

Flexible Choice Modelling based on Bayesian Nonparametric Mixed Multinomial Logit Models

LANCELOT F. JAMES AND JOHN W. LAU
Hong Kong University of Science and Technology
July 7, 2004

This paper develops the first fully implementable nonparametric estimation method for discrete choice models based on the Mixed Multinomial Logit (MMNL) model. McFadden and Train (2000) show that the class of MMNL choice models encompass all discrete choice models derived under the assumption of random utility maximization (RUM), subject to the identification of an unknown distribution G . Noting the mixture model description of the MMNL, we employ a Bayesian nonparametric approach, using the Dirichlet process and related processes as priors on the unknown mixing distribution G , to estimate the unknown choice probabilities and related functionals. Slightly different techniques for non-Panel and Panel data models are discussed. Our framework, in contrast to methods based on parametric specifications for G , allows for the full utilization of the flexibility of MMNL models. We provide a complete description of the posterior distribution and show how it is naturally related to posterior distributions for parametric models. For practical implementation, we describe efficient and relatively easy to use blocked Gibbs sampling procedures which share many similarities to MCMC procedures for Bayesian procedures when G is assumed to have a parametric form.

ADDRESS. Department of Information Systems and Management. The Hong Kong University of Science and Technology. Clear Water Bay, Kowloon, Hong Kong. Corresponding Author: Lancelot F. James.

KEY WORDS. Blocked Gibbs sampler, Choice models, Dirichlet process, Mixed Multinomial Logit, Random Utility Maximization, Stick-breaking priors.

1 Introduction

This paper develops the first fully nonparametric estimation method for discrete choice models. Our procedure utilizes a Bayesian mixture model framework, based on the Dirichlet process and related models, to exploit the richness and flexibility of the Mixed Multinomial Logit (MMNL) model. Before describing our approach we give some background on relevant models. Discrete choice models arise naturally in many fields of applications including marketing and transportation science. Such choice models are based on the neoclassical economic theory of random utility maximization (RUM). Popular choice models include the Multinomial Logit (MNL), Probit and Generalized Extreme Value (GEV) models. The MNL model yields choice probabilities with a simple tractable form derived based on the assumption of independent Gumbel errors. Due to this simple structure the model is a popular choice among practitioners. The MNL possesses the property of independence from irrelevant alternatives (IIA), see Luce (1959). While this property is advantageous when it is known to exist, it is not easily verified and is inappropriate in many situations. Additionally, the independence assumptions on the errors is often viewed as an undesirable and unrealistic property as this is contradictory to the logical notion that one's choices are interdependent. The Probit and GEV models have been proposed as alternatives to the MNL, which do not exhibit the IIA property and are models derived from dependent error structures. The Probit, which uses a multivariate normal assumption for the errors, is more widely applicable than the GEV model. However, in contrast to the MNL, choice probabilities based on the Probit model must be obtained via Monte Carlo procedures. A drawback of the above mentioned procedures is that they are not robust against model miss-specification.

The Mixed Multinomial Logit (MMNL) model first introduced by Cardell and Dunbar (1980) emerges as potentially the most attractive model. The recent book of Train (2003) gives a detailed discussion of this model. The general MMNL choice probabilities are defined by mixing a MNL model over a mixing distribution G . McFadden and Train (2000) establish the important result that in theory all RUM models can be captured by correct specification of G . Thus a robust approach now amounts to being able to employ statistical estimation methods based on a nonparametric assumption on G . However, statistical techniques have only been developed for the case where G is given a parametric form. The most popular choice is where G is specified to be a multivariate normal with unknown mean and covariance structure. This sub-class of MMNL models is often referred to simply as mixed logit models, here we will refer to them as Gaussian Mixed Logit (GML) models. The GML has become a popular choice with recent applications and discussions by, among others, Ben-Akiva, Bolduc and Walker (2001), Bhat (1998), Brownstone and Train (1999), Erdem (1996), Srinivasan and Mahmassani (2000), and Walker (2001). Additionally Dube, Chintagunta, Bronnenburg, Goettler, Petrin, Sudhir, and Zhao (2002), provide a discussion focused on applications to marketing. The GML, as well as general MMNL models require simulations either via simulated maximum likelihood methods or Bayesian parametric MCMC procedures. Train (2003) gives a thorough description of Bayesian parametric Markov Chain Monte Carlo (MCMC) procedures for the GML showing that such procedures, while somewhat sophisticated, are not overly complex and have good convergence properties relative to simulated maximum likelihood methods. Thus one might prefer a Bayesian parametric approach to estimate the unknown quantities in GML models. However, for more general parametric choices of G , such MCMC schemes are more complex with slower convergence rates. Thus from a practical point of view, with the exception of the GML and closely related models, it is not yet a simple matter to use arbitrary but otherwise

user specified parametric G . However, despite the attractive features of the GML, it, along with any other parametric model for G , is not robust against miss-specification. Hence so far one has not been able to utilize the full flexibility of the MMNL model.

Here we propose Bayesian nonparametric theoretical and computational techniques to estimate the choice probabilities under vague assumptions on G . We divide our discussion into methods for non-Panel and Panel data models. Specifically for non-Panel data models we use as a prior for G , a mixture of Dirichlet processes as discussed in Antoniak (1974). Methods for Panel data are then proposed using a Dirichlet process mixture of normals. Our models are analogous to the Bayesian density estimation procedure discussed in Lo (1984). We will show that the mixture of Dirichlet process framework offers a great deal of flexibility as one is able to embed a parametric model such as as GML within the nonparametric framework. One feature of this is that one is able to hedge their bets, as an appropriate guess of the parametric form will lead to good results even for small sample sizes. On the other hand, the Bayesian nonparametric framework dictates that eventually the data corrects for possible miss-specifications in a parametric model. We provide a complete description of the posterior distribution and show how it is naturally related to posterior distributions for parametric models.

The ability to utilize tractable parametric models, such as the GML, within a nonparametric framework is crucial from a practical viewpoint. For practical implementation, we describe efficient and relatively easy to use blocked Gibbs sampling procedures, recently developed in Ishwaran and Zarepour (2000) and Ishwaran and James (2001), which share many similarities to MCMC procedures for Bayesian procedures when G is assumed to have a parametric form. We provide a modest but illustrative simulation study which shows the flexibility and good performance of our procedure versus the parametric GML model.

2 Random Utility Models

Given a finite set of choices $\Phi = \{1, \dots, J\}$, it is assumed that each individual has a utility function

$$U_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij} \text{ for } j \in \Phi.$$

The values $\mathbf{X}_i = (\mathbf{x}, \dots, \mathbf{x}_{iJ})$ are observed covariates, where $\mathbf{x}_{ij} \in \mathcal{R}^q$ denotes the covariates associated with each choice $\{j\} \in \Phi$ and the coefficient $\boldsymbol{\beta}$ is an unknown (preference) vector in \mathcal{R}^q , and $(\varepsilon_{i1}, \dots, \varepsilon_{iJ})$ are random terms. Suppose that all U_{ij} are distinct, an individual makes a choice $\{j\}$ if $U_{ij} = \max\{U_{i1}, \dots, U_{iJ}\}$. The introduction of the random error terms ε_j represents the departure from classical economic utility models. Such models are referred to as random utility models (RUM). The random errors account for the discrepancy between the actual utility which is known by the chooser and that which is deduced by the experimenter who observes the \mathbf{X}_i and the choices made by an individual but not the actual U_{ij} . Hence the deterministic statement of choice j is replaced by the probability of choosing j . That is, the probabilistic version of such a statement is $P(U_{ij} = \max\{U_{i1}, \dots, U_{iJ}\})$ which is denoted as $P(\{j\}|\boldsymbol{\beta})$. For n individuals one observes the choice each individual makes and the covariates $(\mathbf{X}_1, \dots, \mathbf{X}_n)$. The analysis of such a model depends on the specifications of the errors. McFadden (1974) showed that the specification of independent Gumbel error terms leads to the the tractable Multinomial Logit (MNL) Model.

