INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

Pierpaolo De Blasi, Lancelot F. James and John W. Lau

BAYESIAN NONPARAMETRIC ESTIMATION AND CONSISTENCY OF MIXED MULTINOMIAL LOGIT CHOICE MODELS

Working Paper no. 15/2007
April 2007

APPLIED MATHEMATICS AND QUANTITATIVE METHODS
WORKING PAPER SERIES

Bayesian Nonparametric Estimation and Consistency of Mixed Multinomial Logit Choice Models

Pierpaolo De Blasi* — University of Turin

Lancelot F. James — Hong Kong University of Science and Technology

John W. Lau — University of the Witwatersrand

New version: April 2008

Abstract

This paper develops nonparametric estimation for discrete choice models based on the Mixed Multinomial Logit (MMNL) model. It has been shown that MMNL models encompass all discrete choice models derived under the assumption of random utility maximization, subject to the identification of an unknown distribution G. Noting the mixture model description of the MMNL, we employ a Bayesian nonparametric approach, using nonparametric priors on the unknown mixing distribution G, to estimate the unknown choice probabilities. Theoretical support for the use of the proposed methodology is provided by establishing strong consistency of a general nonparametric prior on G under simple sufficient conditions. Consistency is defined according to a $\mathcal{L}_1$-type distance on the space of choice probabilities and is achieved by extending to a regression model framework a recent approach to strong consistency based on the summability of square roots of prior probabilities. Moving to estimation, slightly different techniques for non-panel and panel data models are discussed. For practical implementation, we describe efficient and relatively easy to use blocked Gibbs sampling procedures. A simulation study is also performed to illustrate the proposed methods and the flexibility they achieve with respect to parametric Gaussian MMNL models.

Keyword: Bayesian consistency, Bayesian nonparametrics, Blocked Gibbs sampler, Discrete choice models, Mixed Multinomial Logit, Random probability measures, Stick-breaking priors.

*Address for correspondence: pierpaolo.deblasi@unito.it

1 Introduction

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 models (RUMs). Given a finite set of choices $\Phi = \{1, \ldots, J\}$, it is assumed that each individual has a utility function

$$U_j = \mathbf{x}_j'\boldsymbol{\beta} + \varepsilon_j, \quad \text{for } j \in \Phi.$$

The values $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$ are observed covariates, where $\mathbf{x}_j \in \mathcal{R}^d$ denotes the covariates associated with each choice $\{j\} \in \Phi$, the coefficient $\boldsymbol{\beta}$ is an unknown (preference) vector in $\mathcal{R}^d$, and $(\varepsilon_1, \ldots, \varepsilon_J)$ are random terms. Suppose that all U_j are distinct and that individual makes a choice $\{j\}$ if and only if $U_j > U_l \ \forall l \neq j$. The introduction of the random error terms ε_j's represents the departure from classical economic utility models. 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 $\mathbf{x}$ and the choice made by the individual. Hence, the deterministic statement of choice $\{j\}$ is replaced by the probability of choosing $\{j\}$, that is $\mathrm{P}\{U_j > U_l \ \forall l \neq j\}$. The analysis of such a model depends on the specifications of the errors. McFadden (1974) shows that the specification of independent Gumbel error terms leads to the tractable Multinomial Logit (MNL) Model. This representation is written as

$$\mathrm{P}(\{j\} \mid \boldsymbol{\beta}, \mathbf{x}) = \frac{\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}{\sum_{l\in\Phi} \exp\{\mathbf{x}_l'\boldsymbol{\beta}\}} \qquad \text{for } j \in \Phi.$$

The MNL possesses the property of independence from irrelevant alternatives (IIA), which makes it inappropriate in many situations. The probit and the generalized extreme value models have been proposed as alternatives to the MNL, which do not exhibit the IIA property and are models derived from dependent error structures. 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 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. For a set of covariates $\mathbf{x}$, the MMNL model is written as

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

McFadden and Train (2000) establish the important result that in theory all RUMs can be captured by correct specification of G. Thus, a robust approach 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 model is when G is specified to be multivariate normal with unknown mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\tau}$:

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

where $\phi(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\tau})$ represents a multivariate normal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$. We shall term this model a Gaussian Mixed Logit (GML) model. Here, based on a sample of size n, one estimates the choice probabilities by estimating $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$. Applications and discussions are, among others, in Bhat (1998), Brownstone and Train (1999), Erdem (1996), Srinivasan and Mahmassani (2005) and Walker, Ben-Akiva and Bolduc (2007). Additionally, Dubé *et al.* (2002) provide a discussion focused on applications to marketing. The GML model is 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. However, despite the attractive features of the GML, it does not encompass all RUMs, hence, it is not robust against miss-specification.

