Parametric Discrete Choice Models Based on the Scale Mixtures of Multivariate Normal Distributions*

(running title: Parametric Discrete Choice Models)

By ZHEN CHEN

and

LYNN KUO

University of Connecticut, Storrs

SUMMARY. A rich class of parametric models is proposed for discrete choice data based on the scale mixtures of multivariate normal distributions. With special connections to multinomial probit, the new models can be implemented in a Bayesian framework without much difficulty. The proposed class of models can be extended to panel data where accounting for heterogeneities is needed. This is done by building a Bayesian hierarchical model within the random utility maximization framework. Model comparison using the Bayes factor is developed. The new models and the estimation techniques are illustrated using travel mode choice and panel brand choice data.

1. Introduction

Suppose we observe data on N subjects who independently choose from J mutually exclusive alternatives. It is assumed that each subject can make one and only one choice from the set of alternatives. A central issue in the analysis of such data is to model the probabilities of choosing the alternatives as a function of the attributes associated with each alternative and possibly the characteristics of the subject. For example, in marketing, researchers are interested in understanding how prices of different brands within a certain product category might affect a consumer's choices (Jain *et al.*, 1994); and in transportation, researchers would like to learn the impact of household income on the selection of travel modes (Greene, 1997; McFadden, 1974); in labor economics, re-

* *AMS (1991) subject classification.* 62F15, 62P20, 62P25.

Key words and phrases. Multinomial probit model; MCMC; Random coefficients models; Bayesian analysis

searchers would like to know the impact of education attainment on occupation choices (Schmidt and Strauss, 1975).

A common framework in developing discrete choice models is the random utility maximization theory. A stochastic utility function is assumed for a subject over the alternatives set. The subject chooses the alternative with the highest utility. This utility function is usually written as the sum of an error term and a systematic part relating the utility to the known covariates. The systematic part usually takes a linear-in-parameters form. For the error term, one possibility is to use the extreme value distribution (MacFadden, 1974). This leads to the multinomial logit (MNL) model. Another possibility is to assume a multivariate normal distribution for the error term which leads to the multinomial probit (MNP) model (Hausman and Wise, 1978; McCulloch and Rossi, 1994). We argue that more flexible distributions for the error term are necessary to robustify the discrete choice modeling process. A semi- or non-parametric approach is certainly an option (see Manski, 1984 and Matzkin, 1991, among many others). In this paper, however, we take a strictly parametric approach. But the use of other parametric distributions has been hampered by the computational complexity. From a frequentist perspective, this complexity occurs because the evaluation of the choice probabilities involves a $(J-1)$-dimensional integral of a multivariate distribution and the expression for the choice probabilities usually doesn't have a closed form. From a Bayesian point of view, the complexity arises from the analysis of the posterior distribution: a non-normal specification for the error term usually leads to intractable and complicated posterior distributions.

In this paper, we extend the error distribution to a large family of distributions and demonstrate that the estimation can be implemented without much difficulty. The class of distributions we consider is scale mixtures of multivariate normal (SMMVN) distributions (Andrews and Mallows, 1974; West, 1987) which consists of a wide collection of symmetric and continuous distributions. With SMMVN, the robustification is achieved from the tail behavior of the different distributions. Following Albert and Chib (1993), we estimate the proposed models by the approach of latent variables with Markov chain Monte Carlo (MCMC). The simulation of the posterior is made easy by exploiting the special relationship between SMMVN and the multivariate normal distribution. In fact, multivariate normal is a special case of SMMVN. Some other SMMVN distributions include multivariate t, multivariate logistic, multivariate stable, multivariate double exponential and multivariate exponential power. We term the new discrete choice models corresponding to these distributions multinomial t (MNT), multinomial logistic (MNLOG), multinomial stable (MNS), multinomial double exponential (MNDE), and multinomial exponential power (MNEP), respectively. Given this large class of parametric choice models, we present procedures for model comparison based on Bayes factors (Kass and Raftery, 1995). The contribution of this paper is that we provide a richer class of parametric models for analyzing discrete choice data and thus model robustness can be assessed. Chen and Dey (1998) have used SMMVN in longitudinal binary data analysis. Our work extends theirs to nominal polychotomous data, with and without a longitudinal feature.

We organize the rest of the paper as follows. In Section 2 we introduce the new models based on SMMVN and in Section 3 we consider the sampling of the posterior distribution. Model comparison via marginal likelihood is discussed in Section 4. Section 5 presents a real data application in travel mode choice. In Section 6 we extend the proposed models to a panel data situation and illustrate them in Section 7 by brand choice data in marketing. Section 8 contains brief concluding remarks.

2. Parametric Models with SMMVN

Let y_i denote the discrete choice response on the ith subject, i from 1 to N, where $y_i = j$ indicates that the jth alternative has been chosen by the ith subject, j from 1 to $J+1$. Here we assume the alternatives set is of size $J+1$ instead of J because this will simplify the notation later on. Let $\mathbf{x}_i^* = (\mathbf{x}_{i1}^{*\prime}, \dots, \mathbf{x}_{i,J+1}^{*\prime})'$ be the $(J+1) \times q$ covariate matrix with $\mathbf{x}_{ij}^{*\prime} = (x_{ij1}^*, \dots, x_{ijq}^*)'$ where q is the number of covariates. Note that we write the covariates to be alternative-specific; in fact, they could contain subject-specific variables as well, because we can always convert a subject-specific variable into a set of alternative-specific variables, as suggested by Agresti (1989, Chapter 9).

Let $\mathbf{w}_i^* = (w_{i1}^*, \dots, w_{i,J+1}^*)'$ be the column vector of unobserved random utilities of the $J+1$ alternatives for subject i. It is assumed that the subject chooses the alternative with the highest utility, i.e.

$$y_i = j \text{ if } w_{ij}^* \geq \max(w_{i1}^*, \dots, w_{i,J+1}^*). \tag{1}$$

Our rich class of discrete choice models is constructed by postulating that $\mathbf{w}_i^*$ follows a distribution of the class of scale mixture of multivariate normal distributions:

$$\begin{aligned} &\mathbf{w}_i^* = \mathbf{x}_i^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*, \\ &\boldsymbol{\varepsilon}_i^* | \lambda_i \sim N_{J+1}(\mathbf{0}, \kappa(\lambda_i)\, \Sigma^*), \\ &\lambda_i \sim \pi(\lambda_i) \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}$ is the regression parameter vector of length q, $\boldsymbol{\varepsilon}_i^* = (\varepsilon_{i1}^*, \ldots, \varepsilon_{i,J+1}^*)'$ is the random vector, $\kappa(.)$ is a positive function, and λ_i is the scale mixture parameter with distribution $\pi(\lambda_i)$. Observe

$$f(\mathbf{w}_i^*|\mathbf{x}_i^*\boldsymbol{\beta}, \Sigma^*) = \int_0^\infty \phi_{J+1}(\mathbf{w}_i^*|\mathbf{x}_i^*\boldsymbol{\beta}, \kappa(\lambda_i)\,\Sigma^*)\pi(\lambda_i)d\lambda_i,$$

where $\phi_{J+1}(\mathbf{w}_i^*|\mathbf{x}_i^*\boldsymbol{\beta}, \kappa(\lambda_i)\,\Sigma^*)$ is the density of a $(J+1)$-variate normal distribution with mean $\mathbf{x}_i^*\boldsymbol{\beta}$ and covariance matrix $\kappa(\lambda_i)\,\Sigma^*$.

