribute the problem of fragile identification to the large number of free parameters in the model rather than to the lack

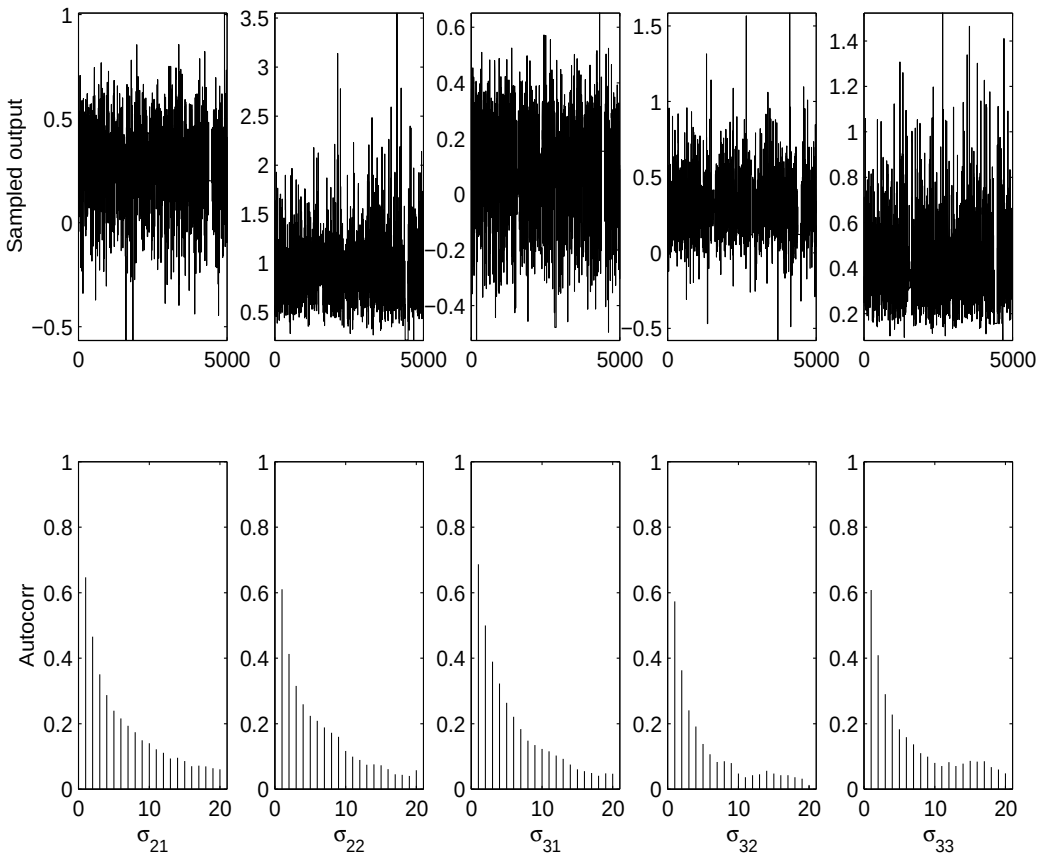

Figure 2: Sampled output and autocorrelation plots of σ from Algorithm 3.

of exclusion restrictions. It is possible to obtain very similar likelihood functions for quite different sets of parameter values whether or not there are exclusion restrictions. The same problem arises in the MNT version of the model, but it is less serious in the MNL model because there are no covariance parameters to estimate.

The case $J = 2$ is examined. Figure 4 displays in the (z_1, z_2) space those regions that lead, respectively, to choices 1 (lightly shaded), 2 (medium shaded), or 3 (heavily shaded). The distribution of a (z_{i1}, z_{i2}) pair depends on its mean $\mathbf{X}_i\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$. To see how fragile identification may arise, consider an observation for which the mean is located deep in the region where $y_i = j$ (i.e., the covariates are very effective in predicting choice). In that case, the probability that the person chooses j is very high. If the covariates are effective predictors for most of the observations in a sample, the observed