In this article, we develop a nonparametric Bayesian method for the estimation of the choice probabilities and we prove consistency for the posterior distribution. The idea is to model the mixing distribution G via a random probability measure in order to fully exploit the flexibility of the MMNL model. Many nonparametric priors are nowadays available for modeling G, like stick-breaking priors, normalized random measures with independent increments and Dirichlet process mixtures. We establish consistency of the posterior distribution of G under neat sufficient conditions which are readily verifiable for all these nonparametric priors. Consistency is defined according to a $\mathcal{L}_1$-type distance on the space of choice probabilities by exploiting the square root approach to strong consistency of Walker (2003a, 2004). We essentially show that the MMNL model is consistent if the prior on G has the true mixing distribution in its weak support and satisfies a mild condition on the tails of the prior predictive distribution. Then, we move to estimation and divide our discussion into methods for non-panel and panel data. Specifically, for non-panel data models we use, as a prior for G, a mixture of Dirichlet processes. Methods for panel data involve instead a Dirichlet mixture of normal densities. For practical implementation, we describe efficient and relatively easy to use blocked Gibbs sampling procedures, developed in Ishwaran and Zarepour (2000) and Ishwaran and James (2001).

The rest of the paper is organized as follows. In Section 2 we describe the Bayesian nonparametric approach by placing a nonparametric prior on the mixing distribution and present the consistency result for the posterior distribution of G. In Section 3 we show how to implement a blocked Gibbs sampling for drawing inference when a discrete nonparametric prior is used. Section 4 deals with panel data with similar Bayesian nonparametric methods, where we define a class of priors for G that preserves the distinct nature of individual preferences and specialize the blocked Gibbs sampler to this setting. In Section 5 we provide an illustrative simulation study which shows the flexibility and good performance of our procedure versus the parametric GML model. Finally, in Section 6 we give a detailed proof of consistency.

2 Bayesian MMNL models

A Bayesian nonparametric model for the MMNL is specified by placing a nonparametric prior on the mixing distribution G in (1):

$$\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}_j' \boldsymbol{\beta}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}_l' \boldsymbol{\beta}\}} \, \tilde{G}(\mathrm{d}\boldsymbol{\beta}) \qquad \text{for } j \in \Phi. \tag{3}$$

Here $\tilde{G}$ denotes a random probability measure which takes values over the space $\mathbb{P}$ of probability measures on $\mathcal{R}^d$, the former endowed with the weak topology. The nonparametric distribution of $\tilde{G}$ is denoted by $\mathcal{P}$. Model (3) can be equivalently expressed in hierarchical form as

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}_i \ &\overset{\text{ind}}{\sim}\ \frac{\exp\{\mathbf{x}_{iY_i}' \boldsymbol{\beta}_i\}}{\sum_{j \in \Phi} \exp\{\mathbf{x}_{ij}' \boldsymbol{\beta}_i\}}, \qquad \text{for } i{=}1, \ldots, n, \ \text{ and } Y_i{\in}\Phi, \\ \boldsymbol{\beta}_i \mid \tilde{G} \ &\overset{\text{iid}}{\sim}\ \tilde{G}, \quad \text{for } i{=}1, \ldots, n, \\ \tilde{G} \ &\sim\ \mathcal{P} \end{aligned} \tag{4}$$

with $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})$ the covariates and Y_i the choice observed for individual i.

One can choose $\tilde{G}$ to be a Dirichlet process (Ferguson 1973), although there exist nowadays other nonparametric priors that can be used, like stick-breaking processes (Ishwaran and James 2001) and normalized random measure with independent increments (NRMI) (Regazzini, Lijoi and Prünster 2003). All these priors select discrete distribution almost surely (a.s.), whereas random probability measures whose support contains continuous distributions can be obtained by using a Dirichlet process mixture of densities in the spirit of Lo (1984). An important role in the sequel will be played by the prior predictive distribution of $\tilde{G}$, say H, which is an element of $\mathbb{P}$ and is defined by

$$H(B) := \mathrm{E}[\tilde{G}(B)] = \mathrm{P}\{\boldsymbol{\beta} \in B\}, \tag{5}$$

for all Borel set B of $\mathcal{R}^d$, where $\mathrm{E}(\cdot)$ denotes expectation. In the next section we show that an essential condition for consistency of the posterior distribution is expressed in terms of H. This yields an easy to use criterion for the choice of the prior for $\tilde{G}$ as H is readily obtained for all the nonparametric priors listed above. Furthermore, one can embed a parametric model, such as the GML, within the nonparametric framework via a suitable specification of the distribution H.

2.1 Posterior consistency

Bayesian consistency deals with the asymptotic behavior of posterior distributions with respect to repeated sampling. The problem can be set in general terms as follows: suppose the existence of a "true" unknown distribution P_0 that generates the data, then check whether the posterior accumulates in suitably-defined neighborhoods of P_0. There exist two main approaches to the study of strong consistency, that is consistency when the neighborhood of P_0 is defined accordingly to the Hellinger metric on the space of density functions. One is based on the metric entropy of the parameter space and was set forth in Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorthi (1999). The second approach was introduced by Walker (2003a, 2004) and has a more Bayesian flavor in the sense that it relies on the summability of square roots of prior probabilities. For discussion the readers is referred to Wasserman (1998), Walker, Lijoi and Prünster (2005) and Choudhuri, Ghosal and Roy (2005). Strong consistency in mixture models for density estimation is addressed by Ghosal, Ghosh and Ramamoorthi (1999) and Lijoi, Prünster and Walker (2005) by using the metric entropy approach and the square root approach, respectively. As for the non-identically distributed case, we mention Choi and Schervish (2007), Ghosal and Roy (2006). All these papers follow the metric entropy approach. The square root approach is adopted by Walker (2003b) for nonparametric regression models and by Ghosal and Tang (2006) for estimating transition densities in the context of Markov processes.