This representation is written as

$$P(\{j\}|\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}'_j\boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l\boldsymbol{\beta}\}} \text{ where } j \in \Phi,$$

where $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_J)$ denotes a set of covariates not necessarily associated with any observed individual. The simplicity of the MNL model makes it a popular choice in many applications. Another popular model is the Multinomial Probit model based on the assumption that the errors are from a multivariate normal distribution. The Probit model is often used in place of the MNL in situations where the (independence of irrelevant alternatives) IIA property of the MNL is undesirable, however simulations must be used to compute the choice probabilities. A drawback of the MNL and Probit models is that they are not robust with respect to error model misspecification.

2.1 Mixed Multinomial Logit (MMNL) models for non-Panel data

MacFadden and Train (2000) propose the usage of the Mixed Multinomial Logit (MMNL) Model. For non-Panel data, the MMNL is defined by assuming that $\boldsymbol{\beta}$ in the MNL model is random with an unknown (mixing) distribution G . For a set of covariates \mathbf{x} , the MMNL model is written as

$$P(\{j\}|G, \mathbf{x}) = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}'_j\boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l\boldsymbol{\beta}\}} G(d\boldsymbol{\beta}) \text{ where } j \in \Phi. \quad (1)$$

The significance of the MMNL is that in theory all RUM models can be captured by the specification of the otherwise unknown mixing distribution G . However, as mentioned in the introduction, statistical techniques have only been developed for the case where G has a specified parametric form. The most popular model is when G is specified to be a multivariate normal distribution with unknown mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\tau}$. We shall term this model a Gaussian mixed logit(GML) model. That is, the GML has the form

$$P(\{j\}|\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{x}) = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}'_j\boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l\boldsymbol{\beta}\}} \phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau}) d\boldsymbol{\beta}, \quad (2)$$

where $\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})$ represents a multivariate normal density with unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$. Here, based on a sample of size n one estimates the choice probabilities by estimating $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ via frequentist(classical) or Bayesian methods. The GML model is a popular since it is flexible and relatively easy to estimate via simulated maximum likelihood techniques or via Bayesian MCMC procedures. Other choices for G , include the lognormal and uniform distributions. Train (2003) discusses the merits and possible drawbacks of Bayesian MCMC procedures versus simulated maximum likelihood procedures for various choices of G . He notes that parametric Bayesian MCMC procedures based on multivariate normal or lognormal specifications for G converge significantly faster than procedures using G with bounded support. See McFadden and Train (2000) for more details. Despite the attractive nature of the GML it does not encompass all RUM models. In the next section we describe how a Bayesian nonparametric approach can be used to model the mixing distribution G , and hence the MMNL. This allows one to fully exploit the flexibility of the MMNL.

3 Bayesian MMNL models

A Bayesian nonparametric model for the MMNL is specified by placing a nonparametric prior on the mixing distribution G in (6). That is to say, we model G as a random probability measure which takes values over the space of probability measures on \mathcal{R}^d . This in turn defines prior models for the $P(\{j\}|G, \mathbf{x})$. We note clearly that we are not interested in estimating G , but rather using the nonparametric nature of the Dirichlet prior on G to estimate the $P(\{j\}|G, \mathbf{x})$ and related quantities. Specifically one can choose G to be a Dirichlet process [Ferguson (1973)] with shape parameter $\alpha H(\cdot)$, where α is a positive scalar and H is a probability measure on \mathcal{R}^d . This is analogous to the Bayesian density estimation framework discussed in Lo (1984). Denote the Dirichlet process law by $\mathcal{P}(dG|\alpha H)$. Formally, a random measure G is said to be a Dirichlet process with shape parameter $\alpha H(\cdot)$ if for each measurable partition (A_1, \dots, A_k) of \mathcal{R}^d , the vector $(G(A_1), \dots, G(A_k))$ is distributed as a Dirichlet random vector with parameters $(\alpha H(A_1), \dots, \alpha H(A_k))$. One can also describe the Dirichlet process via its representation as a stick-breaking process. That is, if $G \sim \mathcal{P}(dG|\alpha H)$, then it is representable as

$$G(d\boldsymbol{\beta}) = \sum_{k=1}^{\infty} p_k \delta_{Z_k}(d\boldsymbol{\beta}) \tag{3}$$

where $p_k = V_k \prod_{i=1}^{k-1} (1 - V_i)$ are random probabilities summing to one, such that (V_k) are independent beta $(1, \alpha)$ random variables and (Z_k) are iid random variables with distribution H , independent of the p_k . The distribution H allows the user to incorporate their prior belief about the unknown mixing distribution G . In particular, the prior predictive distribution is given by

$$E[G(A)] = \Pr\{\boldsymbol{\beta} \in A\} = H(A)$$

for all measurable sets A , where E denotes expectation. For instance, H can be set to have density $\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})$, which is the mixing measure that generates the GML in (2). The parameter α can be interpreted as one's strength of belief in H , with large values indicating more belief in the model. However, an important feature of the Bayesian nonparametric framework is that the observed data eventually dominates the model, thus correcting for misspecifications in H . An important characterization of the Dirichlet process via its prediction rule for G demonstrates this fact. Suppose that $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n|G$ are iid G and $G \sim \mathcal{P}(dG|\alpha H)$. Then the Dirichlet process is characterized by its Blackwell-MacQueen (1973) prediction rule given as,

$$\Pr\{\boldsymbol{\beta}_{n+1} \in du|\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n\} = \frac{\alpha}{\alpha + n} H(du) + \frac{1}{\alpha + n} \sum_{j=1}^n \delta_{\boldsymbol{\beta}_j}(du). \tag{4}$$

The prediction rule in (4) corresponds to the posterior mean of G given $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$. We see that for general n , the prediction rule is a mixture of the prior guess H and the empirical distribution of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$. Moreover, as n gets large, relative to the choice of α , (4) eventually converges to the same limit as the empirical distribution function. One can compare this with a parametric Bayesian framework for the GML where one might place priors on $\boldsymbol{\mu}, \boldsymbol{\tau}$. While one is able to get posterior estimates for the true values of $\boldsymbol{\mu}, \boldsymbol{\tau}$, inference is restricted to the assumption of the GML model. The flexibility of the Bayesian nonparametric approach allows one to choose H based on convenience and ease of use. As we shall show, one can utilize, for instance, the attractive features of GML models while still maintaining the robustness of a nonparametric approach. See Ishwaran and James (2001) for a general discussion on the larger class of stick-breaking priors.