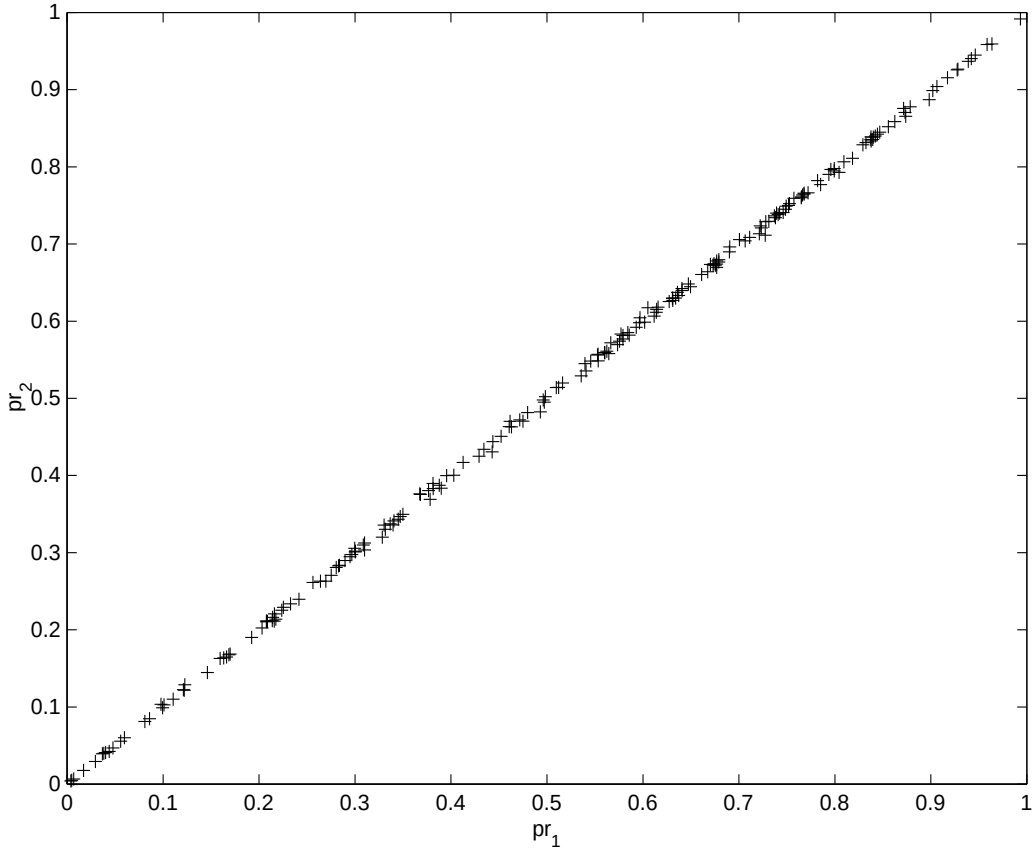

Figure 3: Probability of observed choice: Algorithm 1 vs 2.

choice is consistent with quite different covariance matrices, and the resulting likelihood function is flat. Note that the likelihood contribution of the ith subject is based only on the actual choice made. Figure 4 illustrates the problem in a less extreme case. The dashed 99% contour is plotted around a mean of $(0, -0.5)$ and $\text{vech}^*(\mathbf{\Sigma}) = (0, 1)$, and the solid contour is around mean $(0.39, -0.22)$ and $\text{vech}^*(\mathbf{\Sigma}) = (1.68, 3.00)$. The correlation is zero for the first of these and 0.97 for the second. Although the two sets of parameters are very different, they yield the same probabilities of choices to two decimal places: 0.43, 0.22, and 0.35. Thus, even for observations that are not deep in one of the regions, the parameters may not be well identified, and the extent of the problem would vary for different data sets.

In view of this discussion, we support Keane's ideas that identification may be fragile but believe that for some data sets this fragility will persist even in the presence of exclusion