We face the issue of consistency for the MMNL model (3) by exploiting the square root approach of Walker and its variation proposed in Lijoi, Prünster and Walker (2005) which makes use of metric entropy in an instrumental way. We assume the existence of a $G_0 \in \mathbb{P}$ such that the true distribution of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$P_0(\{j\} \mid \mathbf{x}) = \int_{\mathcal{R}^d} \frac{\exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{l \in \Phi} \exp(\mathbf{x}_l' \boldsymbol{\beta})} \, G_0(\mathrm{d}\boldsymbol{\beta}).$$

The variables $\mathbf{X}_i$'s are taken as independent draws from a common distribution $M(\mathrm{d}\mathbf{x})$ which is supported on $\mathcal{X} \subset \mathcal{R}^{Jd}$. The distribution of an infinite sequence $(Y_i, \mathbf{X}_i)_{i \geq 1}$ will be then denoted by

$\mathrm{P}^{\infty}_{(G_0,M)}$. The posterior distribution of $\tilde{G}$ is given by

$$\mathcal{P}_n(A) = \mathrm{P}\{\tilde{G} \in A \mid (Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)\}$$

for any measurable set A of $\mathbb{P}$. We give conditions on G_0 and the prior predictive distribution of $\mathcal{P}$ such that the posterior distribution $\mathcal{P}_n$ concentrates all probability mass in neighborhoods of G_0 defined according to strong consistency of choice probabilities. To this aim, we look at the vector of choice probabilities as a vector-valued function $\mathbf{q} : \mathcal{X} \to \Delta$, where Δ is the J-dimensional probability simplex. We define

$$\mathbf{q}(\mathbf{x}; G) = \big[\mathrm{P}(\{1\} \mid G, \mathbf{x}), \ldots, \mathrm{P}(\{J\} \mid G, \mathbf{x})\big], \tag{6}$$

for any $G \in \mathbb{P}$. On the space $\mathscr{Q} = \{\mathbf{q}(\mathbf{x}; G) : \; G \in \mathbb{P}\}$ we define the $\mathcal{L}_1$-type distance

$$d(\mathbf{q}_1, \mathbf{q}_2) = \int_{\mathcal{X}} |\mathbf{q}_1(\mathbf{x}) - \mathbf{q}_2(\mathbf{x})| M(\mathrm{d}\mathbf{x}), \tag{7}$$

where $|\cdot|$ stays for the Euclidean norm in Δ.

Definition 1 . *$\mathcal{P}$ is consistent at G_0 if, for any $\epsilon > 0$,*

$$\mathcal{P}_n\{G : \; d\big(\mathbf{q}(\cdot\,; G), \mathbf{q}(\cdot\,; G_0)\big) > \epsilon\} \to 0 \qquad a.s.\text{–}\, \mathrm{P}^{\infty}_{(G_0,M)}.$$

The main result is stated in the following theorem.

Theorem 1 . *Let $\mathcal{P}$ be a prior on $\mathbb{P}$ with predictive distribution H and G_0 be in the weak support of $\mathcal{P}$. Suppose that $\mathcal{X}$ is a compact subset of $\mathcal{R}^{Jd}$. If*

(i) $P_0(\{j\} \mid \mathbf{x}) > 0$ for any $j \in \Phi$ and $\mathbf{x} \in \mathcal{X}$,

(ii) $\int_{\mathcal{R}^d} |\boldsymbol{\beta}| H(\mathrm{d}\boldsymbol{\beta}) < +\infty$,

then $\mathcal{P}$ is consistent at G_0.

The compactness of the covariate space is a standard assumption in nonparametric regression problems. Condition (i) is fairly reasonable, since it is guaranteed by a correct specification of the RUM: one can always redefine the set of choices or the covariate space to fulfilled this requirement. Condition (ii) is a mild condition on the tails of the prior predictive distribution H: it is satisfied by any distribution with tails lighter than the Cauchy distribution.

2.2 Illustration

It is worth considering Condition (ii) more in details for a variety of Bayesian MMNL models, obtained from different specification of $\mathcal{P}$. When $\tilde{G}$ is taken to be a Dirichlet process with base measure $\alpha = aF$, where $a > 0$ is a constant and $F \in \mathbb{P}$, then F coincides with H in (5). A larger class of Bayesian MMNL models arise when $\tilde{G}$ is chosen to be a stick-breaking prior:

$$\tilde{G}(\cdot) = \sum_{k\geq 1} p_k \delta_{Z_k}(\cdot) \tag{8}$$

where the p_k are positive random probabilities chosen to be independent of Z_k and such that $\sum_{k\geq 1} p_k = 1$ a.s.. The Z_k's are random locations taken as independent draws from some nonatomic distribution F in $\mathbb{P}$. What characterizes a stick-breaking prior is that the random weights are expressible as $p_k = V_k \prod_{i=1}^{k-1}(1 - V_i)$, where V_k's are independent Beta(a_k, b_k) random variables for $a_k, b_k > 0$. Examples of random probability measures in this class are given in Ishwaran and James (2001), see also Pitman and Yor (1997) and Ishwaran and Zarepour (2000). They represent extensions of the Dirichlet process, which has $a_k = 1$ and $b_k = a$ $\forall k$, and they all have in common that the prior predictive distribution H coincides with F.