Comparison between the new class of models and the well-known MNP model is in order. Clearly, if we take $\kappa(\lambda_i) = 1$, and $\pi(\lambda_i = 1) = 1$, then we have the multivariate normal distribution for $\mathbf{w}_i^*$, and hence obtain MNP for the corresponding discrete choice model. By choosing different κ functions and different distributions for λ_i, we can have distributions of $\mathbf{w}_i^*$ other than multivariate normal, and hence models other than MNP. For example, the multivariate t distribution is a member of SMMVN, and thus an MNT model based on the multivariate t can be formed. In Table 1, we list some members of the class of scale mixtures of multivariate normal distributions and the corresponding discrete choice models. Column 1 gives the names of some distributions in the class of SMMVN. The κ functions and the mixing distributions are in columns 2 and 3, respectively; the last column gives the names of the discrete choice models corresponding to each distribution in column 1. Note that in some models (MNT, MNS, and MNEP), the mixture distribution has parameters to be specified. One way to do this is to put some priors on these parameters and let the data determine their most probable values. As was mentioned by Chen and Dey (1998), this will bring extra complexities to the computations. For simplicity, we only consider the cases with fixed parameters. More detailed descriptions can be found in Choy (1995) and Chen and Dey (1998). We refer readers to the Appendix for the notation used in Table 1 and thereafter.

(insert Table 1 around here)

As in the case of MNP, an inherent unidentification problem exists for the new class of models based on SMMVN. Scaling the $\mathbf{w}_i$ in (2) by a positive constant c doesn't change the value of y_i, but does result in a different parameterization $(c\boldsymbol{\beta}, c\kappa(\lambda_i)\,\Sigma^*)$. Then we have $\Pr(y_i|\boldsymbol{\beta}, \kappa(\lambda_i)\,\Sigma^*) = \Pr(y_i|c\boldsymbol{\beta}, c^2\kappa(\lambda_i)\,\Sigma^*)$. The usual approaches to fixing the unidentification problem are to impose restrictions on the covariance matrix. We explain these approaches in Section 3 in detail.

Following the common practice in the literature, we work with the difference form of model (2). To do that, we choose a reference alternative, usually the $(J+1)$th one and express each w_{ij}^*

relative to $w^*_{i,J+1}$ by letting $w_{ij} = w^*_{ij} - w^*_{i,J+1}$, and collect all the w_{ij} in $\mathbf{w}_i$. Now (2) becomes

$$\begin{aligned} &\mathbf{w}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \\ &\boldsymbol{\varepsilon}_i|\lambda_i \sim N_J(\mathbf{0}, \kappa(\lambda_i)\,\Sigma), \\ &\lambda_i \sim \pi(\lambda_i) \end{aligned} \tag{3}$$

where $\mathbf{x}_i$ is the $J \times q$ matrix obtained from the original $\mathbf{x}^*_i$ by subtracting the $(J+1)$th row from the first J rows; and (1) becomes

$$y_i = \begin{cases} j & \text{if } w_{ij} > \max(\mathbf{w}_{i(-j)}, 0), \\ J+1 & \text{if } \mathbf{w}_i < 0, \end{cases} \tag{4}$$

where $\mathbf{w}_{i(-j)}$ is the $\mathbf{w}_i$ vector with the jth element deleted, $\max(\mathbf{w}_{i(-j)}, 0)$ serves as a short hand for $\max(w_{i1}, \ldots, w_{i,j-1}, w_{i,j+1}, \ldots, w_{iJ}, 0)$, and $\mathbf{w}_i < 0$ implies each element of the vector $\mathbf{w}_i$ is less than 0.

The probability that subject i chooses alternative j may be expressed as

$$\begin{aligned} P_i(j|\boldsymbol{\beta}, \Sigma) &= \int_{\mathcal{A}_i} f(\mathbf{w}_i|\mathbf{x}_i\boldsymbol{\beta}, \Sigma)d\mathbf{w}_i \\ &= \int_{\mathcal{A}_i}\int_0^\infty \phi_J(\mathbf{w}_i|\mathbf{x}_i\boldsymbol{\beta}, \kappa(\lambda_i)\,\Sigma)\pi(\lambda_i)d\lambda_i d\mathbf{w}_i. \end{aligned} \tag{5}$$

where $\mathcal{A}_i$ is the set-valued inverse of the mapping in (4), i.e.

$$\mathcal{A}_i = \begin{cases} \{\mathbf{w}_i : w_{ij} > \max(\mathbf{w}_{i(-j)}, 0)\} & \text{if } y_i = 1, \ldots, J \\ \{\mathbf{w}_i : \mathbf{w}_i < 0\} & \text{if } y_i = J+1. \end{cases} \tag{6}$$

Clearly, $\mathcal{A}_i$ depends on the data y_i.

3. Posterior Analysis

In deriving the posterior distribution of the unknown parameters given the data $\mathbf{y} = (y_1, \ldots, y_N)$, we can consider $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ and $\mathbf{w} = (\mathbf{w}'_1, \ldots, \mathbf{w}'_N)'$ as latent variables. Then the joint posterior distribution of the parameters and the latent variables $\mathbf{w}$ and $\boldsymbol{\lambda}$ is given by

$$\pi(\boldsymbol{\beta}, \Sigma, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\lambda}) f(\mathbf{w}|\boldsymbol{\beta}, \Sigma, \boldsymbol{\lambda})\pi(\boldsymbol{\beta}, \Sigma)\pi(\boldsymbol{\lambda}),$$

where $\pi(\boldsymbol{\beta}, \Sigma)$ is the prior on $(\boldsymbol{\beta}, \Sigma)$. Note that

$$f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\lambda}) = \prod_{i=1}^N f(y_i|\mathbf{w}_i, \boldsymbol{\beta}, \Sigma, \lambda_i) = \prod_{i=1}^N I(\mathbf{w}_i \in \mathcal{A}_i),$$

where $I(A)$ is the indicator function of the event A. Hence,

$$\pi(\boldsymbol{\beta}, \Sigma, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}) \propto \pi(\boldsymbol{\beta}, \Sigma) \prod_{i=1}^{N} f(\mathbf{w}_i|\boldsymbol{\beta}, \Sigma,\lambda_i, y_i) I(\mathbf{w}_i \in \mathcal{A}_i)\pi(\lambda_i), \tag{7}$$

where

$$f(\mathbf{w}_i|\boldsymbol{\beta}, \Sigma, \lambda_i) = \phi_J(\mathbf{x}_i\boldsymbol{\beta}, \kappa(\lambda_i)\Sigma),$$

independently across subjects.

An MCMC sampling scheme to evaluating (7) can be constructed from the full conditional distributions $\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{w}, \Sigma, \boldsymbol{\lambda})$, $\pi(\Sigma|\mathbf{y}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda})$, $\pi(\mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\lambda})$, and $\pi(\boldsymbol{\lambda}|\mathbf{y}, \boldsymbol{\beta}, \Sigma, \mathbf{w})$. We now provide details for the full conditional distributions and the sampling schemes. We use the notation $\theta|\cdot$ to denote the conditional distribution of θ given all other parameters as well as the data.