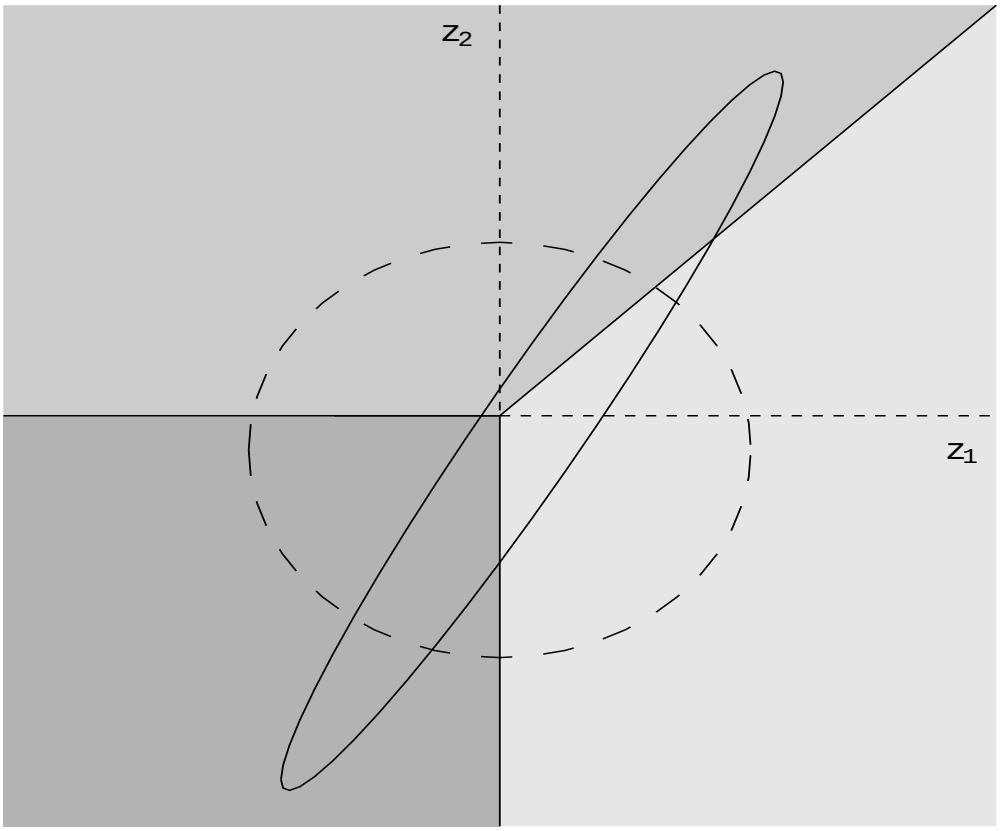

Figure 4: Means at (0,-0.50) and (0.39, -0.22), vech*(Σ)s at(0,1) and (1.68,3.00), and two 99% contours.

restrictions. (Our model implies several restrictions; for example, the variable IND1 is not contained in the T and B equations.) The problem seems less severe for estimates of $\boldsymbol{\beta}$ than for $\boldsymbol{\Sigma}$, and although the coefficients are somewhat different, the predicted probabilities are very close.

4 Comparing alternative models with Bayes factors

An important question that we now address is the comparison of alternative multinomial models. One complication in this context is the fact that the models under consideration are typically non-nested. In such cases, the Bayesian framework is particularly convenient. Let the collection of models be denoted by $\mathcal{M}_k, k = 1, \ldots, K$, and let the marginal likelihood

of model $\mathcal{M}_k$ be given by

$$m(\mathbf{y}|\mathcal{M}_k) = \int p(\mathbf{y}|\mathcal{M}_k, \boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))\pi(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathcal{M}_k)\, d\boldsymbol{\beta}\, d\boldsymbol{\theta}, \tag{6}$$

where we have adopted the $\boldsymbol{\theta}$ parameterization for $\boldsymbol{\Sigma}$ and suppressed the dependence of the parameters on $\mathcal{M}_k$. The MNL marginal likelihood has the same form except for the integration over $\boldsymbol{\theta}$. Given the marginal likelihood of each model, model evidence in favor of $\mathcal{M}_k$ over $\mathcal{M}_r$ is measured by the Bayes factor B_{kr}, which is given by the ratio $m(\mathbf{y}|\mathcal{M}_k)/m(\mathbf{y}|\mathcal{M}_r)$.