In the case of the Dirichlet process, the parameters associated with H , for instance $\boldsymbol{\mu}$, and $\boldsymbol{\tau}$, are considered fixed. A method to introduce more flexibility in the model is to treat such parameters as random. This corresponds to the mixture of Dirichlet process models described in Antoniak (1974). Formally, G is said to be a mixture of Dirichlet processes if $G|\theta$ is a Dirichlet process with shape parameter αH_θ and θ is a random vector in a Euclidean space Θ with distribution $\pi(d\theta)$. The law of G is given by the mixture

$$\int_{\Theta} \mathcal{P}(dG|\alpha H_\theta)\pi(d\theta).$$

Equivalently, using (10), a mixture of Dirichlet processes is defined by specifying each $Z_k|\theta$ to be iid H_θ . The specification for G translates into a Bayesian model for the MMNL as,

$$P(\{j\}|G, \mathbf{x}) = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}'_j \boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l \boldsymbol{\beta}\}} G(d\boldsymbol{\beta}) = \sum_{k=1}^{\infty} p_k \frac{\exp\{\mathbf{x}'_j Z_k\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l Z_k\}}. \quad (5)$$

Notice that conditional on θ , a prior guess for the choice probabilities is

$$E[P(\{j\}|G, \mathbf{x})|\theta] = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}'_j \boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l \boldsymbol{\beta}\}} H_\theta(d\boldsymbol{\beta}) = P(\{j\}|H_\theta, \mathbf{x}) \text{ where } j \in \Phi. \quad (6)$$

Specifying $\theta = (\boldsymbol{\mu}, \boldsymbol{\tau})$ and $H_\theta(d\boldsymbol{\beta})$ to have density $\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})$ equates to a GML prior guess for $P(j|G, \mathbf{x})$. If $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ were observables from the distribution G then one could obtain the prediction rule for the choice probabilities given $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n$ as follows,

$$E[P(\{j\}|G, \mathbf{x})|\theta, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n] = \frac{\alpha}{\alpha + n} P(\{j\}|H_\theta, \mathbf{x}) + \sum_{i=1}^n \frac{1}{\alpha + n} \frac{\exp\{\mathbf{x}'_j \boldsymbol{\beta}_i\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l \boldsymbol{\beta}_i\}}. \quad (7)$$

Unfortunately $\boldsymbol{\beta}_i$'s are not observable, having an interpretation instead as missing values from the distribution G . This leads to a more complex posterior distribution and needs implementing computational procedures to draw from the posterior distribution.

One can exploit the description above to implement various computational procedures for approximating the posterior distribution of the choice probabilities $P(\{j\}|G, \mathbf{x})$ and related functionals. In principle the Pólya urn Gibbs samplers developed in Escobar (1994) and Escobar and West (1995) can be used. However, these methods often result in a slowly mixing Markov Chain. For conjugate models one can resort to the accelerated Pólya urn Gibbs sampler as described in MacEachern (1998), which generally improve mixing. However the present situation involves non-conjugate models for $\boldsymbol{\beta}$, which requires much more delicate modifications of those procedures. A more general Monte Carlo method is therefore needed to address these issues. In the next section, we discuss a flexible Gibbs sampling technique, the blocked Gibbs sampler developed in Ishwaran and Zarepour (2000) and Ishwaran and James (2001), that can be used in general. This procedure is fairly easy to use even in the presence of non-conjugate models and enjoys good mixing properties. For a systematic comparison of Pólya urn Gibbs sampling to blocked Gibbs sampling see Ishwaran and James (2001)

3.1 Blocked Gibbs sampling

The blocked Gibbs sampler for Dirichlet process models utilizes the fact that a truncated Dirichlet process, based on almost sure truncation which is exponentially accurate, discussed in for instance Ishwaran and Zarepour (2000) and Ishwaran and James (2001) serves as a good approximation to the Dirichlet process. That is, the blocked Gibbs procedure works by replacing the Dirichlet process specification for G with a truncated Dirichlet process. Since in the present setting G is modelled as a mixture of Dirichlet process, we work instead with a law which is a mixture of truncated Dirichlet process. The joint distribution of the augmented data can be expressed using a hierarchical model as follows:

$$\begin{aligned} \mathbf{Y}_i | \beta_i &\stackrel{\text{ind}}{\sim} \frac{\exp \left\{ \mathbf{x}'_{iY_i} \beta_i \right\}}{\sum_{l_i \in \Phi} \exp \left\{ \mathbf{x}'_{il_i} \beta_i \right\}}, \text{ for } i=1, \dots, n, \text{ and } Y_i \in \Phi, \\ \beta_i | G &\stackrel{\text{iid}}{\sim} G, \text{ for } i=1, \dots, n \\ G &\sim \mathcal{P}_N(dG | \alpha H_\theta), \\ \theta &\sim \pi(d\theta). \end{aligned} \tag{8}$$

That is, instead of the Dirichlet process we use the random probability measure,

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^N p_k \delta_{Z_k}(\cdot) \tag{9}$$

where the random vector $\mathbf{p} = \{p_1, \dots, p_N\}$ is constructed by the stick-breaking construction

$$p_1 = V_1 \text{ and } p_k = (1 - V_1) \cdots (1 - V_{k-1}) V_k, \quad k = 2, \dots, N \tag{10}$$

where V_1, V_2, \dots, V_{N-1} are iid $\text{Beta}(1, \alpha)$ random variables and we set $V_N = 1$ to ensure that $\sum_{k=1}^N p_k = 1$ and $Z_k | \theta$ are iid H_θ and θ has law $\pi(d\theta)$. The law of $G | \theta$ is referred to as a truncated Dirichlet process. Denote this conditional law for G as $\mathcal{P}_N(dG | \alpha H_\theta)$. Note, that for accuracy, N is actually chosen as a function of the sample size n , that is $N := N(n)$. Hence this method is not to be confused with a finite mixture model, but is rather more in line with a sieve based approach. According to Ishwaran and James (2001), one can choose N based on an \mathcal{L}_1 error bound for the approximation of conditional density of \mathbf{Y} given θ , which is described next. Let

$$\mu_N(\mathbf{Y} | \theta) = \int \left[\prod_{i=1}^n \int_{\mathcal{R}^d} L_i(Y_i, \beta_i) G(d\beta_i) \right] \mathcal{P}_N(dG | \alpha H_\theta)$$

denote the conditional density of \mathbf{Y} given θ and its limit is denoted as

$$\mu_\infty(\mathbf{Y} | \theta) = \int \left[\prod_{i=1}^n \int_{\mathcal{R}^d} L_i(Y_i, \beta_i) G(d\beta_i) \right] \mathcal{P}(dG | \alpha H_\theta)$$

THEOREM 1. [Ishwaran and James (2001)] If $\|\cdot\|_1$ denotes the \mathcal{L}_1 distance, then $\|\mu_N - \mu_\infty\|_1 \sim 4ne^{-(N-1)/\alpha}$. That is,

$$\int |\mu_N(\mathbf{Y} | \theta) - \mu_\infty(\mathbf{Y} | \theta)| d\mathbf{Y} \sim 4ne^{-(N-1)/\alpha}.$$

For example, if $\alpha = 2.5$, $N = 50$ and $n = 1000$, then $\|\mu_N - \mu_\infty\|_1 \sim 1.229952 \times 10^{-05}$. We use those values in our simulation study since the bound is small. The truncated Dirichlet approximation is effectively indistinguishable from one based on the Dirichlet process. See Ishwaran and James (2001) for more details.