Sampling $\mathbf{w}|\cdot$

We begin with the sampling of the latent variable $\mathbf{w}$ from $\pi(\mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\lambda})$. We can achieve this by drawing $\mathbf{w}_i$ parallelly for $i = 1, \ldots, N$ since they are conditionally independent. From (7) it follows that

$$\pi(\mathbf{w}_i|y_i, \boldsymbol{\beta}, \Sigma, \lambda_i) \propto N_J(\mathbf{w}_i|\mathbf{x}_i\boldsymbol{\beta},\kappa(\lambda_i)\Sigma) I(\mathbf{w}_i \in \mathcal{A}_i).$$

This is a multivariate normal density truncated to the region specified by $\mathcal{A}_i$. We can sample from this distribution efficiently by composing a sub-cycle of J Gibbs steps through the components of $\mathbf{w}_i$ (Geweke, 1991). In the jth step of this sub-cycle, w_{ij} is simulated from $\pi(w_{ij}|y_i, \boldsymbol{\beta}, \Sigma, \lambda_i, \mathbf{w}_{i(-j)})$, which is a truncated univariate normal distribution with density proportional to

$$N(m_{ij}, \tau_{ij}^2)\left\{I(y_i = j)I(w_{ij} > \max(\mathbf{w}_{i(-j)}, 0)) + I(y_i \neq j)I(w_{ij} < \max(\mathbf{w}_{i(-j)}, 0))\right\} \tag{8}$$

where, m_{ij} and τ_{ij}^2 are the conditional mean and variance of w_{ij}, respectively. To derive m_{ij} and τ_{ij}^2, let g_{jj} be the (j, j)th element of Σ^{-1}; let $\mathbf{g}_{j(-j)}$ be the jth row of Σ^{-1} with the jth element deleted; let $\mathbf{x}_{i(-j)}$ be $\mathbf{x}_i$ with the jth row deleted. Then by properties of the multivariate normal distribution and the inverse of the partitioned matrix formula, we have

$$\begin{aligned} \tau_{ij}^2 &= \frac{\kappa(\lambda_i)}{g_{jj}}, \\ m_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} - \frac{\mathbf{g}_{j(-j)}'}{g_{jj}}(\mathbf{w}_{i(-j)} - \mathbf{x}_{i(-j)}\boldsymbol{\beta}). \end{aligned}$$

Sampling $\boldsymbol{\beta}|\cdot$

To describe the sampling distribution of $\boldsymbol{\beta}$, we assume prior independence between $\boldsymbol{\beta}$ and Σ and let $\pi(\boldsymbol{\beta}) = N_q(\boldsymbol{\beta}_0, \mathbf{B}_0^{-1})$. From a standard linear model theorem (Lindley and Smith, 1972), it is easy to see that $\boldsymbol{\beta}|\mathbf{y}, \mathbf{w}, \Sigma, \boldsymbol{\lambda} \sim N_q(\hat{\boldsymbol{\beta}}, \hat{\mathbf{B}}^{-1})$ where

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \hat{\mathbf{B}}^{-1}\left(\mathbf{B}_0\boldsymbol{\beta}_0 + \sum\nolimits_{i=1}^{N} \frac{1}{\kappa\left(\lambda_i\right)}\mathbf{x}_i'\Sigma^{-1}\mathbf{w}_i\right) \text{ and} \\
\hat{\mathbf{B}} &= \mathbf{B}_0 + \sum\nolimits_{i=1}^{N} \frac{1}{\kappa\left(\lambda_i\right)}\mathbf{x}_i'\Sigma^{-1}\mathbf{x}_i.
\end{aligned}$$

To minimize the influence of the prior on the posterior distribution, we specify a very diffuse prior on $\boldsymbol{\beta}$ by taking the precision matrix $\mathbf{B}_0$ to be very "small" (for example, $\varepsilon\mathbf{I}$ with ε very small, where $\mathbf{I}$ denotes the identity matrix with dimension q). With a very small precision matrix, the choice of the prior mean of $\boldsymbol{\beta}$ is immaterial; we can simply set $\boldsymbol{\beta}_0 = \mathbf{0}$.

Sampling $\Sigma|\cdot$

Next we consider the sampling of Σ from $\pi(\Sigma|\mathbf{y}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda})$. From (7),

$$\begin{aligned}
\pi(\Sigma|\mathbf{y}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda}) &\propto \pi(\Sigma)\prod_{i=1}^{N} N_J(\mathbf{w}_i|\mathbf{x}_i\boldsymbol{\beta}, \kappa\left(\lambda_i\right)\Sigma) \\
&\propto \pi(\Sigma)\prod_{i=1}^{N} N_J(\boldsymbol{\varepsilon}_i|\mathbf{0}, \kappa\left(\lambda_i\right)\Sigma). \qquad (9)
\end{aligned}$$

As mentioned in Section 2, we need to take care of the unidentification problem by imposing restriction on Σ. Following McCulloch and Rossi (1994) and McCulloch, Polson and Rossi (2000), we achieve the identification by fixing the first diagonal element of the covariance matrix at value 1 ($\sigma_{11} = 1$). For the purpose of computing Bayes factors to compare models, we take the full identified parameters approach of McCulloch *et al.* (2000). Denote the first column of Σ without the first element by $\boldsymbol{\gamma}$, i.e., $\boldsymbol{\gamma} = (\sigma_{21}, \ldots, \sigma_{J1})'$. Also let $\boldsymbol{\varepsilon}_{i(-1)}$ be the vector of the error terms without the first element ε_{i1}. Observe that

$$\begin{aligned}
&\boldsymbol{\varepsilon}_{i(-1)}|\varepsilon_{i1} \sim N_{J-1}\left(\frac{\varepsilon_{i1}}{\sigma_{11}}\boldsymbol{\gamma}, \kappa\left(\lambda_i\right)\left[\mathrm{Var}(\boldsymbol{\varepsilon}_{i(-1)}) - \frac{1}{\sigma_{11}}\boldsymbol{\gamma}\boldsymbol{\gamma}'\right]\right), \\
&\varepsilon_{i1} \sim N(0, \sigma_{11}),
\end{aligned}$$

where 'Var' is the variance operator. Let $\Phi = \mathrm{Var}(\boldsymbol{\varepsilon}_{i(-1)}) - \frac{1}{\sigma_{11}}\boldsymbol{\gamma}\boldsymbol{\gamma}'$. Since σ_{11} is fixed, (9) can be rewritten as

$$\pi(\Sigma|\mathbf{y}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda}) \propto \pi(\boldsymbol{\gamma}, \Phi)\prod_{i=1}^{N} N_{J-1}(\boldsymbol{\varepsilon}_{i(-1)}|\frac{\varepsilon_{i1}}{\sigma_{11}}\boldsymbol{\gamma}, \kappa\left(\lambda_i\right)\Phi),$$

where we see that both Φ and $\boldsymbol{\gamma}$ can be viewed as parameters of the multivariate regression of $\boldsymbol{\varepsilon}_{i(-1)}$ on ε_{i1}. Note

$$\Sigma = \begin{pmatrix} \sigma_{11} & \boldsymbol{\gamma}' \\ \boldsymbol{\gamma} & \Phi + \frac{1}{\sigma_{11}}\boldsymbol{\gamma}\boldsymbol{\gamma}' \end{pmatrix}. \tag{10}$$

Thus the correspondence between Σ and $(\sigma_{11}, \boldsymbol{\gamma}, \Phi)$ is one-to-one. Define the prior of $\boldsymbol{\gamma}$ as $N_{J-1}(\boldsymbol{\gamma}_0, \mathbf{R}_0^{-1})$, and the prior for Φ^{-1} as $W_{J-1}(\upsilon_0, \Phi_0^{-1})$. Then the conditional distribution of Φ^{-1} given the data and all other unknown parameters is given by $W_{J-1}\left(\hat{\upsilon}, \hat{\Phi}^{-1}\right)$ where $\hat{\upsilon} = N + \upsilon_0$ and

$$\hat{\Phi} = \Phi_0 + \sum_{i=1}^{N} \frac{1}{\kappa(\lambda_i)} \left(\boldsymbol{\varepsilon}_{i(-1)} - \frac{\varepsilon_{i1}}{\sigma_{11}}\boldsymbol{\gamma}\right)\left(\boldsymbol{\varepsilon}_{i(-1)} - \frac{\varepsilon_{i1}}{\sigma_{11}}\boldsymbol{\gamma}\right)';$$

and the conditional distribution of $\boldsymbol{\gamma}$ given the data and all other parameters is $N_{J-1}(\hat{\boldsymbol{\gamma}}, \hat{\mathbf{R}}^{-1})$ with

$$\begin{aligned} \hat{\mathbf{R}} &= \mathbf{R}_0 + \frac{1}{\sigma_{11}^2} \sum_{i=1}^{N} \frac{\varepsilon_{i1}^2}{\kappa(\lambda_i)} \Phi^{-1}, \\ \hat{\boldsymbol{\gamma}} &= \hat{\mathbf{R}}^{-1} \left(\mathbf{R}_0 \boldsymbol{\gamma}_0 + \frac{1}{\sigma_{11}} \sum_{i=1}^{N} \frac{\varepsilon_{i1}}{\kappa(\lambda_i)} \Phi^{-1} \boldsymbol{\varepsilon}_{i(-1)} \right). \end{aligned}$$