The class of NRMI is another valid choice for $\mathcal{P}$. Specifically, one can take $\tilde{G}(\cdot) = \tilde{\mu}(\cdot)/\tilde{\mu}(\mathcal{R}^d)$, where $\tilde{\mu}$ is a completely random measure with Poisson intensity measure $\nu(\mathrm{d}v, \mathrm{d}z) = \rho(\mathrm{d}v|z)\alpha(\mathrm{d}z)$ on $(0, +\infty) \times \mathcal{R}^d$. Here $\rho(\cdot\,|z)$ is a Lévy density on $(0, +\infty)$ for any z and α is a finite measure on $\mathcal{R}^d$ such that $\psi(u) := \int_{\mathcal{R}^d \times \mathcal{R}^+} (1 - \mathrm{e}^{-uv})\rho(\mathrm{d}v|z)\alpha(\mathrm{d}z) < \infty$, which is needed for guaranteeing that $\tilde{\mu}(\mathcal{R}^d) < \infty$ a.s.. It can be shown that $H(B) = \int_B \int_0^{+\infty} \mathrm{e}^{-\psi(u)} \big\{ \int_0^{+\infty} \mathrm{e}^{-uv} v \rho(\mathrm{d}v|z) \big\} \mathrm{d}u\, \alpha(\mathrm{d}z)$ for any Borel set B of $\mathcal{R}^d$. See also James, Lijoi and Prünster (2005). When $\rho(\mathrm{d}v|z) = \rho(\mathrm{d}v)$ for each z (homogeneous case), the prior predictive distribution reduces to

$$H(B) = \frac{\alpha(B)}{\alpha(\mathcal{R}^d)} \qquad \text{for any Borel } B \subset \mathcal{R}^d. \tag{9}$$

The homogeneous NRMI includes, as a special case, the Dirichlet process and belongs, together with the stick-breaking priors, to the class of *species sampling models*, for which (9) holds for some finite measure α. Note that all the nonparametric priors belonging to this class allow an easy verification of condition (ii).

The specification of the nonparametric prior in terms of a base measure α as in (9) allows to introduce more flexibility via an additional level in the hierarchal structure (4). If we let the base measure be indexed by a parameter θ, say α_θ, and θ be random with probability density $\pi(\theta)$ on some

Euclidean space Θ, then we obtain a mixture of Dirichlet process in the spirit of Antoniak (1974). Then, Condition (ii) has to be verified for the convolution

$$H(B) = \int_{\Theta} \int_{B} H_{\theta}(\mathrm{d}z)\ \pi(\theta)\mathrm{d}\theta, \quad \text{where } H_{\theta}(\mathrm{d}z) = \frac{\alpha_{\theta}(\mathrm{d}z)}{\alpha_{\theta}(\mathcal{R}^d)}. \tag{10}$$

It is quite straightforward to check that condition (ii) holds for the mixture of Dirichlet processes implemented in the analysis of non-panel data of Section 3.

Finally, consider the case of Dirichlet process mixture models of Lo (1984), where $\tilde{G}$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{R}^d$ with random density function specified as $\int_{\Theta} K(\boldsymbol{\beta}, \theta)\tilde{\Pi}(\mathrm{d}\theta)$. Here $K(\boldsymbol{\beta}, \theta)$ is a nonnegative kernel defined on $\mathcal{R}^d \times \Theta$ such that, for each $\theta \in \Theta$, $\int_{\mathcal{R}^d} K(z, \theta)\mathrm{d}z = 1$, while $\tilde{\Pi}$ is a Dirichlet process prior with base measure aF and F a probability measure on Θ. The distribution H is then absolutely continuous and is given by

$$H(B) = \int_{B} \int_{\Theta} K(z, \theta)F(\mathrm{d}\theta)\ \mathrm{d}z.$$

As in (10), verifying condition (ii) requires to study the tail properties of a convolution, this time of $K(z, \theta)$ with respect to $F(\mathrm{d}\theta)$. In the analysis of panel data, see Section 4, we adopt a Dirichlet mixture model as continuous nonparametric prior for $\tilde{G}$ where the verification of (ii) can be readily established.