Let $\mathbf{K} = \{K_1, \dots, K_n\}$ denote random variables such that conditional on \mathbf{W} each K_i is independent with distribution

$$\Pr\{K_i \in \cdot | \mathbf{W}\} = \sum_{k=1}^N W_k \delta_k(\cdot).$$

That is $\Pr\{K_i = k | \mathbf{W}\} = W_k$ for $k = 1, \dots, N$. A key feature used in the blocked Gibbs procedure is that if $\beta_1, \dots, \beta_n | G, \theta$ are draws from G , it follows that

$$\beta_i = Z_{K_i},$$

where the K_i act as random classification variables taking on possible values $\{1, \dots, N\}$. Here \mathbf{K} plays a similar role to \mathbf{p} . In this setting a sample β_1, \dots, β_n from G produces $n_0 \leq \min(n, N)$ distinct values. The blocked Gibbs algorithm is based on sampling $\mathbf{Z}, \mathbf{W}, \mathbf{K}, \theta$ from the distribution proportional to

$$\left[\prod_{i=1}^n L_i(Y_i, \beta_i) \right] \left[\prod_{i=1}^n \sum_{k=1}^N W_k \delta_{Z_k}(d\beta_i) \right] \pi(\mathbf{W}) \left[\prod_{k=1}^N H(dZ_k | \theta) \right] \pi(d\theta).$$

This augmented likelihood is an expression of the augmented density when $\mathcal{P}(dG | \alpha H_\theta)$ is replaced by $\mathcal{P}_N(dG | \alpha H_\theta)$. Before describing the algorithm we specify choices for H_θ and θ which agree with a natural choice for a parametric Bayesian treatment of the GML model.

Set $\theta = (\boldsymbol{\mu}, \boldsymbol{\tau})$ and specify the density of H_θ to be $\phi(\beta | \boldsymbol{\mu}, \boldsymbol{\tau})$. Let λ denote a positive scalar. We choose a Multivariate Normal-Inverse Wishart distribution for $\boldsymbol{\mu}, \boldsymbol{\tau}$, where specifically $\boldsymbol{\mu} | \boldsymbol{\tau}$ is a Multivariate Normal vector with mean parameter \mathbf{m} and scaled covariance matrix $\lambda^{-1} \boldsymbol{\tau}$. $\boldsymbol{\tau}$ is drawn from an Inverse-Wishart distribution with degrees of freedom ν_0 and scale matrix S_0 . Denote this distribution for $\boldsymbol{\mu}, \boldsymbol{\tau}$ as $G\text{-IW}(\mathbf{m}, \lambda^{-1} \boldsymbol{\tau}, S_0, \nu_0)$. Our specifications are quite similar to those used in Train (2003, Chapter 12) for a parametric GML model for panel data.

3.2 Blocked Gibbs algorithm

To approximate the posterior law, given \mathbf{Y} , for a function $g(\mathbf{B}, G, \theta)$, cycle through the following steps:

1. *Conditional draw for \mathbf{K} .* Independently sample K_i according to

$$\Pr\{K_i \in \cdot | \mathbf{p}, \mathbf{Z}, \mathbf{Y}\} = \sum_{k=1}^N p_{k,i} \delta_k(\cdot), \quad \text{for } i = 1, \dots, n,$$

where $(W_{1,i}, \dots, p_{N,i}) \propto (p_1 L_i(Y_i, Z_1), \dots, p_N L_i(Y_i, Z_N))$.

2. *Conditional draw for \mathbf{p} .* $p_1 = V_1^*$ and $p_k = (1 - V_1^*) \cdots (1 - V_{k-1}^*) V_k^*$, $k = 2, \dots, N - 1$ where

$$V_k^* \stackrel{\text{ind}}{\sim} \text{Beta} \left(1 + e_k, \alpha + \sum_{l=k+1}^N e_l \right), \quad k = 1, \dots, N - 1$$

and e_k records the number of K_i values which equal k .

3. *Conditional draw for \mathbf{Z} .* Let $\{K_1^*, \dots, K_{n_0}^*\}$ denote the unique set of K_i values. For each $k \notin \{K_1^*, \dots, K_{n_0}^*\}$ draw $Z_k | \boldsymbol{\mu}, \boldsymbol{\tau}$ from the prior Multivariate Normal density

$$H(dZ | \boldsymbol{\mu}, \boldsymbol{\tau}) := \phi(Z | \boldsymbol{\mu}, \boldsymbol{\tau}).$$

For $j = 1, \dots, n_0$, draw $Z_{K_j^*} := \beta_j^*$ from the density proportional to

$$\left[\prod_{\{i: K_i = K_j^*\}} \frac{\exp\{\mathbf{x}'_{iY_i} \beta_j^*\}}{\sum_{l_i \in \Phi} \exp\{\mathbf{x}'_{il_i} \beta_j^*\}} \right] \phi(\beta_j^* | \boldsymbol{\mu}, \boldsymbol{\tau})$$

For this step, draws are obtained by using a standard Metropolis-Hastings procedure.

4. *Conditional draw for $\theta = (\boldsymbol{\mu}, \boldsymbol{\tau})$.* Conditional on $\boldsymbol{\tau}, \mathbf{B}, \mathbf{Y}, \mathbf{K}$, draw $\boldsymbol{\mu}$ from a Multivariate Normal distribution with parameters

$$\frac{\lambda \mathbf{m} + n_0 \bar{\boldsymbol{\beta}}_{n_0}}{\lambda + n_0} \text{ and } \frac{\boldsymbol{\tau}}{\lambda + n_0}$$

where

$$\bar{\boldsymbol{\beta}}_{n_0} = \frac{1}{n_0} \sum_{j=1}^{n_0} \beta_j^* \quad (11)$$

Conditional on \mathbf{B}, \mathbf{Y} , draw $\boldsymbol{\tau}$ from an Inverse-Wishart distribution with parameters

$$\nu_0 + n_0 \text{ and } \frac{\nu_0 \mathbf{S}_0 + n_0 \mathbf{S}_{n_0} + R(\bar{\boldsymbol{\beta}}_{n_0}, \mathbf{m})}{\nu_0 + n_0}$$

where

$$\mathbf{S}_{n_0} = \frac{1}{n_0} \sum_{j=1}^{n_0} (\beta_j^* - \bar{\boldsymbol{\beta}}_{n_0}) (\beta_j^* - \bar{\boldsymbol{\beta}}_{n_0})' \text{ and } R(\bar{\boldsymbol{\beta}}_{n_0}, \mathbf{m}) = \frac{\lambda n_0}{\lambda + n_0} (\bar{\boldsymbol{\beta}}_{n_0} - \mathbf{m}) (\bar{\boldsymbol{\beta}}_{n_0} - \mathbf{m})'$$