For sampling from the Wishart distribution, we employ the algorithm of Odell and Feiveson (1966). The sampling of $\boldsymbol{\gamma}$ from its conditional distribution is straightforward.

By using this reparametrization of Σ, one has to be careful in choosing the prior parameters of $\boldsymbol{\gamma}$ and Φ^{-1}. The goal is to choose $(\boldsymbol{\gamma}_0, \mathbf{R}_0^{-1}, \upsilon_0, \Phi_0^{-1})$ so that the induced prior on Σ is as diffuse as possible. As suggested by McCulloch *et al.* (2000), if we want the prior on the correlations to be spread out on the interval $[-1, 1]$, the choice $\mathbf{R}_0^{-1} = (d/4)\mathbf{I}$ and $\Phi_0 = (3d/4)(\upsilon_0 - J + 1)\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, J is the dimension of Σ and d is the expectation of the diagonal elements of Σ, seems to work well. In practice $\boldsymbol{\gamma}_0$ is usually chosen to be $\mathbf{0}$, and υ_0 is chosen to be small relative to the sample size.

Sampling $\boldsymbol{\lambda}|\cdot$

For $i = 1, \ldots, N$, the λ_i are conditionally independent. So the sampling of $\boldsymbol{\lambda}$ can be made through component-wise drawing of λ_i from $\pi(\lambda_i | y_i, \boldsymbol{\beta}, \Sigma, \mathbf{w}_i)$. In general, as can be seen from (7),

$$\pi(\lambda_i | y_i, \boldsymbol{\beta}, \Sigma, \mathbf{w}_i) \propto N_J(\mathbf{w}_i | \mathbf{x}_i \boldsymbol{\beta}, \kappa(\lambda_i)\Sigma) \pi(\lambda_i), \tag{11}$$

where the scale parameter $\kappa(\lambda_i)$ and mixing distribution $\pi(\lambda_i)$ for the models have been specified in Table 1. It is easy to see that under MNT,

$$\lambda_i|y_i, \boldsymbol{\beta}, \Sigma, \mathbf{w}_i \sim \mathcal{G}\left(\frac{J+v}{2}, \frac{1}{2}\left\{(\mathbf{w}_i - \mathbf{x}_i\boldsymbol{\beta})'\Sigma^{-1}(\mathbf{w}_i - \mathbf{x}_i\boldsymbol{\beta}) + v\right\}\right).$$

For other models, however, no standard distributions exist for $\lambda_i|y_i, \boldsymbol{\beta}, \Sigma, \mathbf{w}_i$. We deal with this problem by using the Metropolis-Hastings algorithm (Hastings, 1970; Chib and Greenberg, 1995). Chen and Dey (1998) give a comprehensive treatment. For completeness, we list the target and proposal densities used in the Metropolis-Hastings algorithms in Table 2. In the case of MNEP, to draw from the inverse normal distribution $\mathcal{IN}\left((4c_0\boldsymbol{\varepsilon}_i'\Sigma^{-1}\boldsymbol{\varepsilon}_i)^{-\frac{1}{2}}, \frac{1}{2}\right)$, we adopt the algorithm of Devroye (1986).

(insert Table 2 around here)

4. Model Comparison

Now we consider the question of comparing competing discrete choice models. The models may be compared in a Bayesian context by Bayes factors, or ratios of model marginal likelihoods (Kass and Raftery, 1995; Chib, 1995). Assuming that each model is *a priori* equally likely, we simply choose the model which yields the largest marginal likelihood.

Suppose there are L models $\mathcal{M}_l$ $(l = 1, \ldots, L)$ under our consideration. Following Chib (1995), the marginal likelihood of model $\mathcal{M}_l$ can be written as

$$m(\mathbf{y}|\mathcal{M}_l) = \frac{f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\beta}_l, \Phi_l^{-1}, \boldsymbol{\gamma}_l)\pi(\boldsymbol{\beta}_l, \Phi_l^{-1}, \boldsymbol{\gamma}_l|\mathcal{M}_l)}{\pi(\boldsymbol{\beta}_l, \Phi_l^{-1}, \boldsymbol{\gamma}_l|\mathcal{M}_l, \mathbf{y})}. \tag{12}$$

Since in (12), the left-hand side is free of the parameters, the marginal likelihood may be calculated by evaluating the terms on the right-hand side at any point $(\boldsymbol{\beta}_l^*, \Phi_l^{-1*}, \boldsymbol{\gamma}_l^*)$ for a fixed model $\mathcal{M}_l$. In practice, a high density point such as the posterior mean is usually used. Thus, suppressing the dependence on $\mathcal{M}_l$, we can rewrite the marginal likelihood on the log scale:

$$\begin{aligned} \log m(\mathbf{y}) \quad = \quad & \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) \\ & + \log \pi(\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) - \log \pi(\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*|\mathbf{y}). \end{aligned} \tag{13}$$

To obtain an estimate of the marginal likelihood on the log scale, we need to estimate the first term and the last term on the right-hand side of (13). We explain this estimation below in turn.

Estimate $f(y_i|\boldsymbol{\beta}^*, \Phi^{*-1}, \boldsymbol{\gamma}^*)$

Observe that

$$f(y_i|\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) = \int\int f(y_i|\lambda_i, \mathbf{w}_i, \boldsymbol{\beta}^*, \Phi^{*-1}, \boldsymbol{\gamma}^*) \\ \times f(\mathbf{w}_i|\lambda_i, \boldsymbol{\beta}^*, \Phi^{*-1}, \boldsymbol{\gamma}^*)\pi(\lambda_i)d\lambda_i d\mathbf{w}_i.$$

The estimation of $f(y_i|\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*)$ can be done by repeating the following two steps for a total of S times.

Step 1. Generate $\lambda_i^{(s)} \sim \pi(\lambda_i)$ and then simulate $\mathbf{w}_i^{(s)} \sim N_J(\mathbf{x}_i\boldsymbol{\beta}^*, \kappa(\lambda_i^{(s)})\Sigma^*)$ where Σ^* can be constructed by (10) with Φ^{-1} and $\boldsymbol{\gamma}$ replaced by Φ^{-1*} and $\boldsymbol{\gamma}^*$;

Step 2. Calculate $f_i(y_i|\boldsymbol{\beta}^*, \Sigma^*, \mathbf{w}_i^{(s)}, \lambda_i^{(s)}) = I(\mathbf{w}_i^{(s)} \in \mathcal{A}_i)$. This quantity is either 0 or 1.

Take the average of the quantities in step 2 to obtain an estimate of $f(y_i|\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*)$.