3 Implementation for non-panel data

Assume we have a single observation for each individual and we want to account for the possibility of ties among different individuals' preferences. Therefore, we use a discrete nonparametric prior for the mixing distribution. Take $\tilde{G}$ to be a Dirichlet process with base measure aF and denote its law by $\mathcal{P}(\mathrm{d}G|aF)$, although the treatment can be easily extended to any other stick breaking prior. Then, representation (8) holds with random probabilities $p_1, p_2, \ldots$ at locations $Z_1, Z_2, \ldots$ which are iid draws from F. This translates into a Bayesian model for the MMNL as

$$\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) = \sum_{k \geq 1} p_k \frac{\exp\{\mathbf{x}_j' Z_k\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}_l' Z_k\}} \qquad \text{for } j \in \Phi. \tag{11}$$

One can then center $\tilde{G}$ on a parametric model like the GML in (2) by taking F to have normal density $\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})$. In a parametric Bayesian framework, by placing priors on $\boldsymbol{\mu}, \boldsymbol{\tau}$, one is able to get posterior estimates of $\boldsymbol{\mu}, \boldsymbol{\tau}$, but inference is restricted to the assumption of the GML model. The flexibility of the Bayesian nonparametric approach allows one to choose F based on convenience and ease of use and to utilize, for instance, the attractive features of GML models while still maintaining the robustness of a nonparametric approach.

In the case of the Dirichlet process, the parameters associated with F, for instance $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$, are considered fixed. As observed in Section 2, one can introduce more flexibility in the model by treating such parameters as random. Specifying $\theta = (\boldsymbol{\mu}, \boldsymbol{\tau})$, $F_\theta(\mathrm{d}\boldsymbol{\beta})$ to have density $\phi(\boldsymbol{\beta}|\theta)\mathrm{d}\boldsymbol{\beta}$ and $\pi(\theta)$ the density function for θ, the law of $\tilde{G}$ is given by the mixture $\int_\Theta \mathcal{P}(\mathrm{d}G|aF_\theta)\pi(\mathrm{d}\theta)$. Equivalently, using (8), a mixture of Dirichlet processes is defined by specifying each $Z_k \mid \theta$ to be iid F_θ. Notice that, conditional on θ, a prior guess for the choice probabilities is

$$\mathrm{E}\big[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \theta\big] = \int_{\mathcal{R}^d} \frac{\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}{\sum_{l\in\Phi}\exp\{\mathbf{x}_l'\boldsymbol{\beta}\}} F_\theta(\mathrm{d}\boldsymbol{\beta}) \qquad \text{for } j \in \Phi. \tag{12}$$

By the properties of the Dirichlet process, the prediction rule for the choice probabilities given $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ is given by

$$\mathrm{E}\big[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \theta, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n\big] = \frac{a}{a+n}\mathrm{P}(\{j\} \mid F_\theta, \mathbf{x}) + \sum_{i=1}^n \frac{1}{a+n} \frac{\exp\{\mathbf{x}_j'\boldsymbol{\beta}_i\}}{\sum_{l\in\Phi}\exp\{\mathbf{x}_l'\boldsymbol{\beta}_i\}}. \tag{13}$$

where $\mathrm{P}(\{j\} \mid F_\theta, \mathbf{x}) := \mathrm{E}\big[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \theta\big]$ is given in (12) with a notation consistent with (1). However, the variables $\boldsymbol{\beta}_i$'s are not observable, and hence one needs to implement computational procedures to draw from their posterior distribution.

In this framework, a reasonable algorithm to use is the blocked Gibbs sampler developed in Ishwaran and Zarepour (2000) and Ishwaran and James (2001). Indeed, since the multinomial logistic kernel does not form a conjugate pair for $\boldsymbol{\beta}$, marginal algorithms suffer from slow convergence, although strategies for overcoming this problem can be found in MacEachern and Muller (1998).

3.1 Blocked Gibbs sampling

In this section we discuss how to implement a blocked Gibbs sampling algorithm for drawing inference on a nonparametric hierarchical model with the structure

$$\begin{aligned}
Y_i \mid \boldsymbol{\beta}_i &\overset{\text{ind}}{\sim} L(Y_i, \boldsymbol{\beta}_i), \quad \text{for } i = 1, \ldots, n, \text{ and } Y_i \in \Phi, \\
\boldsymbol{\beta}_i \mid \tilde{G} &\overset{\text{iid}}{\sim} \tilde{G}, \quad \text{for } i = 1, \ldots, n, \\
\tilde{G} \mid \theta &\sim \mathcal{P}(\mathrm{d}G|aF_\theta), \\
\theta &\sim \pi(\mathrm{d}\theta),
\end{aligned} \tag{14}$$

where $L(Y_i, \boldsymbol{\beta}) = \exp\{\mathbf{x}'_{iY_i}\boldsymbol{\beta}\}/\sum_{j\in\Phi}\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta}\}$ is the probability for Y_i conditional on $\boldsymbol{\beta}_i$. The blocked Gibbs sampler utilizes the fact that a truncated Dirichlet process, discussed in Ishwaran and Zarepour (2000) and Ishwaran and James (2001), serves as a good approximation to the random probability measure $\tilde{G} \mid \theta$ in (14). We replace the conditional law $\mathcal{P}(\mathrm{d}G|aF_\theta)$ with the law of the random probability measure

$$\tilde{G}(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot) \qquad 1 \le N < \infty, \tag{15}$$

where $Z_k \mid \theta$ are iid F_θ and the random probabilities $p_1, \ldots, p_N$ are defined by the stick-breaking construction

$$p_1 = V_1 \text{ and } p_k = (1 - V_1)\cdots(1 - V_{k-1})V_k,\ k = 2, \ldots, N \tag{16}$$