4.1 Computation of marginal likelihood

A straightforward way to estimate the integral (6) is by the method of Chib (1995) [see DiCiccio, Kass, Raftery, and Wasserman (1997) for this and other methods of computing the marginal likelihood]. The Chib method utilizes Bayes theorem to obtain

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

where all normalizing constants are included and $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ are arbitrary points, taken to be high density values such as the posterior means. Transforming to the log scale and utilizing conditional/marginal decompositions yields

$$\begin{aligned} \log m(\mathbf{y}|\mathcal{M}_k) &= \log p(\mathbf{y}|\mathcal{M}_k, \boldsymbol{\beta}^*, \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) + \log \pi(\boldsymbol{\beta}^*|\mathcal{M}_k) + \log \pi(\boldsymbol{\theta}^*|\mathcal{M}_k) \\ &\quad - \log \pi(\boldsymbol{\beta}^*|\mathcal{M}_k, y, \boldsymbol{\theta}^*) - \log \pi(\boldsymbol{\theta}^*|\mathcal{M}_k, y). \end{aligned} \tag{7}$$

A key desirable feature of this approach is that the likelihood function $p(\mathbf{y}|\mathcal{M}_k, \boldsymbol{\beta}^*, \boldsymbol{\Sigma}(\boldsymbol{\theta}^*))$ needs to be computed only once. In the appendix we explain how each term in (7) is computed.

In order to implement this Bayesian model selection approach it is necessary to think carefully about the prior inputs. One criterion is that the prior distributions lead a priori to the same distribution of observable responses across models. Another possible requirement on the prior is that the choice between different models depends primarily on the data and only slightly on the details of the prior. We offer two suggestions for choosing such priors for $\boldsymbol{\Sigma}$.

One approach to specifying a prior on $\boldsymbol{\Sigma}_k$, where k indicates $\mathcal{M}_k$, is based on the preposterior distribution of the data under $\mathcal{M}_k$:

$$\Pr(\mathbf{y}|\mathcal{M}_k) = \int f_k(z|\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)\pi_k(\boldsymbol{\beta}_k)\pi_k(\boldsymbol{\Sigma}_k)\, dz\, d\boldsymbol{\beta}_k\, d\boldsymbol{\Sigma}_k,$$

where $\boldsymbol{\beta}_k \sim N(0, cI)$ and $\text{vech}^*(\boldsymbol{\Sigma}_k) \propto N(\mathbf{s}_0, \mathbf{S}_0)$. In this approach, c, $\mathbf{s}_0$, and $\mathbf{S}_0$ are chosen to make $\Pr(\mathbf{y}|\mathcal{M}_k)$ approximately equal for the models to be compared and approximately equal to what is known about $\Pr(\mathbf{y}|\mathcal{M}_k)$. For example, for the travel data discussed in the example, the approximate percentage breakdown of people traveling by the various modes may be known from previous studies, or information may be available for trips between comparable destinations. Under this approach, the priors for the two models are comparable in the sense that they produce the same probabilities of choice.

An alternative prior can be based on a method that uses a training sample. For model $\mathcal{M}_k$, assume that the prior distribution is $\pi_k(\boldsymbol{\beta}, \boldsymbol{\Sigma}_k|\mathbf{c}_k)$, where $\mathbf{c}_k$ is a vector of hyperparameters. Let $\mathbf{y}^t$ be a vector of n_1 observations selected at random from $\mathbf{y}$, and let $\mathbf{y}^r$ be the remainder of the sample. The training prior distribution is defined as

$$\pi_k(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k|\mathbf{y}^t) = \pi_k(\mathbf{y}^t|\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)\pi_k(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k|\mathbf{c}_k).$$

The ratio of marginal likelihoods for $\mathcal{M}_k$ and $\mathcal{M}_j$ based on $\mathbf{y}^t$ is

$$B^*_{kj} = \frac{m_k(\mathbf{y}^t)}{m_j(\mathbf{y}^t)} = \frac{\int p_k(\mathbf{y}^t|\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)\pi_k(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k|\mathbf{c}_k)\, d\boldsymbol{\beta}_k\, d\boldsymbol{\Sigma}_k}{\int p_j(\mathbf{y}^t|\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)\pi_j(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j|\mathbf{c}_j)\, d\boldsymbol{\beta}_j\, d\boldsymbol{\Sigma}_j}.$$

This expression represents the Bayes factor before seeing the data in $\mathbf{y}^r$. Our suggestion is to choose $\mathbf{c}_k$ and $\mathbf{c}_j$ so that $B^*_{kj} = 1$. This choice makes the first stage priors $\pi_k(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k|\mathbf{c}_k)$ and $\pi_j(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j|\mathbf{c}_j)$ comparable in the sense that the Bayes factor based on them and the training sample does not favor either model.