In generating $\lambda_i^{(s)}$ from $\pi(\lambda_i)$, Devroye's algorithms (Devroye, 1986) are used to generate from the Kolmogorov-Smirnov distribution (for MNLOG model) and the positive stable distribution (for MNS, MNEP and MNDE models).

Estimate $\pi(\boldsymbol{\beta}^*, \Phi^{-1*}, \boldsymbol{\gamma}^*|\mathbf{y})$

First observe

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

The marginal density $\pi(\Phi^{-\mathbf{1}*}|\mathbf{y})$ ordinate can be estimated by the Rao-Blackwell device (Tanner and Wong 1987; Gelfand and Smith 1990)

$$\hat{\pi}(\Phi^{-\mathbf{1}*}|\mathbf{y}) = G^{-1}\textstyle\sum_{g=1}^{G} \pi(\Phi^{-\mathbf{1}*}|\mathbf{y}, \boldsymbol{\beta}^{(g)}, \boldsymbol{\gamma}^{(g)}, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)})$$

where

$$\{\boldsymbol{\beta}^{(g)}, \boldsymbol{\gamma}^{(g)}, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)}\} \sim \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y})$$

is the gth draw from the effective MCMC simulation with size G.

Next consider the estimation of the conditional posterior ordinate

$$\pi(\boldsymbol{\gamma}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}) = \int\int\int \pi(\boldsymbol{\gamma}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda})\pi(\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*})d\boldsymbol{\beta}d\mathbf{w}d\boldsymbol{\lambda} \tag{14}$$

where $\pi(\boldsymbol{\gamma}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda})$ is the full conditional density of $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ and $\Phi^{-1} = \Phi^{-\mathbf{1}*}$. Here a different approach is necessary since $\{\boldsymbol{\beta}^{(g)}, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)}\}$ is not a draw from the distribution $[\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*}]$. Chib (1995) suggests we continue sampling for an additional G (not necessarily the same G as above) iterations with the conditional distributions $\pi(\boldsymbol{\gamma}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda})$, $\pi(\boldsymbol{\beta}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\lambda})$, $\pi(\mathbf{w}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$, and $\pi(\boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w})$. It follows from MCMC theory that $\{\boldsymbol{\beta}^{(g)}, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)}\}$ is now a draw from the distribution $[\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*}]$. Consequently,

$$\hat{\pi}(\boldsymbol{\gamma}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}) = G^{-1}\sum_{g=1}^{G} \pi(\boldsymbol{\gamma}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\beta}^{(g)}, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)})$$

is a simulation consistent estimate of (14).

The same argument applies to the estimation of $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*)$ as well. Now,

$$\pi(\boldsymbol{\beta}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*) = \int\int \pi(\boldsymbol{\beta}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*, \mathbf{w}, \boldsymbol{\lambda})\pi(\mathbf{w}, \boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*)d\mathbf{w}d\boldsymbol{\lambda}. \tag{15}$$

We make the additional G draws from the reduced run with components $\pi(\boldsymbol{\beta}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*, \boldsymbol{\lambda})$, $\pi(\mathbf{w}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \boldsymbol{\lambda})$, and $\pi(\boldsymbol{\lambda}|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{w})$. The simulation consistent estimate of (15) is then

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*) = G^{-1}\sum_{g=1}^{G} \pi(\boldsymbol{\beta}^*|\mathbf{y}, \Phi^{-\mathbf{1}*}, \boldsymbol{\gamma}^*, \mathbf{w}^{(g)}, \boldsymbol{\lambda}^{(g)}).$$

5. Example with Travel Mode Choice Data

We illustrate the new class of parametric choice models and carry out the model comparison techniques using real travel mode choice data previously analyzed by Greene (1997, pp. 868-869). The data consist of 210 observations on choices among four travel modes: air, train, bus, and car. The choice-specific variables used in this example include GC and TTME. The former is a continuous measure of the generalized cost of the travel constructed from multiplying measures of in-vehicle cost and a wage-like measure with amount of time spent traveling and the latter is the terminal waiting time for each travel mode (0 for car). Household income (HINC) is used as a chooser-specific variable. Three choice-specific intercepts are also included in this example. We use car as the reference alternative.

We take the prior of $\boldsymbol{\beta}$ as a 5-dimensional normal distribution centered at $\mathbf{0}$ with precision matrix $0.0001\mathbf{I}$. This choice ensures a fairly diffuse prior for $\boldsymbol{\beta}$. Ideally, we also would like to put such a diffuse prior on the covariance matrix Σ. This needs some elegant balancing work on choices of the prior distribution of $\boldsymbol{\gamma}$ and $\Phi^{-\mathbf{1}}$. Following the discussion of McCulloch *et al.* (2000), we choose $N_2(\mathbf{0},\mathbf{I})$ as the prior for $\boldsymbol{\gamma}$, and $W_2(22,\mathbf{I}/60)$ as the prior for $\Phi^{-\mathbf{1}}$. These choices amount to prior expectation of the diagonal elements of Σ as 4 ($d=4$), and a fairly diffuse prior on the correlations on the interval $[-1,1]$. For the parameters in the mixture distributions, we choose $v=8.0$ (for MNT) and $\alpha=0.75$ (for MNS and MNEP).

After an initial burn-in of 1000 iterations, we use the next 100,000 Gibbs iterations to compute the posterior estimates for all six models. The posterior means and posterior standard deviations of the coefficient parameters are shown in Table 3 under each model. Although the estimated coefficients are slightly different for different models, they all have the same signs and similar patterns.

The 100,000 original Gibbs iterations used to estimate parameters is also used to compute $\hat{\pi}(\Phi^{-\mathbf{1}*}|\mathbf{y})$. Then we separately generate 100,000 iterations from the first and second reduced runs to estimate $\pi(\boldsymbol{\gamma}^*|\mathbf{y},\Phi^{-\mathbf{1}*})$ and $\pi(\boldsymbol{\beta}^*|\mathbf{y},\Phi^{-\mathbf{1}*},\boldsymbol{\gamma}^*)$, respectively. In the algorithm to estimate $f(y_i|\boldsymbol{\beta}^*,\Phi^{*-1},\boldsymbol{\gamma}^*)$, we generate a random sample of size 10,000 from $\pi(\lambda_i)$. The resulted marginal likelihood estimates on the log scale are reported in the last row of Table 3. It can be seen that the MNS model fits this travel mode choice data the best, followed by the MNDE, MNLOG, MNT, and MNEP. Surprisingly enough, the MNP model is the worst among the six proposed models.

(insert Table 3 around here)

6. Extension to Panel Data

So far, we have assumed that the observations in the choice data are independent. This assumption is usually valid if each observation is from a different subject and there is no "clustering" pattern among subjects, as it is in the travel mode example above. However, many times the data may come in with an obvious violation of this assumption. Failing to take into account the dependence in the data will likely result in biased estimates. In this section, we extend our models based on SMMVN to explicitly accommodate heterogeneity through a random coefficients approach.

Suppose we observe the choice data of subject i at occasion t, $(y_{it},\mathbf{x}_{it})$, $i=1,\ldots,N$, $t=1,\ldots,T_i$. For ease of notation, let us assume the $\mathbf{x}_{it}$ are already in their difference form. Consider the latent utility model based on SMMVN for fixed i and t:

$$\begin{aligned}
&\mathbf{w}_{it}=\mathbf{x}_{it}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_{it},\\
&\boldsymbol{\varepsilon}_{it}|\lambda_{it}\sim N_J(\mathbf{0},\kappa(\lambda_{it})\,\Sigma),\\
&\lambda_{it}\sim\pi(\lambda_{it}),
\end{aligned}$$

while the connection between the observed choices and the latent utilities is

$$y_{it} = \begin{cases} j & \text{if } w_{itj} > \max(\mathbf{w}_{it(-j)}, 0) \\ J+1 & \text{if } \mathbf{w}_{it} < 0. \end{cases}$$

The set-valued inverse of the above mapping can be defined similarly as in (6). That is

$$\mathcal{A}_{it} = \begin{cases} \{\mathbf{w}_{it} : w_{itj} > \max(\mathbf{w}_{it(-j)}, 0)\} & \text{if } y_{it} = 1, \ldots, J \\ \{\mathbf{w}_{it} : \mathbf{w}_{it} < 0\} & \text{if } y_{it} = J+1. \end{cases}$$

Again, $\mathcal{A}_{it}$ depends on the data y_{it}.