with $V_1, V_2, \ldots, V_{N-1}$ iid Beta$(1, a)$ and $V_N = 1$, which ensures that $\sum_{k=1}^{N} p_k = 1$. The law of $\tilde{G} \mid \theta$ in (15) is referred to as truncated Dirichlet process and will be denoted as $\mathcal{P}^N(dG|\alpha F_\theta)$. Moreover, the limit as $N \to \infty$ will converge to a random probability measure with law $\mathcal{P}(\mathrm{d}G|aF_\theta)$. Indeed, the method yields an accurate approximation of the Dirichlet process for N moderately large since the truncation is exponentially accurate. Theorem 2 in Ishwaran and James (2001) provides an $\mathcal{L}_1$–error bound for the approximation of conditional density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ given θ. Let

$$\mu^N(\mathbf{Y}|\theta) = \int \left[\prod_{i=1}^{n} \int_{\mathcal{R}^d} L(Y_i, \boldsymbol{\beta}_i) G(\mathrm{d}\boldsymbol{\beta}_i)\right] \mathcal{P}^N(\mathrm{d}G|aF_\theta)$$

and $\mu(\mathbf{Y}|\theta)$ its limit under the prior $\mathcal{P}(\mathrm{d}G|aF_\theta)$. Then, one has

$$\left\|\mu^N - \mu\right\|_1 := \int \left|\mu^N(\mathbf{Y}|\theta) - \mu(\mathbf{Y}|\theta)\right| \mathrm{d}\mathbf{Y} \sim 4n\mathrm{e}^{-(N-1)/a},$$

where the integral above is considered over the counting measure on the n-fold product space Φ^n. See Ishwaran and James (2001) for more details.

The key to working with random probability measures like (15) is that it allows to perform blocked updates for $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)$ by recasting the hierarchical model (14) completely in terms of random variables. To this aim, define the classification variables $\mathbf{K} = \{K_1, \ldots, K_n\}$ such that, conditional on $\mathbf{p}$, each K_i is independent with distribution

$$\mathrm{P}\{K_i \in \cdot \mid \mathbf{p}\} = \sum_{k=1}^{N} p_k \delta_k(\cdot).$$

That is $\mathrm{P}\{K_i = k \mid \mathbf{p}\} = p_k$ for $k = 1, \ldots, N$ so that K_i identifies the Z_k associated with each $\boldsymbol{\beta}_i$: $\boldsymbol{\beta}_i = Z_{K_i}$. In this setting a sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ from (15) produces $n_0 \le \min(n, N)$ distinct values.

The blocked Gibbs algorithm is based on sampling $\mathbf{K}, \mathbf{p}, \mathbf{Z}, \theta$ from the distribution proportional to

$$\left[\prod_{i=1}^{n} L(Y_i, \boldsymbol{\beta}_i)\right] \left[\prod_{i=1}^{n} \sum_{k=1}^{N} p_k \delta_{Z_k}(\mathrm{d}\boldsymbol{\beta}_i)\right] \pi(\mathbf{p}) \left[\prod_{k=1}^{N} F_\theta(\mathrm{d}Z_k)\right] \pi(\mathrm{d}\theta),$$

where $\pi(\mathbf{p})$ denotes the distribution of $\mathbf{p}$ defined in (16). This augmented likelihood is an expression of the augmented density when $\mathcal{P}(\mathrm{d}G|aF_\theta)$ is replaced by $\mathcal{P}^N(\mathrm{d}G|aF_\theta)$.

Before describing the algorithm, we specify choices for F_θ and θ which agree with the GML model. Set $\theta = (\boldsymbol{\mu}, \boldsymbol{\tau})$ and specify the density of F_θ to be $\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})$. Let λ denote a positive scalar. We choose a Multivariate Normal-Inverse Wishart distribution for $\boldsymbol{\mu}, \boldsymbol{\tau}$, where specifically $\boldsymbol{\mu} \mid \boldsymbol{\tau}$ is a Multivariate Normal vector with mean parameter $\mathbf{m}$ and scaled covariance matrix $\lambda^{-1}\boldsymbol{\tau}$ and $\boldsymbol{\tau}$ is drawn from an Inverse-Wishart distribution with degrees of freedom ν_0 and scale matrix $\mathbf{S}_0$. Denote this distribution for $\boldsymbol{\mu}, \boldsymbol{\tau}$ as N-IW$(\mathbf{m}, \lambda^{-1}\boldsymbol{\tau}, \nu_0, \mathbf{S}_0)$. Our specification is similar to that used in Train (2003, Chapter 12) for a parametric GML model for panel data.