Bayes factors are now computed for the data of our example. For the purpose of this illustration we have chosen proper priors for each model that imply approximately the same prior probability distribution on the outcomes. The consequences of a particular prior for the outcomes are determined by simulation. This requires the simulation of parameters from the prior distribution followed by a simulation of the outcomes given the parameters. These two steps are repeated a large number of times and the hyperparameters are adjusted until the implied empirical distribution of the outcomes is roughly similar across models.

4.2 Example (cont.)

Let the model fitted in Section 3.3 be denoted as $\mathcal{M}_1$, let $\mathcal{M}_2$ denote the MNP model that adds two covariates—in vehicle cost for all stages (INVC) and in vehicle time for all stages (INVT)—to model $\mathcal{M}_1$, and let $\mathcal{M}_3$ denote the MNP model in which $\boldsymbol{\Sigma}$ has equal covariances. This patterned covariance arises from the assumption that the original set of four latent variables are independent. Finally, let $\mathcal{M}_4$ denote the MNL model and let $\mathcal{M}_5$ denote the MNT model with $\nu = 10$ (both with the same covariates as $\mathcal{M}_1$).

We begin with the posterior distribution of $\boldsymbol{\sigma}$ in model $\mathcal{M}_3$. Due to the restriction on the covariances, Algorithm 2 cannot be applied in this case, but one can use a version of Algorithm 1 where the $\boldsymbol{\sigma}$ parameters are sampled directly through an MH step. The posterior distribution is summarized in Table 4. The posterior distribution of $\boldsymbol{\beta}$ in this model is close to that of $\mathcal{M}_1$ and is not reported.

	Algorithm 1	
Covariance	Mean	Std Dev
$\sigma_{ij}, i \neq j$	0.267	0.112
σ_{22}	0.800	0.310
σ_{33}	0.445	0.184

Table 4: Posterior distribution for Σ under $\mathcal{M}_3$.

The first set of model comparison results, computed as a by-product of Algorithm 1 (with a reduced run of 10,000 iterations), is contained in Tables 5. From this table it is clear that the data strongly favor $\mathcal{M}_4$, that the support for $\mathcal{M}_1$ and $\mathcal{M}_3$ is approximately the same, and that $\mathcal{M}_2$ is decisively rejected. This comparison of alternative non-nested models nicely illustrates the usefulness of the Bayes factors approach.

Next, Table 6 allocates the marginal likelihood into its components. Table 6 reveals that $\mathcal{M}_2$ and $\mathcal{M}_4$ have similar likelihood values, even though the former has nine more parameters, and that the larger value of the marginal likelihood of $\mathcal{M}_4$ arises from the value of its posterior ordinate. Moreover, since the prior ordinates except for $\mathcal{M}_2$ are very similar, the dominance of $\mathcal{M}_4$ cannot be ascribed to incomparable priors. Further understanding of the dominance of the MNL model comes from Figure 5. This figure provides a scatter

Model	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$\mathcal{M}_2$	-4.63	–	–	–
$\mathcal{M}_3$	0.54	5.17	–	–
$\mathcal{M}_4$	7.81	12.43	7.27	–
$\mathcal{M}_5$	2.23	6.86	1.69	-5.58

Table 5: Log (base 10) of Bayes factors for row model against column model

Model	Data Likelihood	Prior Ordinate	Posterior Ordinate	Marginal Likelihood	S. E.
$\mathcal{M}_1$	-83.37	-10.92	-9.55	-103.72	0.05
$\mathcal{M}_2$	-79.76	-13.65	-14.94	-108.35	0.06
$\mathcal{M}_3$	-83.11	-10.54	-9.53	-103.18	0.04
$\mathcal{M}_4$	-80.75	-9.99	-5.17	-95.91	0.04
$\mathcal{M}_5$	-83.82	-8.78	-8.89	-101.49	0.05

Table 6: Log (base 10) of the marginal likelihood and its components. Numerical standard error in the last column is computed as in Chib (1995).