In this extended model formula, we specify a subject-specific regression coefficient $\boldsymbol{\beta}_i$ to account for the heterogeneity feature of the data. We assume that $\boldsymbol{\beta}_i$ is a realization of a multivariate normal distribution with mean $\boldsymbol{\beta}_0$ and variance $\mathbf{B}_0^{-1}$, namely, $\boldsymbol{\beta}_i \overset{iid}{\sim} N_q(\boldsymbol{\beta}_0, \mathbf{B}_0^{-1})$. We postulate that the hyperparameters $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ follow a multivariate normal-Wishart distribution (Bernardo and Smith, 1994), i.e.

$$\pi(\boldsymbol{\beta}_0, \mathbf{B}_0) = N_q(\boldsymbol{\beta}_0 | \boldsymbol{\beta}_{00}, (n_0 \mathbf{B}_0)^{-1}) W_q(\mathbf{B}_0 | \upsilon_{00}, \mathbf{B}_{00}^{-1})$$

where $n_0, \boldsymbol{\beta}_{00}, \mathbf{B}_{00}^{-1}$, and υ_{00} are known quantities. The simulation method we devised in Section 3 can be readily extended to this situation. We list the full conditional distributions here.

- $\pi(\Phi^{-1}|\cdot) = W_{J-1}\left(\hat{\upsilon}, \hat{\Phi}^{-1}\right)$, where $\hat{\upsilon} = \upsilon_0 + \sum_{i=1}^{N} T_i$ and

$$\hat{\Phi} = \Phi_0 + \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{1}{\kappa(\lambda_{it})} \left(\boldsymbol{\varepsilon}_{it(-1)} - \frac{\varepsilon_{it1}}{\sigma_{11}} \boldsymbol{\gamma}\right) \left(\boldsymbol{\varepsilon}_{it(-1)} - \frac{\varepsilon_{it1}}{\sigma_{11}} \boldsymbol{\gamma}\right)';$$

- $\pi(\boldsymbol{\gamma}|\cdot) = N_{J-1}(\hat{\boldsymbol{\gamma}}, \hat{\mathbf{R}}^{-1})$ with

$$\begin{aligned} \hat{\mathbf{R}} &= \mathbf{R}_0 + \frac{1}{\sigma_{11}^2} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\varepsilon_{it1}^2}{\kappa(\lambda_{it})} \Phi^{-1}, \\ \hat{\boldsymbol{\gamma}} &= \hat{\mathbf{R}}^{-1} \left(\mathbf{R}_0 \boldsymbol{\gamma}_0 + \frac{1}{\sigma_{11}} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\varepsilon_{it1}}{\kappa(\lambda_{it})} \Phi^{-1} \boldsymbol{\varepsilon}_{it(-1)} \right). \end{aligned}$$

- $\pi(\boldsymbol{\beta}_i|\cdot) = N_q(\hat{\boldsymbol{\beta}}_i, \hat{\mathbf{B}}_i^{-1})$, $i = 1, \ldots, N$, where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_i &= \hat{\mathbf{B}}_i^{-1} \left(\mathbf{B}_0 \boldsymbol{\beta}_0 + \sum_{t=1}^{T_i} \frac{1}{\kappa(\lambda_{it})} \mathbf{x}_{it}' \Sigma^{-1} \mathbf{w}_{it} \right), \\ \hat{\mathbf{B}}_i &= \mathbf{B}_0 + \sum_{t=1}^{T_i} \frac{1}{\kappa(\lambda_{it})} \mathbf{x}_{it}' \Sigma^{-1} \mathbf{x}_{it}. \end{aligned}$$

- $\pi(w_{itj}|\cdot)$ is proportional to

$$\begin{aligned} N(m_{itj}, \tau^2_{itj})[I(y_{it} &= j)I(w_{itj} > \max(\mathbf{w}_{it(-j)}, 0)) \\ +I(y_{it} &\neq j)I(w_{itj} < \max(\mathbf{w}_{it(-j)}, 0))], \end{aligned}$$

  for $i = 1, \ldots, N$, $t = 1, \ldots, T_i$ and $j = 1, \ldots, J$, where

$$\tau^2_{itj} = \frac{\kappa(\lambda_{it})}{g_{jj}},$$

$$m_{itj} = \mathbf{x}'_{itj}\boldsymbol{\beta}_i - \frac{\mathbf{g}'_{j(-j)}}{g_{jj}}(\mathbf{w}_{it(-j)} - \mathbf{x}_{it(-j)}\boldsymbol{\beta}_i).$$

- $\pi(\lambda_{it}|\cdot) = \mathcal{G}\left(\frac{J+v}{2}, \frac{1}{2}\left\{(\mathbf{w}_{it} - \mathbf{x}_{it}\boldsymbol{\beta}_i)'\Sigma^{-1}(\mathbf{w}_{it} - \mathbf{x}_{it}\boldsymbol{\beta}_i) + v\right\}\right)$, $i = 1, \ldots, N$, $t = 1, \ldots, T_i$ and $j = 1, \ldots, J$, for MNT.

  For other models involving the generation of the λ_{it}, the sampling scheme is the same as that used in the nonpanel case except that the subscripts for $\mathbf{w}$ and $\mathbf{x}$ are now it (i before), and that $\boldsymbol{\beta}$ now has subscript i.

- $\pi(\boldsymbol{\beta}_0|\cdot) = N_q(\hat{\boldsymbol{\beta}}_0, \hat{\mathbf{B}}_0^{-1})$, where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_0 &= (n_0 + N)^{-1}(n_0\boldsymbol{\beta}_{00} + N\overline{\boldsymbol{\beta}}), \ \overline{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i \\ \hat{\mathbf{B}}_0 &= (n_0 + N)\mathbf{B}_0. \end{aligned}$$

- $\pi(\mathbf{B}_0|\cdot) = W_q(\hat{v}_{00}, \hat{\mathbf{B}}_{00}^{-1})$, where $\hat{v}_{00} = N + v_{00}$, and

$$\hat{\mathbf{B}}_{00} = \mathbf{B}_{00} + \sum_{i=1}^{N}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)'.$$

To implement model comparison, we estimate the marginal density in the log scale at the posterior mean $(\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*)$

$$\begin{aligned} \log(m(\mathbf{y})) &= \log f(\mathbf{y}|\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) + \log f(\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) \\ &\quad - \log \pi(\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*|\mathbf{y}) \end{aligned}$$

where we have suppressed the dependence on model $\mathcal{M}_l$. To estimate the first term, observe that

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) = \prod_{i=1}^{N}\prod_{t=1}^{T_i}\int\int\int & f(y_{it}|\lambda_{it}, \mathbf{w}_{it}, \boldsymbol{\beta}_i, \boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) \\ & \times f(\mathbf{w}_{it}|\lambda_{it}, \boldsymbol{\beta}_i, \boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) f(\boldsymbol{\beta}_i|\lambda_{it}, \boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) \\ & \times f(\lambda_{it}|\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*) d\lambda_{it} d\mathbf{w}_{it} d\boldsymbol{\beta}_i. \end{aligned}$$

So, for each combination of i and t, we perform the following two steps for a total of $s = 1, \ldots, S$ times:

Step 1. Generate $\lambda_{it}^{(s)} \sim \pi(\lambda_{it})$, $\boldsymbol{\beta}_i^{(s)} \sim N_q(\boldsymbol{\beta}_0^*, \mathbf{B}_0^{-1*})$ and $\mathbf{w}_{it}^{(s)} \sim N_J(\mathbf{x}_{it}\boldsymbol{\beta}_i^{(s)}, \kappa(\lambda_{it}^{(s)})\Sigma^*)$;

Step 2. Calculate $f(y_{it}|\boldsymbol{\beta}_i^*, \Sigma^*, \mathbf{w}_{it}^{(s)}, \lambda_{it}^{(s)}) = I(\mathbf{w}_{it}^{(s)} \in \mathcal{A}_{it})$.