Algorithm 1

1. *Conditional draw for* $\mathbf{K}$. Independently sample K_i according to $\mathrm{P}\{K_i \in \cdot|\mathbf{p}, \mathbf{Z}, \mathbf{Y}\} = \sum_{k=1}^{N} p_{k,i}\, \delta_k(\cdot)$, for $i = 1, \ldots, n$, where

$$(p_{1,i}, \ldots, p_{N,i}) \propto \big(p_1 L(Y_i, Z_1), \ldots, p_N L(Y_i, Z_N)\big).$$

2. *Conditional draw for* $\mathbf{p}$. $p_1 = V_1^*$, $p_k = (1 - V_1^*) \cdots (1 - V_{k-1}^*) V_k^*$, $k = 2, \ldots, N-1$ and $V_N^* = 1$ where, if e_k records the number of K_i values which equal k,

$$V_k^* \overset{\text{ind}}{\sim} \mathrm{Beta}\Big(1 + e_k,\ a + \textstyle\sum_{l=k+1}^{N} e_l\Big), \quad k = 1, \ldots, N-1.$$

3. *Conditional draw for* $\mathbf{Z}$. Let $\{K_1^*, \ldots, K_{n_0}^*\}$ denote the unique set of K_i values.
For each $k \notin \{K_1^*, \ldots, K_{n_0}^*\}$ draw $Z_k \mid \boldsymbol{\mu}, \boldsymbol{\tau}$ from the prior Multivariate Normal density $\phi(Z|\boldsymbol{\mu}, \boldsymbol{\tau})$.
For $j = 1, \ldots, n_0$, draw $Z_{K_j^*} := \boldsymbol{\beta}_j^*$ from the density proportional to $\phi(\boldsymbol{\beta}_j^*|\boldsymbol{\mu}, \boldsymbol{\tau}) \prod_{\{i\,:\,K_i = K_j^*\}} L(Y_i, \boldsymbol{\beta}_j^*)$ by using, for example, a standard Metropolis-Hastings procedure.

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

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

where $\bar{\boldsymbol{\beta}}_{n_0} = n_0^{-1} \sum_{j=1}^{n_0} \boldsymbol{\beta}_j^*$. Conditional on $\mathbf{K}, \mathbf{Z}, \mathbf{Y}$, draw $\boldsymbol{\tau}$ from and Inverse–Wishart distribution with parameters

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

where

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

Notice that, when $n_0 = 1$, Steps 3 and 4 reduce to the MCMC steps for a parametric Bayesian model. Iterating the steps above produces a draw from the distribution $\mathbf{Z}, \mathbf{K}, \mathbf{p}, \theta \mid \mathbf{Y}$. Thus, each iteration m defines a probability measure $G^{(m)}(\cdot) = \sum_{k=1}^{N} p_k^{(m)} \delta_{Z_k^{(m)}}(\cdot)$, which eventually approximates draws from the posterior distribution of $\tilde{G} \mid \mathbf{Y}$. Consequently, one can approximate the posterior distributional properties of the choice probabilities $\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x})$ by constructing (iteratively)

$$\mathrm{P}(\{j\} \mid G^{(m)}, \mathbf{x}) = \sum_{k=1}^{N} p_k^{(m)} \frac{\exp\{\mathbf{x}_j' Z_k^{(m)}\}}{\sum_{l \in \Phi} \exp\{\mathbf{x}_l' Z_k^{(m)}\}} \ ,$$

see (11). For instance, an histogram of the $\mathrm{P}(\{j\} \mid G^{(m)}, \mathbf{x})$, for $m = 1, \ldots, M$, approximates the posterior distribution. An approximation to the posterior mean $\mathrm{E}[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \mathbf{Y}]$ is obtained by $M^{-1} \sum_{m=1}^{M} \mathrm{P}(\{j\} \mid G^{(m)}, \mathbf{x})$ or, alternatively, by

$$\widehat{P}(\{j\} \mid \mathbf{x}) := \frac{1}{M} \sum_{m=1}^{M} \mathrm{E}\big[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \theta^{(m)}, \boldsymbol{\beta}_1^{(m)}, \ldots, \boldsymbol{\beta}_n^{(m)}\big]. \tag{17}$$

where $\mathrm{E}\big[\mathrm{P}(\{j\} \mid \tilde{G}, \mathbf{x}) \mid \theta, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n\big]$ is given in (13) and $\boldsymbol{\beta}_i^{(m)} = Z_{K_i^{(m)}}^{(m)}$.

4 Bayesian modeling 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 at different time points. The random utility for choosing j for individual i in choice situation t is given by

$$U_{ij\,t} = \mathbf{x}_{ij\,t}' \boldsymbol{\beta}_i + \varepsilon_{ij\,t}, \quad j \in \Phi$$

for times $t = 1, \ldots, T_i$. The MMNL model can be described as follows [see Train (2003, Section 6.7)]: given $\boldsymbol{\beta}_i$, the probability that a person makes the sequence of choices $\mathbf{Y}_i = \{Y_{i1}, \ldots, Y_{iT_i}\}$ is the product of logit formulas