plot of the predicted probabilities from models $\mathcal{M}_1$ and $\mathcal{M}_4$ of the choice made by each subject. In the case of the MNL model, the probability of the observed outcome is larger than that from the MNP model in about 80% of the observations. Thus, in this case, the MNL model is more successful than the MNP model in predicting the choices made by the individuals. The MNT model was included because it represents a compromise between MNP and MNL in the sense that it allows for correlated errors but has thicker tails than the normal. Interestingly, Greene (1997) obtains a parallel result with a classical nested models test. He compares the MNL model and the nested logit model (a model that is similar to the MNP in that both relax the independence of irrelevant attributes property) and finds that the MNL model cannot be rejected for these data.

5 Conclusions

This paper has presented a set of new MCMC-based algorithms and inference procedures for the Bayesian analysis of the MNP model. One contribution is the comparison for the first time of different MCMC algorithms for simulating the posterior distribution of the

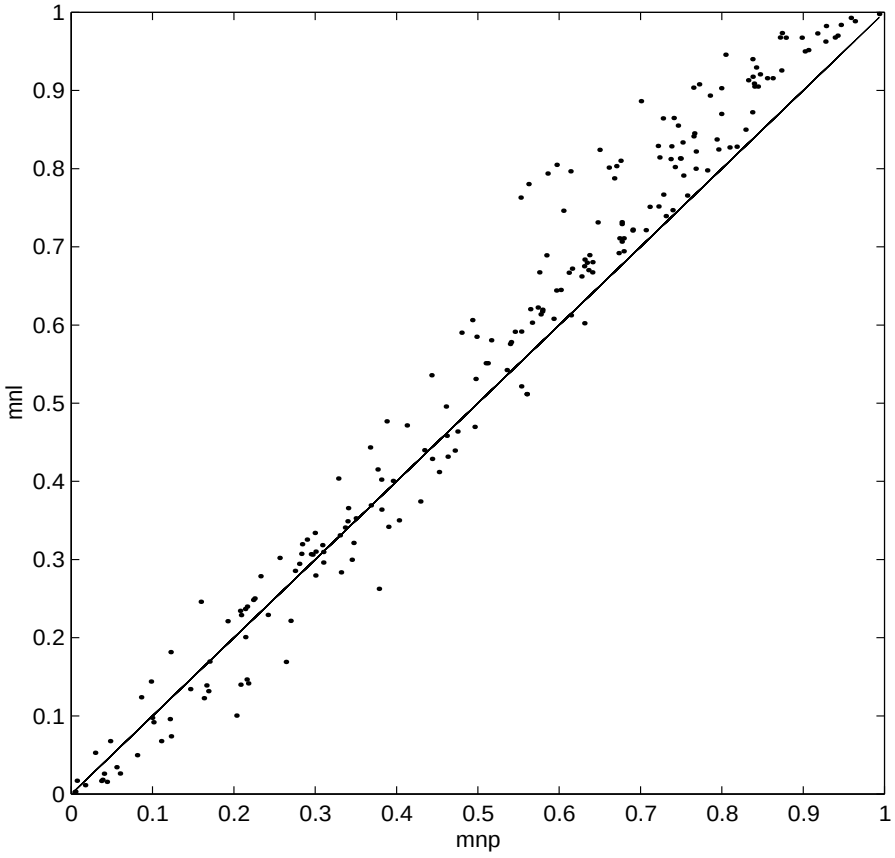

Figure 5: Predicted posterior mean of observed choices from the MNL model (vertical axis) and the MNP model (horizontal axis)

parameters. Another contribution is the study of the MNT model and its analysis by MCMC methods. A general comment based on our experience is that the fitting of these models requires some care and that the covariance parameters can be particularly difficult to estimate, regardless of the algorithm that may be used in the fitting.

An important concern of this paper is the question of comparing the fit of alternative MNP models and the fit of the MNP model with that of the MNT and the simpler MNL model. We show that the Bayes factors framework is quite useful for this purpose and that the marginal likelihood of competing models can be computed from the MCMC output as a by-product of the simulation procedure. One interesting result is that the MNP model is not guaranteed to fit better than the MNL model once model complexity is taken into account. Finally, the paper reports on a probability plot for comparing the fit of alternative

multinomial response models that should be useful in the practical fitting of these models.