Notice that Steps 3 and 4 are quite similar to the MCMC steps for a parametric Bayesian model. In fact when $n_0 = 1$, Steps 3 and 4 reduce to calculations for a parametric model. Iterating the steps above eventually produces a draw from the distribution $\mathbf{Z}, \mathbf{K}, \mathbf{p}, \theta | \mathbf{Y}$. Thus, each iteration, m , defines a random probability measure,

$$G_{(m)}(\cdot) = \sum_{k=1}^N W_{k,m} \delta_{Z_{k,m}}(\cdot)$$

which eventually approximates draws from the posterior distribution of $G | \mathbf{Y}$. Consequently one can approximate the posterior distributional properties of the choice probabilities $P(\{j\} | G, \mathbf{x})$ by constructing (iteratively) values ,

$$P(\{j\} | G_{(m)}, \mathbf{x}) = \sum_{k=1}^N W_{k,m} \frac{\exp\{\mathbf{x}'_j Z_{k,m}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l Z_{k,m}\}}$$

for $m = 1, \dots, M$ iterations. For instance, a histogram of the $P(\{j\}|G_{(m)}, \mathbf{x})$ for $m = 1, \dots, M$, approximates the posterior distribution. An approximation to the posterior mean

$$E[P(\{j\}|G, \mathbf{x})|\mathbf{Y}] = \int E[P(\{j\}|G, \mathbf{x})|\mathbf{B}, \theta]\pi(d\mathbf{B}, d\theta|\mathbf{Y}), \quad (12)$$

where $E[P(\{j\}|G, \mathbf{x})|\mathbf{B}, \theta]$ is given in (7), is obtained by

$$\frac{1}{M} \sum_{m=1}^M P^m(\{j\}), \text{ or alternatively by } \frac{1}{M} \sum_{m=1}^M E[P(\{j\}|G, \mathbf{x})|\mathbf{B}^m, \theta^m]. \quad (13)$$

While it seems reasonable to use the the posterior mean of $P(\{j\}|G, \mathbf{x})$, given in (13), for estimating the choice probabilities, we instead work with a plug-in type estimator defined as,

$$\frac{1}{M} \sum_{m=1}^M \frac{\exp\{\mathbf{x}'_j \bar{\boldsymbol{\beta}}_m\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}'_l \bar{\boldsymbol{\beta}}_m\}} \quad (14)$$

where

$$\bar{\boldsymbol{\beta}}_m = \frac{1}{n_0} \sum_{j=1}^{n_0} e_{K_j^*} \boldsymbol{\beta}_{j,m}^* = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\beta}_{i,m}$$

is the sample mean of $\mathbf{B} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n\}$ obtained from each iteration. We have found that in simulations the estimator (14) performs significantly better than posterior mean estimate.

4 Bayesian modelling for Panel data

The MMNL framework may also be used to model choice probabilities based on Panel data. In the Panel data setting, each individual i is observed to make a sequence of choices (or purchases) at different time points. The random utility for choosing j for each individual i now additionally depends on time, t , and is given by

$$U_{ijt} = \mathbf{x}'_{ijt} \boldsymbol{\beta}_i + \varepsilon_{ijt} \quad (15)$$

for times $t = 1, \dots, T_i$ and individuals $i = 1, \dots, n$. The MMNL model can be described as follows [see Train (2003, Section 6.7)], given $\boldsymbol{\beta}_i$, the probability that a person makes a sequence of choices $\mathbf{Y}_i = \{Y_{i1}, \dots, Y_{i,T_i}\}$ at times $t = 1, \dots, T_i$ is given by

$$L_i(\mathbf{Y}_i, \boldsymbol{\beta}_i) = \prod_{t=1}^{T_i} \frac{\exp\{\mathbf{x}'_{iY_{it}} \boldsymbol{\beta}_i\}}{\sum_{l_{it} \in \Phi} \exp\{\mathbf{x}'_{il_{it}} \boldsymbol{\beta}_i\}}$$

The MMNL model is completed by choosing $\boldsymbol{\beta}_i$ to be from a distribution F and the unconditional choice probability is specified by

$$P_i(\mathbf{Y}_i|F) = \int_{\mathcal{R}^d} L_i(\mathbf{Y}_i, \boldsymbol{\beta}_i) F(d\boldsymbol{\beta}_i).$$

Similar to the non-Panel data setting we wish to model F using a Bayesian framework. While it is possible to choose F to follow a Dirichlet process, this would result in possible ties among the individual preferences β_i . In order to preserve the distinct nature of each individuals preference, we assume that each $\beta_1, \dots, \beta_n | F$ are iid F where F is a mixture of multivariate normal distributions with unknown mixing distribution G . That is, F has density $f(\beta_i | G)$ given by

$$f(\beta_i | G) = \int_{\Theta} \phi(\beta_i | \mu_i, \tau_i) G(d\mu_i, d\tau_i) \quad (16)$$

where, $\phi(\beta_i | \mu_i, \tau_i)$ is a multivariate normal density with mean parameter μ_i and covariance matrix τ_i . The Bayesian MMNL is then specified by choosing G to be a Dirichlet process with shape $\alpha H(\cdot)$. Hence the Bayesian MMNL model for each individual i is expressible as

$$P_i(\mathbf{Y}_i | F) = \int_{\mathcal{R}^d} L(\mathbf{Y}_i, \beta_i) F(d\beta_i | G) = \sum_{k=1}^{\infty} p_k \left[\int_{\mathcal{R}^d} L_i(\mathbf{Y}_i, \beta_i) \phi(\beta_i | Z_k) d\beta_i \right]$$

While one may use any choice for H , we choose $H(d\mu, d\tau)$ to be the Multivariate Normal-Inverse-Wishart distribution $G\text{-IW}(\mathbf{m}, \lambda^{-1}\tau, \mathbf{S}_0, \nu_0)$ described in section 5.

4.1 Blocked Gibbs algorithm for Panel data

The explicit posterior analysis for the Panel data case is quite similar to the non-Panel case. The main difference is that the (μ_i, τ_i) for $i = 1, \dots, n$ rather than β_1, \dots, β_n are drawn from the Dirichlet process. Here we will briefly focus on the relevant data structure and then proceed to a description of how to implement the blocked Gibbs procedure. One will note many similarities to the analogous parametric MCMC procedure for GML Panel data models in Train (2003, Section 12.6).

The joint distribution of the augmented data can be expressed using a hierarchical model as follows:

$$\begin{aligned} \mathbf{Y}_i | \beta_i &\stackrel{\text{ind}}{\sim} L_i(\mathbf{Y}_i, \beta_i) = \prod_{t=1}^{T_i} \frac{\exp\{\mathbf{x}'_{iY_{it}} \beta_i\}}{\sum_{l_{it} \in \Phi} \exp\{\mathbf{x}'_{il_{it}} \beta_i\}}, \text{ for } i=1, \dots, n, \text{ and } Y_{it} \in \Phi, \\ \beta_i | \mu_i, \tau_i &\stackrel{\text{ind}}{\sim} \phi(\beta_i | \mu_i, \tau_i), \text{ for } i=1, \dots, n \\ \mu_i, \tau_i | G &\stackrel{\text{iid}}{\sim} G, \text{ for } i=1, \dots, n \\ G &\sim \mathcal{P}(dG | \alpha H) \end{aligned} \quad (17)$$

Similar to the non-Panel case, the blocked Gibbs sampler works by using the $\mathcal{P}_N(dG | \alpha H)$ law in place of the Dirichlet process. We now sample $(\mathbf{B}, \mathbf{Z}, \mathbf{p}, \mathbf{K})$ from the distribution proportional to

$$\left[\prod_{i=1}^n L_i(\mathbf{Y}_i, \beta_i) \phi(\beta_i | \mu_i, \tau_i) \right] \left[\prod_{i=1}^n \sum_{k=1}^N W_k \delta_{Z_k}(d\mu_i, d\tau_i) \right] \pi(\mathbf{p}) \prod_{k=1}^N H(dZ_k).$$

Here we use the fact that $(\mu_i, \tau_i) = Z_{K_i}$, for $i = 1, \dots, n$.