Taking the average of the quantities in step 2 gives rise an estimate of $f(y_{it}|\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*)$.

To estimate the posterior ordinates $\pi(\boldsymbol{\beta}_0^*, \mathbf{B}_0^*, \Phi^{-1*}, \boldsymbol{\gamma}^*|\mathbf{y})$, we rewrite it as

$$\pi(\mathbf{B}_0^*|\mathbf{y})\boldsymbol{\pi}(\boldsymbol{\beta}_\mathbf{0}^*|\mathbf{B}_\mathbf{0}^*, \mathbf{y})\boldsymbol{\pi}(\Phi^{-1*}|\mathbf{B}_0^*, \boldsymbol{\beta}_\mathbf{0}^*, \mathbf{y})\boldsymbol{\pi}(\boldsymbol{\gamma}^*|\boldsymbol{\beta}_0^*, \Phi^{-1*}, \mathbf{B}_0^*, \mathbf{y}).$$

These terms can be estimated either directly from the original MCMC output or from some reduced Gibbs, in the same spirit as in Section 4.

7. Example with Panel Brand Choice Data

We use a brand choice data previously analyzed by Jain *et al.* (1994) to demonstrate the extended version of the new class of models based on SMMVN. For illustrative purposes, we report the results only for the MNP, MNT, and MNLOG models. We also skip the model comparison.

This data set consists of purchases of yogurts by a panel of 100 households in Springfield, Missouri. Collected by optical scanners over a period of about two years, the data contain information on, for each purchase made by each household in the panel, the brand purchased, store environment (e.g., prices of all brands in the product category and variables such as special displays of brands in the store), and the marketing environment (e.g., presence of absence of newspaper feature advertisements). There are four brands of yogurts: Yoplait, Dannon, Weight Watchers, and Hiland, with market share 34%, 40%, 23% and 3%, respectively. In their paper, Jain *et al.* studied this data set under a random coefficients multinomial logit model, with the (discretized) distribution of the random coefficients arbitrary and estimated by the data. The two covariates used in their models are price (for the brand purchased) and feature (binary variable indexing whether there is a newspaper feature advertisement for the brand). Three alternative-specific intercepts are also included. Brand Hiland was used as the base. One of their basic findings is that the models with random coefficients on either intercepts only or both intercepts and slopes significantly improve model performance (based on AIC and BIC) over the one without random coefficients at all.

To illustrate our models with the yogurt data, we estimate the same effects as they did. We keep the same mixture distribution as in the travel mode choice example and also use the same priors for $\boldsymbol{\gamma}$ and Φ^{-1}. Furthermore, we let the prior parameters in the normal-Wishart distribution be

$\boldsymbol{\beta}_{00} = \mathbf{0}$, $n_0 = 5$, $v_{00} = 5$ and $\mathbf{B}_{00} = 0.0001\mathbf{I}$. We use 200,000 MCMC cycles beyond 3000 burn-in iterations to obtain the posterior distribution of the regression parameters $\boldsymbol{\beta}_i$, $i = 1, \ldots, N$. In Table 4 we report the posterior mean and standard deviation of the averaged $\boldsymbol{\beta}_i$ ($\sum_{i=1}^{N} \boldsymbol{\beta}_i / N$).

In general, all three models indicate that feature advertisement has a small positive effect on the choice probability while price has a strong opposite impact. Also the preference ordering of the brands are the same across the three models; Yoplait is most preferred, followed by Dannon, Weight Watchers, and Hiland.

(insert Table 4 around here)

8. Discussion

In this paper, we have proposed a rich class of parametric models for analyzing discrete choice data. Based on the scale mixtures of multivariate normal (SMMVN) distributions, the new class of models make it possible for researchers analyzing discrete choice data to apply models other than multinomial logit (MNL) or multinomial probit (MNP). A model selection method, like the one we used, can be then applied to select the best model for particular data. As we have shown above, the new models can be straightforwardly implemented under a Bayesian framework due to their special connection to MNP. We also demonstrated that the new class of models can be readily extended to panel data.

One more word about the justification of using SMMVN distributions on the error term. A subject's choice y_i from the J alternatives can be thought of as the result of sequential choices: the subject decides whether to choose the first alternative ($y_{i1} = 0$ or 1), then the second alternative ($y_{i2} = 0$ or 1), and so on, constrained by the condition $\sum_{j=1}^{J} y_{ij} = 1$. So, in a way, the single data y_i from a subject can be viewed as J (correlated) binary observations from the same subject. Thus the various distributions we put on $\mathbf{w}_i$ play similar role as various link functions do in the longitudinal binary models. Therefore, a new class of distributions for the error term has the effect of enriching the link functions.

There are some future work we can carry out. First, in the above modeling of MNT, MNS, and MNEP, we have only considered the cases where the parameters of mixture distribution $\pi(\lambda)$ are fixed. We did this simply to limit model complexity in the current paper. Future work could consider a prior for those parameters and let the data determine their most probable values.

The second future direction is to consider some other model comparison criteria. The Bayes factor is certainly useful in some contexts. However other Bayesian model comparison methods, such as the predictive distribution, might render more robust results. We also note that even by using the Bayes factor, with the advances of Chib and Jeliazkov (2000), nonconjugate prior distributions can be considered.

Appendix: Some Distributions and Their Notations

- Normal distribution, $N_q(x|\boldsymbol{\theta}, \Sigma)$, $\mathrm{E}(X) = \boldsymbol{\theta}$, $\mathrm{Var}(X) = \Sigma$. The density of it is denoted by $\phi_q(\boldsymbol{\theta}, \Sigma)$.

- Gamma distribution, $\mathcal{G}(x|\alpha, \beta)$, $\mathrm{E}(X) = \frac{\alpha}{\beta}$, $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

- Inverse Gamma, $\mathcal{IG}(x|\alpha, \beta)$, $\mathrm{E}(X) = \frac{\beta}{\alpha-1}$ if $\alpha > 1$, $\mathrm{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$.

- Kolmogorov-Smirnov distribution, $\pi_{KS}(x)$, density

$$8\sum_{k=1}^{\infty}(-1)^{k+1}k^2 x \exp(-2k^2x^2).$$

- Standardized positive stable distribution, $\pi_{PS}(x|\alpha, 1)$, density

$$\pi_{PS}(x|\alpha, 1) = ax^{-(a+1)}\int_0^1 s(u)\exp\{-\frac{s(u)}{x^a}\}du, \ x > 0,$$

where

$$a = \frac{\alpha}{1-\alpha} \quad \text{and} \quad s(u) = \left(\frac{\sin(\alpha\pi u)}{\sin(\pi u)}\right)^a \left(\frac{\sin[(1-\alpha)\pi u]}{\sin(\pi u)}\right), \ \alpha \in (0, 1).$$

- Inverse normal distribution, $\mathcal{IN}(x|\mu, \sigma)$, density

$$\sqrt{\frac{\sigma}{2\pi x^3}}\exp\left\{-\frac{\sigma(x-\mu)^2}{2\mu^2 x}\right\}.$$

- Wishart distribution, $W_J(\Sigma|\nu, \Phi^{-1})$, mean $\nu\Phi^{-1}$, density proportional to

$$|\Sigma|^{\frac{v-J-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}(\Phi\Sigma)\right).$$

References

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

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. and Mallows, C. (1974). Scale mixtures of normal distributions. *Journal of Royal Statistical Society,* **Ser. B, 36**, 99-102.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Wiley, Chichester, England.