A Appendix

A.1 Computing $p(y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with the GHK algorithm

We compute $p(y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ by the Geweke-Keane-Hajivassiliou (GHK) method [see Geweke (1991), Hajivassiliou (1990), and Keane (1994)] for the MNP model as follows. Let $\mu_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = LL'$, where $L = \{l_{km}\}$ is lower triangular, and repeat the following steps for $r = 1, 2, \ldots, R$.

- If $y_i = j$ $(j < J+1)$, reorder the variables so that outcome j appears in the first row. Let
$$Q_{i1}^{(r)} = 1 - \Phi(A_{i1}^{(r)}),$$
where $A_{i1}^{(r)} = -\mu_{i1}/l_{11}$ and $\Phi(\cdot)$ is the cdf of the standard normal distribution. Draw $\epsilon_{i1}^{(r)}$ from $\mathrm{TN}(A_{i1}^{(r)}, \infty)$, where $\mathrm{TN}(B, U)$ is the standard normal distribution truncated to (B, U).
  - For $j = 2, \ldots, J$, let
$$Q_{ij}^{(r)} = \Phi(B_{ij}^{(r)}),$$
where
$$B_{ij}^{(r)} = \frac{\mu_{i1} - \mu_{ij} + l_{11}\epsilon_{i1}^{(r)} - \sum_{m=1}^{j-1} l_{jm}\epsilon_{im}^{(r)}}{l_{jj}}.$$
  - Draw $\epsilon_{ij}^{(r)}$ from $\mathrm{TN}(-\infty, B_{ij}^{(r)})$.
- If $y_i = J+1$, repeat the following steps.
  - For $j = 1, 2, \ldots, J$, let
$$Q_{ij}^{(r)} = \Phi(B_{ij}^{(r)}),$$
where
$$B_{ij}^{(r)} = \frac{-(\mu_{ij} + \sum_{m=1}^{j-1} l_{jm}\epsilon_{im}^{(r)})}{l_{jj}}.$$
  - Draw $\epsilon_{ij}^{(r)}$ from $\mathrm{TN}(-\infty, B_{ij}^{(r)})$.

- Compute

$$Q_i^{(r)} = \prod_{j=1}^{J} Q_{ij}^{(r)}.$$

The GHK estimate of the probability $p(y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is then given by

$$Q_i = R^{-1} \sum_{r=1}^{R} Q_i^{(r)}.$$

To apply this method to the MNT model, one draws the $\epsilon_{ij}^{(r)}$ from the standard univariate-t distribution, truncated as above, and replaces the cdf of the normal in the above calculations by the cdf of the t distribution.

A.2 Computing the marginal likelihood using Chib's method

Details for computing the marginal likelihood for the MNP model of equation (7) follow, and the necessary modifications for the MNT model are obvious. Calculations for the MNL model are discussed at the end of this subsection. Note that in this section $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ refer to values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ at high density points. The dependence of the parameters on $\mathcal{M}_k$ is suppressed.

1. The likelihood contribution of the ith observation $f(y_i|\mathcal{M}_k, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ is calculated by the GHK algorithm as described above. The $\log p(y_i|\mathcal{M}_k, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ are then added to obtain $\log p(\mathbf{y}|\mathcal{M}_k, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$.

2. The next term, $\pi(\boldsymbol{\beta}^*|\mathcal{M}_k)$, is the ordinate of a normal distribution.

3. The term $\pi(\boldsymbol{\theta}^*|\mathcal{M}_k)$ is the ordinate of a normal distribution.

4. The fourth term,

$$\pi(\boldsymbol{\beta}^*|\mathcal{M}_k, \mathbf{y}, \boldsymbol{\theta}^*) = \int \pi(\boldsymbol{\beta}^*|\mathcal{M}_k, \mathbf{y}, \mathbf{Z}, \boldsymbol{\theta}^*) f(\mathbf{Z}|\mathcal{M}_k, \mathbf{y}, \boldsymbol{\theta}^*)\, d\mathbf{Z},$$