To approximate the posterior law of various functionals cycle through the following steps:

1. *Conditional draw for $\mathbf{K}|\mathbf{p}, \mathbf{B}, \mathbf{Z}, \mathbf{Y}$.* Independently sample K_i according to

$$\Pr\{K_i \in \cdot | \mathbf{p}, \mathbf{Z}, \mathbf{B}, \mathbf{X}\} = \sum_{k=1}^N p_{k,i} \delta_k(\cdot), \quad \text{for } i = 1, \dots, n,$$

where $(p_{1,i}, \dots, p_{N,i}) \propto (p_1 \phi(\boldsymbol{\beta}_i | Z_1), \dots, p_N \phi(\boldsymbol{\beta}_i | Z_N))$.

2. *Conditional draw for \mathbf{p} .* $p_1 = V_1^*$ and $p_k = (1 - V_1^*) \cdots (1 - V_{k-1}^*) V_k^*$, $k = 2, \dots, N - 1$ where

$$V_k^* \stackrel{\text{ind}}{\sim} \text{Beta} \left(1 + e_k, \alpha + \sum_{l=k+1}^N e_l \right), \quad k = 1, \dots, N - 1$$

and e_k records the number of K_i values which equal k .

3. *Conditional draw for $\mathbf{Z}|\mathbf{B}, \mathbf{Y}$.* Let $\{K_1^*, \dots, K_{n_0}^*\}$ denote the unique set of K_i values. For each $k \notin \{K_1^*, \dots, K_{n_0}^*\}$ draw $Z_k = (\boldsymbol{\mu}_k, \boldsymbol{\tau}_k)$ from the prior, $\text{G-IW}(\mathbf{m}, \lambda^{-1} \boldsymbol{\tau}, \mathbf{S}_0, \nu_0)$.

Draw $Z_{K_j^*} = (\boldsymbol{\mu}_j^*, \boldsymbol{\tau}_j^*)$, for $j = 1, \dots, n_0$, as follows:

Conditional on $\boldsymbol{\tau}_j^*, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n, \mathbf{Y}, K$, draw $\boldsymbol{\mu}_j^*$ from a Multivariate Normal distribution with parameters

$$\frac{\lambda \mathbf{m} + e_{K_j^*} \bar{\boldsymbol{\beta}}_j^*}{\lambda + e_{K_j^*}} \quad \text{and} \quad \frac{\boldsymbol{\tau}_j^*}{\lambda + e_{K_j^*}}$$

where

$$\bar{\boldsymbol{\beta}}_j^* = \frac{1}{e_{K_j^*}} \sum_{\{i: K_i = K_j^*\}} \boldsymbol{\beta}_i$$

Conditional on $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n, \mathbf{Y}$, draw $\boldsymbol{\tau}_j^*$ from an Inverse-Wishart distribution with parameters

$$\nu_0 + e_{K_j^*} \quad \text{and} \quad \frac{\nu_0 \mathbf{S}_0 + e_{K_j^*} \mathbf{S}_j + R(\bar{\boldsymbol{\beta}}_j^*, \mathbf{m})}{\nu_0 + e_{K_j^*}}$$

where

$$\mathbf{S}_j = \frac{1}{e_{K_j^*}} \sum_{\{i: K_i = K_j^*\}} (\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_j^*) (\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_j^*)' \quad \text{and} \quad R(\bar{\boldsymbol{\beta}}_j^*, \mathbf{m}) = \frac{\lambda e_{K_j^*}}{\lambda + e_{K_j^*}} (\bar{\boldsymbol{\beta}}_j^* - \mathbf{m}) (\bar{\boldsymbol{\beta}}_j^* - \mathbf{m})'$$

4. *Conditional draw for $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n | \mathbf{Z}, \mathbf{Y}, \mathbf{K}$.* For each fixed $j = 1, \dots, n_0$, draw independently $\boldsymbol{\beta}_i$ for $i \in \{l : K_l = K_j^*\}$ from the density proportional to

$$\left[\prod_{t=1}^{T_i} \frac{\exp \{ \mathbf{x}'_{iY_{it}} \boldsymbol{\beta}_i \}}{\sum_{l_{it} \in \Phi} \exp \{ \mathbf{x}'_{il_{it}} \boldsymbol{\beta}_i \}} \right] \phi(\boldsymbol{\beta}_i | \boldsymbol{\mu}_j^*, \boldsymbol{\tau}_j^*)$$

Draws are obtained by using a standard Metropolis-Hastings procedure.

When $n_0 = 1$, Steps 3 and 4 equate with a parametric MCMC procedure for Panel data models quite similar to the algorithm described in Train (2003, section 12.6).

4.2 Extensions

The similarities between the nonparametric and parametric MCMC procedures suggest that parametric MCMC methods to account for more general MMNL models can be easily adapted to the present setting. For instance, it is straightforward to adapt Train's (2003, 12.7.3) description on how to estimate models with additional fixed coefficients. That is suppose that each individual i has utility,

$$U_{ijt} = \mathbf{r}'_{ijt}\Omega + \mathbf{x}'_{ijt}\boldsymbol{\beta}_i + \varepsilon_{ijt}, \quad (18)$$

with common parameter Ω and additional covariates \mathbf{r}_{ijt} . Placing a prior on Ω , say $\pi(d\Omega)$ the blocked gibbs procedure proceeds by replacing Step 4 with a draw of \mathbf{B} given $\mathbf{Z}, \mathbf{Y}, \mathbf{K}, \Omega$ described as: For each fixed $j = 1, \dots, n_0$, draw independently $\boldsymbol{\beta}_i$ for $i \in \{l : K_l = K_j^*\}$ from the density proportional to

$$\left[\prod_{t=1}^{T_i} \frac{\exp \left\{ \mathbf{r}'_{iY_{it}} \Omega + \mathbf{x}'_{iY_{it}} \boldsymbol{\beta}_i \right\}}{\sum_{l_{it} \in \Phi} \exp \left\{ \mathbf{r}'_{iY_{it}} \Omega + \mathbf{x}'_{l_{it}} \boldsymbol{\beta}_i \right\}} \right] \phi(\boldsymbol{\beta}_i | \boldsymbol{\mu}_j^*, \boldsymbol{\tau}_j^*).$$

A Step 5 is added to obtain a draw of Ω given $\mathbf{B}, \mathbf{Z}, \mathbf{Y}, \mathbf{K}$ proportional to

$$\left[\prod_{i=1}^n \prod_{t=1}^{T_i} \frac{\exp \left\{ \mathbf{r}'_{iY_{it}} \Omega + \mathbf{x}'_{iY_{it}} \boldsymbol{\beta}_i \right\}}{\sum_{l_{it} \in \Phi} \exp \left\{ \mathbf{r}'_{iY_{it}} \Omega + \mathbf{x}'_{l_{it}} \boldsymbol{\beta}_i \right\}} \right] \pi(d\Omega).$$