Chen, M. and Dey, D. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhya,* **Spl. Ser A**, **60**, 322-343.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association,* **99**, 1313-1320.

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

Chib, S. and Jeliazkov, I. (2000). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association,* forthcoming.

Choy, S. (1995). *Robust Bayesian Analysis Using Scale Mixture of Normals Distributions.* Ph.D. dissertation. Imperial College, London.

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

Gelfand, A. and Smith, A.F.M. (1990). Sampling-based approaches to calculation 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 and the evaluation of constraint probabilities. In *Proceedings of 23rd Symposium on the Interface Between Computing Science and Statistics*, ed. E. Kermanidas, 571-578.

Greene, W. (1997). *Econometric Analysis.* Prentice-Hall, New Jersey.

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

Hausman, J. and Wise, D. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica,* **45**, 319-339.

Jain, D., Vilcassim, N., and Chintagunta, P. (1994). A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics,* **12**, 317-328.

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

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

Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics,* **3**, 205-228.

Matzkin, R.L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica*, **59**, 1315-1327.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics,* **64**, 207-240.

McCulloch, R., Polson, N., and Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics,* forthcoming.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics,* ed. P. Zarembka, Academic Press, New York.

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

Odell, P.L. and Feiveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association,* **61**, 199-203.

Schmit, P., and Strauss, R. (1975). The prediction of occupation using multinomial logit models. *International Economic Review*, **16**, 471-486.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association,* **82**, 528-550.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, **74**, 646-648.

ZHEN CHEN AND LYNN KUO

DEPARTMENT OF STATISTICS, U-120

UNIVERSITY OF CONNECTICUT

STORRS, CT 06269

USA

email: zhen@stat.uconn.edu

email: lynn@stat.uconn.edu

Table 1: Some Members of the Scale Mixtures of Multivariate Normal Distributions and the Corresponding Discrete Choice Models (See Appendix for the Notation)

SMMVN	$\kappa(\lambda)$	$\pi(\lambda)$	Model
Multivariate normal	1	$\pi(\lambda=1)=1$	MNP
Multivariate t	$1/\lambda$	$\mathcal{G}(v/2, v/2)$	MNT
Multivariate logistic	$4\lambda^2$	$\pi_{KS}(\lambda_i)$	MNLOG
Multivariate stable	2λ	$\pi_{PS}(\lambda_i \vert \alpha, 1)$	MNS
Multivariate exponential power	$1/(2c_0\lambda)$, $c_0 = \Gamma(\frac{3}{2\alpha})/\Gamma(\frac{1}{2\alpha})$	$(1/\lambda)^{\frac{J}{2}}\,\pi_{PS}(\lambda_i \vert \alpha, 1)$	MNEP
Multivariate double exponential	$1/(4\lambda)$	$(1/\lambda)^{\frac{J}{2}}\,\pi_{PS}(\lambda_i \vert 1/2, 1)$	MNDE

Table 2: Conditional Posterior Densities and Proposal Densities for λ_i in the Implementation of Metropolis-Hastings Algorithms under MNLOG, MNS and MNEP Models (See Appendix for the Notation)

Model	Posterior density of λ_i (proportional to)	Proposal density
MNLOG	$\phi_J(\mathbf{w}_i \vert \mathbf{x}_i\boldsymbol{\beta}, 4\lambda_i^2\Sigma)\pi_{KS}(\lambda_i)$	$\lambda_i^{*2} \sim \mathcal{IG}(\frac{J+0.5}{2}, \frac{1}{8}\left[\varepsilon_i'\Sigma^{-1}\varepsilon_i + \frac{0.5}{0.75^2}\right])$
MNS	$\phi_J(\mathbf{w}_i \vert \mathbf{x}_i\boldsymbol{\beta}, 2\lambda_i\Sigma)\pi_{PS}(\lambda_i \vert a, 1)$	$\lambda_i^* \sim \mathcal{IG}(\frac{J+1}{2}, \frac{1}{4}\left[\varepsilon_i'\Sigma^{-1}\varepsilon_i + 1\right])$
MNEP	$\phi_J(\mathbf{w}_i \vert \mathbf{x}_i\boldsymbol{\beta}, 2\lambda_i\Sigma)\left(\frac{1}{\lambda i}\right)^{\frac{J}{2}}\pi_{PS}(\lambda_i \vert a, 1)$	$\lambda_i^* \sim \mathcal{IN}(\left[4c_0\varepsilon_i'\Sigma^{-1}\varepsilon_i\right]^{-\frac{1}{2}}, \frac{1}{2})$

Table 3: Posterior Means, Standard Deviations (in Parentheses) and Marginal Likelihoods (in the Last Row) of the Travel Mode Choice Data under Different Models

Parameter	MNP	MNT	MNLOG	MNS	MNEP	MNDE
Intercept: Air	3.564	4.633	7.283	9.059	4.162	5.763
	(0.572)	(0.756)	(1.179)	(1.568)	(0.652)	(0.909)
Intercept: Train	3.664	4.532	7.152	8.241	4.204	5.545
	(0.536)	(0.692)	(1.098)	(1.280)	(0.595)	(0.829)
Intercept: Bus	2.169	2.963	4.590	5.801	2.615	3.671
	(0.567)	(0.661)	(1.042)	(1.210)	(0.606)	(0.773)
TTME	-0.061	-0.082	-0.129	-0.162	-0.073	-0.103
	(0.007)	(0.011)	(0.017)	(0.024)	(0.009)	(0.013)
GC	-0.005	-0.006	-0.009	-0.008	-0.006	-0.007
	(0.003)	(0.004)	(0.006)	(0.005)	(0.003)	(0.004)
HINC: Air	0.000	0.002	0.003	0.005	0.001	0.003
	(0.007)	(0.008)	(0.013)	(0.013)	(0.008)	(0.01)
HINC: Train	-0.054	-0.061	-0.097	-0.098	-0.060	-0.074
	(0.013)	(0.016)	(0.026)	(0.027)	(0.015)	(0.020)
HINC: Bus	-0.021	-0.023	-0.034	-0.033	-0.022	-0.026
	(0.013)	(0.014)	(0.022)	(0.023)	(0.013)	(0.017)
log $m(\mathbf{y})$	-272.47	-266.53	-262.86	-249.42	-268.72	-258.35

Table 4: Posterior Means and Standard Deviations (in Parentheses) of the Average Regression Coefficients of the Yogurt Data under Different Models

Parameter	MNP	MNT	MNLOG
Intercept: Yoplait	3.026	3.366	4.966
	(0.262)	(0.309)	(0.489)
Intercept: Dannon	2.299	2.542	3.730
	(0.218)	(0.242)	(0.377)
Intercept: Weight Watchers	0.318	0.154	0.048
	(0.377)	(0.416)	(0.738)
Feature	0.613	0.678	0.979
	(0.170)	(0.170)	(0.259)
Price	-29.102	-33.722	-50.107
	(3.012)	(3.897)	(6.241)