where $\pi(\boldsymbol{\beta}|\mathcal{M}_k, \mathbf{y}, \boldsymbol{\theta}^*, \mathbf{Z})$, is a normal distribution with parameters $(\boldsymbol{\beta}_1, \mathbf{B}_1)$ evaluated at $\boldsymbol{\theta}^*$. This integral can be accurately estimated by drawing a large sample of $\mathbf{Z}$ values from a reduced MCMC run consisting of draws from

$$[\mathbf{Z}_1|\mathcal{M}_k, y_1, \boldsymbol{\beta}, \boldsymbol{\theta}^*], \ldots, [\mathbf{Z}_n|\mathcal{M}_k, y_n, \boldsymbol{\beta}, \boldsymbol{\theta}^*], \quad \text{and} \quad [\boldsymbol{\beta}|\mathcal{M}_k, \mathbf{Z}_1, \ldots, \mathbf{Z}_n, \boldsymbol{\theta}^*].$$

From the G values of $\mathbf{Z}$ drawn from this run, an estimate of the desired ordinate is given by

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathcal{M}_k, \mathbf{y}, \boldsymbol{\theta}^*) = G^{-1} \sum_{g=1}^{G} \phi_p(\boldsymbol{\beta}^*|\boldsymbol{\beta}_1^{(g)}, B_1^{(g)-1}),$$

where $\boldsymbol{\beta}_1^{(g)} = \mathbf{B}_1^{(g)-1}(\sum_i^n \mathbf{X}_i'\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)^{-1}\mathbf{Z}_i^{(g)} + \mathbf{B}_0\boldsymbol{\beta}_0)$ and $\mathbf{B}_1^{(g)} = \mathbf{B}_0 + \sum_i^n \mathbf{X}_i'\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)^{-1}\mathbf{X}_i$.

5. Kernel smoothing may be applied to the sample of $\boldsymbol{\theta}$ generated by the original MCMC run to obtain the ordinate at $\boldsymbol{\theta}^*$. If $\boldsymbol{\theta}$ is high-dimensional it may be desirable to find the ordinate by applying the kernel smoothing to several blocks of the θ_{ij} [Chib and Greenberg (1995b)]. Note that the kernel smoothing steps suggested here and above can be made as accurate as desired by increasing the number of simulated values. This option is, of course, not available when kernel smoothing is employed on data for which the sample size is fixed.

Finally, the calculation of the marginal likelihood for the MNL model proceeds in a similar fashion. The likelihood function at $\boldsymbol{\beta}^*$ is available in closed form. The prior ordinate for $\boldsymbol{\beta}$ is a normal distribution, and the posterior ordinate is computed by kernel smoothing.

References

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

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

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

Chib, S. and E. Greenberg (1995b), Analysis of multivariate probit models. *Biometrika,* forthcoming.

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

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

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

GEWEKE, J. (1991), Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* 571–578.

GEWEKE, J., M. KEANE, and D. RUNKLE (1994), Alternative computational approaches to inference in the multinomial probit model, *Review of Economics and Statistics*, 76, 609–632.

GREENE, W. (1997), *Econometric Analysis,* 3rd ed., Upper Saddle River, NJ: Prentice-Hall.

HAJIVASSILIOU, V. A. (1990), Smooth simulation estimation of panel LDV models. Manuscript.

HAUSMAN, J.A. and D.A. WISE (1978), A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogenous preferences. *Econometrica*, 46, 403-426.

KEANE, M. P. (1992), A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics,* 10, 193–200.

KEANE, M. P. (1994), A computationally practical simulation estimator for panel data. *Econometrica,* 62, 95–116.

Louis, T. A. (1982), "Finding the observed information matrix using the EM algorithm," *Journal of the Royal Statistical Society* B, 44, 226–233.

MCCULLOCH, R. E. and P. E. ROSSI (1994), Exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64, 207–240.

MCFADDEN, D (1989), A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 995-1026.

NATARAJAN, R., C. E. MCCULLOCH, and N. M. KIEFER (1995), Maximum likelihood for the multinomial probit model. *Manuscript.*

NOBILE, A. (1995), A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Manuscript.*

PINHEIRO, J. C., and D. M. BATES (1996), Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing,* 6, 289–296.

STERN, S. (1997), Simulation-based estimation. *Journal of Economic Literature,* 35, 2006–2039.