5 Simulation and Real Data

5.1 Simulation Study

In this section we present some empirical evidence that shows how the MMNL procedures perform overall, and relative to the parametric Normal mixed logit (GML) models, in recovering the choice probabilities based on simulated data. Four different artificial datasets are generated for the simulation study. First two datasets (dataset 1 and dataset 2) are produced for studying non-Panel data model. Two datasets (dataset 3 and dataset 4) are also designed in order to study models with Panel data. We first briefly describe the relevant models and how data was generated from them. All models are based on the following random utility model with three possible responses relative to the utilities U_1, U_2 and U_3 ,

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} = \begin{bmatrix} x_{11}\beta_1 + x_{21}\beta_2 \\ x_{12}\beta_1 + x_{22}\beta_2 \\ x_{13}\beta_1 + x_{23}\beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \quad (19)$$

For the first artificial dataset (dataset 1), we choose $\varepsilon_1, \varepsilon_2, \varepsilon_3 \stackrel{\text{iid}}{\sim}$ Standard Gumbel, $\beta_1 \sim \text{Uniform}(0, 2)$, $\beta_2 \sim \text{Uniform}(0, 2)$, with β_1 and β_2 independent. For each individual i we randomly generate (componentwise) their covariates $\mathbf{x}_i = (x_{i11}, x_{i12}, x_{i13}, x_{i21}, x_{i22}, x_{i23})$ independently from a Uniform $(-1, 1)$ distribution. Set $Y_i = j$ if $U_i = \max \{U_1, U_2, U_3\}$ for $j = 1, 2, 3$. Repeat this procedure n times independently to get a dataset with (Y_i, \mathbf{x}_i) for $i = 1, \dots, n$. The second artificial dataset (dataset 2) is the modified version of the first artificial dataset (dataset 1). The only change is

assuming $\varepsilon_1, \varepsilon_2, \varepsilon_3 \stackrel{\text{iid}}{\sim} \text{Uniform}(-5, 5)$. Set $n = 1000$ for both dataset 1 and dataset 2. For panel data, we assume there are $n = 100$ individuals each making $T_i = 10$ choices. The artificial dataset (dataset 3) consists of panel data drawn from the model corresponding otherwise to dataset 1. Similarly dataset 4 consists of panel data drawn relative to the model used to generate dataset 2. We apply our procedures to choice probabilities $P(\{j\}|G, \mathbf{x})$ for $j = 1, 2, 3$ based on two different sets of covariates, take $\mathbf{x} = (x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}) = (1.0, 1.0, 1.0, -0.5, 0.2, 0.5)$ and $\mathbf{x} = (1.0, 1.0, 1.0, -0.9, 0.0, 0.9)$. The prior parameters for the specifications of the Bayesian MMNL models for panel and non-panel data (pertaining to the explicit models in 5.1 and 6.2) are set to be $\alpha = 2.5$, $\nu_0 = 10$, $\mathbf{m} = (0, \dots, 0)'$ and $t = 0.001$. Additionally we use $N = 50$ and discuss results for two choices of the scale matrix, $\mathbf{S}_0 = 0.01\mathbf{I}$ and $\mathbf{S}_0 = 0.001\mathbf{I}$ where \mathbf{I} is denoted as identity matrix. A parametric GML model is also estimated for comparison with the same specifications for $\nu_0 = 10$, \mathbf{m} , λ and \mathbf{S}_0 . In all cases we use the estimator in (14) based on an initial burn-in of 10000 cycles an additional 10000 Gibbs cycles ($M = 10000$) are then repeated for the estimation. Simulation results using datasets 1, 2, 3, 4 are summarized respectively in Tables 1, 2, 3, 4.

Table 1 and Table 2 show that the performance of the nonparametric MMNL estimator is better than that of the parametric GML estimators. The results for the nonparametric MMNL are better in the case where $\mathbf{S}_0 = 0.01\mathbf{I}$. Table 3 and Table 4 show more dramatically that the nonparametric MMNL for Panel data performs better than the parametric GML model. Notice that the nonparametric estimator varies according to the choice of \mathbf{S}_0 . Although we have not done so, it seems produce to place an additional prior on \mathbf{S}_0 . This extra layer of complexity is easily incorporated into the blocked Gibbs procedures.

5.2 Travel Mode Choice Data

The data set [Greene (2003, page 729)] we considered in this study contains 210 observations on choice among four travel modes, which are *air*, *train*, *bus* and *car*, for travel between Sydney and Melbourne, Australia. The covariates used here are Terminal waiting time (*Ttme*), In vehicle cost (*Invc*), Travel time in vehicle (*Invt*), Generalized cost measure (*GC*) and household income (*Hinc*). The utility function of each choice is written as

$$U_{ij} = Ttime_{ij}\beta_1 + Invc_{ij}\beta_2 + Invt_{ij}\beta_3 + GC_{ij}\beta_4 + Hinc_{ij}\beta_5 + \varepsilon_{ij} \text{ for } j \in \{air, train, bus, car\}.$$

Estimates of the choice probabilities are shown in Table 5. Here non-parametric non-panel data model is implemented. The parameters setting of the prior here are the same as those of simulation datasets in the last section. 20000 iterations are run and the draws in last 10000 iterations are used to calculate those choice probabilities. The given independent variables for calculating the choice probabilities are selected from the original dataset, which is the last observation (independent variables only). The result tells us that the choice probability of *car* is highest among that of other choices. It is consistent with the observation (the choice).

INSERT TABLES 1-5 here consecutively

References

- [1] Antoniak, C. E., 1974 Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2 1152-1174.
- [2] Blackwell, D. and J. B. MacQueen, 1973, Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1, 353-355.
- [3] Ben-Akiva M., D. Bolduc, and J. Walker, 2001, Specification, identification, & estimation of the logit kernel (or continuous mixed logit) model. Working Paper, Massachusetts Institute of Technology, Cambridge.
- [4] Bhat, C., 1998, Accommodating variations in responsiveness to level-of-service variables in travel mode choice models. *Transportation Research A*, 32, 455-507.
- [5] Brownstone, D. and K. Train, 1999, Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics*, 89, 109-129.
- [6] Cardell, N. and F. Dunbar, 1980, Measuring the Societal Impacts of Automobile Downsizing. *Transportation Research A*, 14, 423-434.
- [7] Dubé, J. P., Chintagunta, P., Bronnenburg, B., Goettler, R., Petrin, A., Sudhir, K., Zhao Y., 2002, Structural Applications of the Discrete Choice Model. *Marketing Letters*, 13, 207-220.
- [8] Erdem, T., 1996, A dynamic analysis of market structure based on panel data. *Marketing Science*, 15, 359-378.
- [9] Escobar, M. D., 1994, Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.
- [10] Escobar, M. D. and M. West, 1995, Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.
- [11] Ferguson, T. S., 1973, A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.
- [12] Greene, W., 2003, *Econometric Analysis*, Prentice Hall, 5th Edition.
- [13] Ishwaran, H. and L. F. James, 2001, Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96, 161-173.
- [14] Ishwaran, H and M. Zarepour, 2000, Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 371-390.
- [15] Ishwaran, H and M. Zarepour, 2002, Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12, 941-963.
- [16] Kamakura, W. A., and G. J. Russell, 1989, A probabilistic choice for market segmentation and elasticity structure. *Journal of Marketing Research*, 26, 379-390.

- [17] Lo, A. Y., 1984, On a class of Bayesian Nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12, 351-257.
- [18] Luce, R., 1959, *Individual Choice Behavior: A Theoretical Analysis*. New York, John Wiley & Sons.
- [19] MacEachern, S. N., 1998, Computational methods for mixtures of Dirichlet processes. In D. Dey, P. Mueller and D. Sinha, eds. *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer-Verlag, New York.
- [20] McFadden, D., 1974, Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed. *Frontiers of Econometrics*, Academic Press, New York, 105-142.
- [21] McFadden, D. and Train, K., 2000, Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15, 447-470.
- [22] Srinivasan, K. and Mahmassani, 2000, Dynamic kernel logit model for the analysis of longitude discrete choice data: Properties and computational assessment. Working paper, Department of Civil Engineering, University of Texas, Austin.
- [23] Walker J., 2001, *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures and Latent Variables*. Ph.D. Dissertation in Transportation Systems, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

| | | Parametric GML | | Nonparametric MMNL | |
|-------------------------|--------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|
| | | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_1)$ | 0.1846 | 0.2176 | 0.2176 | 0.1904 | 0.1878 |
| $P(\{2\} \mathbf{x}_1)$ | 0.3443 | 0.3510 | 0.3511 | 0.3504 | 0.3505 |
| $P(\{3\} \mathbf{x}_1)$ | 0.4712 | 0.4312 | 0.4313 | 0.4591 | 0.4617 |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_2)$ | 0.1302 | 0.1597 | 0.1597 | 0.1265 | 0.1232 |
| $P(\{2\} \mathbf{x}_2)$ | 0.2538 | 0.2945 | 0.2945 | 0.2703 | 0.2681 |
| $P(\{3\} \mathbf{x}_2)$ | 0.6160 | 0.5458 | 0.5458 | 0.6031 | 0.6087 |

Table 1: Simulation results for artificial dataset 1 with $n = 1000$ where $\mathbf{x}_1 = (1.0, 1.0, 1.0, -0.5, 0.2, 0.5)$ and $\mathbf{x}_2 = (1.0, 1.0, 1.0, -0.9, 0.0, 0.9)$

| | | Parametric GML | | Nonparametric MMNL | |
|-------------------------|--------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|
| | | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_1)$ | 0.2569 | 0.2808 | 0.2808 | 0.2590 | 0.2626 |
| $P(\{2\} \mathbf{x}_1)$ | 0.3482 | 0.3439 | 0.3439 | 0.3463 | 0.3460 |
| $P(\{3\} \mathbf{x}_1)$ | 0.3950 | 0.3753 | 0.3753 | 0.3947 | 0.3915 |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_2)$ | 0.2189 | 0.2512 | 0.2512 | 0.2197 | 0.2247 |
| $P(\{2\} \mathbf{x}_2)$ | 0.3196 | 0.3254 | 0.3254 | 0.3147 | 0.3167 |
| $P(\{3\} \mathbf{x}_2)$ | 0.4616 | 0.4233 | 0.4233 | 0.4656 | 0.4586 |

Table 2: Simulation results for artificial dataset 2 with $n = 1000$ where $\mathbf{x}_1 = (1.0, 1.0, 1.0, -0.5, 0.2, 0.5)$ and $\mathbf{x}_2 = (1.0, 1.0, 1.0, -0.9, 0.0, 0.9)$

| | | Parametric GML | | Nonparametric MMNL | |
|-------------------------|--------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|
| | | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_1)$ | 0.1846 | 0.2089 | 0.2089 | 0.1784 | 0.1881 |
| $P(\{2\} \mathbf{x}_1)$ | 0.3443 | 0.3515 | 0.3515 | 0.3450 | 0.3502 |
| $P(\{3\} \mathbf{x}_1)$ | 0.4712 | 0.4397 | 0.4397 | 0.4766 | 0.4617 |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_2)$ | 0.1302 | 0.1480 | 0.1480 | 0.1221 | 0.1327 |
| $P(\{2\} \mathbf{x}_2)$ | 0.2538 | 0.2882 | 0.2882 | 0.2504 | 0.2432 |
| $P(\{3\} \mathbf{x}_2)$ | 0.6160 | 0.5639 | 0.5639 | 0.6275 | 0.6241 |

Table 3: Simulation results for dataset 3 with $n = 100$ and $T = 10$ where $\mathbf{x}_1 = (1.0, 1.0, 1.0, -0.5, 0.2, 0.5)$ and $\mathbf{x}_2 = (1.0, 1.0, 1.0, -0.9, 0.0, 0.9)$

| | | Parametric GML | | Nonparametric MMNL | |
|-------------------------|--------|---------------------------------|----------------------------------|---------------------------------|----------------------------------|
| | | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_1)$ | 0.2569 | 0.2208 | 0.2208 | 0.2492 | 0.2398 |
| $P(\{2\} \mathbf{x}_1)$ | 0.3482 | 0.3499 | 0.3499 | 0.3481 | 0.3489 |
| $P(\{3\} \mathbf{x}_1)$ | 0.3950 | 0.4292 | 0.4292 | 0.4027 | 0.4111 |
| | True | Estimates | Estimates | Estimates | Estimates |
| $P(\{1\} \mathbf{x}_2)$ | 0.2189 | 0.1781 | 0.1781 | 0.2044 | 0.1912 |
| $P(\{2\} \mathbf{x}_2)$ | 0.3196 | 0.3023 | 0.3023 | 0.3122 | 0.3070 |
| $P(\{3\} \mathbf{x}_2)$ | 0.4616 | 0.5195 | 0.5195 | 0.4834 | 0.5018 |

Table 4: Simulation results for dataset 4 with $n = 100$ and $T = 10$ where $\mathbf{x}_1 = (1.0, 1.0, 1.0, -0.5, 0.2, 0.5)$ and $\mathbf{x}_2 = (1.0, 1.0, 1.0, -0.9, 0.0, 0.9)$

| Estimates | $\mathbf{S}_0 = 0.01\mathbf{I}$ | $\mathbf{S}_0 = 0.001\mathbf{I}$ |
|---------------------------|---------------------------------|----------------------------------|
| $P(\{Air\} \mathbf{x})$ | 0.4665 | 0.3769 |
| $P(\{Train\} \mathbf{x})$ | 0.0007 | 0.0206 |
| $P(\{Bus\} \mathbf{x})$ | 0.0046 | 0.0090 |
| $P(\{Car\} \mathbf{x})$ | 0.5283 | 0.5934 |

Table 5: Simulation results for Train Mode Choice Data with $n = 210$

where $\mathbf{x} = (\mathbf{x}_{Air}, \mathbf{x}_{Train}, \mathbf{x}_{Bus}, \mathbf{x}_{Car})$,

$\mathbf{x}_{Air} = (Ttme, Invc, Invt, GC, Hinc)_{Air} = (64, 66, 140, 87, 70)$,

$\mathbf{x}_{Train} = (Ttme, Invc, Invt, GC, Hinc)_{Train} = (44, 54, 670, 156, 70)$,

$\mathbf{x}_{Bus} = (Ttme, Invc, Invt, GC, Hinc)_{Bus} = (53, 33, 664, 134, 70)$ and

$\mathbf{x}_{Car} = (Ttme, Invc, Invt, GC, Hinc)_{Car} = (0, 12, 540, 94, 70)$