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

The MMNL model is completed by taking the $\boldsymbol{\beta}_i$'s to be from a distribution G so that the unconditional choice probability is specified by

$$\mathrm{P}(\mathbf{Y}_i \mid G, \mathbf{x}_i) = \int_{\mathcal{R}^d} \prod_{t=1}^{T_i} \frac{\exp\{\mathbf{x}'_{iY_{it}t}\boldsymbol{\beta}\}}{\sum_{j\in\Phi}\exp\{\mathbf{x}'_{ij\,t}\boldsymbol{\beta}\}}\, G(\mathrm{d}\boldsymbol{\beta}) = \int_{\mathcal{R}^d} L(\mathbf{Y}_i, \boldsymbol{\beta})\, G(\mathrm{d}\boldsymbol{\beta}),$$

where $\mathbf{x}_i = \{\mathbf{x}_{ij\,t},\ j \in \Phi, t = 1, \ldots, T_i\}$ denotes the array of covariates associated to the sequence of choices of individual i. Similar to the non-panel data setting, we wish to model $\tilde{G}$ as a random probability measure in a Bayesian framework. While it is possible to choose $\tilde{G}$ to follow a Dirichlet process, this would result in possible ties among the individual's preferences $\boldsymbol{\beta}_i$. In order to preserve the distinct nature of each individual's preference, we assume that, given $\tilde{G}$, the $\boldsymbol{\beta}_i$'s are iid with distribution $\tilde{G}$, where $\tilde{G}$ is a mixture of multivariate normal distributions with random mixing distribution $\tilde{\Pi}$. That is, $\tilde{G}$ has random density $\int_\Theta \phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})\tilde{\Pi}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\tau})$, where $\Theta = \mathcal{R}^d \times \mathcal{S}$ with $\mathcal{S}$ the space of covariance matrices. Specifically, we take $\tilde{\Pi}$ to be a Dirichlet process with shape aF, F a probability measure on Θ. Hence, the Bayesian MMNL model for individual i is expressible as

$$\mathrm{P}(\mathbf{Y}_i \mid \tilde{G}, \mathbf{x}_i) = \int_{\mathcal{R}^d} L(\mathbf{Y}_i, \boldsymbol{\beta})\tilde{G}(\mathrm{d}\boldsymbol{\beta}) = \int_{\mathcal{R}^d}\int_\Theta L(\mathbf{Y}_i, \boldsymbol{\beta})\phi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\tau})\tilde{\Pi}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\tau})\, \mathrm{d}\boldsymbol{\beta}.$$

While one may use any choice for F, we take $F(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\tau})$ to be the Multivariate Normal-Inverse-Wishart distribution N-IW$(\mathbf{m}, \lambda^{-1}\boldsymbol{\tau}, \mathbf{S}_0, \nu_0)$ described in Section 3.

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 $(\boldsymbol{\mu}_i, \boldsymbol{\tau}_i)$, $i = 1, \ldots, n$, rather than $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$, 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 sampler. The joint distribution of the augmented data can be expressed using a hierarchical model as follows:

$$\begin{aligned}
\mathbf{Y}_i \mid \boldsymbol{\beta}_i &\overset{\text{ind}}{\sim} L(\mathbf{Y}_i, \boldsymbol{\beta}_i), \quad \text{for } i = 1, \ldots, n, \text{ and } Y_{it} \in \Phi,\\
\boldsymbol{\beta}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\tau}_i &\overset{\text{ind}}{\sim} \phi(\boldsymbol{\beta}_i|\boldsymbol{\mu}_i, \boldsymbol{\tau}_i) \quad \text{for } i = 1, \ldots, n\\
\boldsymbol{\mu}_i, \boldsymbol{\tau}_i \mid \tilde{\Pi} &\overset{\text{iid}}{\sim} \tilde{\Pi} \quad \text{for } i = 1, \ldots, n\\
\tilde{\Pi} &\sim \mathcal{P}(\mathrm{d}\Pi|aF)
\end{aligned} \tag{18}$$

Similar to the non-panel case, the blocked Gibbs sampler works by using the $\mathcal{P}^N(\mathrm{d}\Pi|aF)$ in place of the law of the Dirichlet process $\mathcal{P}(\mathrm{d}\Pi|aF)$. We now sample $(\mathbf{K}, \mathbf{p}, \mathbf{Z}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$ from the distribution proportional to

$$\left[\prod_{i=1}^{n} L(\mathbf{Y}_i, \boldsymbol{\beta}_i)\phi(\boldsymbol{\beta}_i|\boldsymbol{\mu}_i, \boldsymbol{\tau}_i)\right]\left[\prod_{i=1}^{n}\sum_{k=1}^{N} p_k \delta_{Z_k}(\mathrm{d}\boldsymbol{\mu}_i, \mathrm{d}\boldsymbol{\tau}_i)\right]\pi(\mathbf{p})\prod_{k=1}^{N} F(\mathrm{d}Z_k).$$

Here we use the fact that $(\boldsymbol{\mu}_i, \boldsymbol{\tau}_i) = Z_{K_i}$, for $i = 1, \ldots, n$. To approximate the posterior law of various functionals cycle through the following steps:

ALGORITHM 2

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

$$\mathrm{P}\{K_i \in \cdot \mid \mathbf{p}, \mathbf{Z}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \mathbf{Y}\} = \sum_{k=1}^{N} p_{k,i}\, \delta_k(\cdot), \quad \text{for } i = 1, \ldots, n,$$

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

2. *Conditional draw for* $\mathbf{p}$. $p_1 = V_1^*$, $p_k = (1 - V_1^*) \cdots (1 - V_{k-1}^*)V_k^*$, $k = 2, \ldots, N-1$ and $V_N^* = 1$ where, for e_k records the number of K_i values which equal k,

$$V_k^* \overset{\text{ind}}{\sim} \mathrm{Beta}\Big(1 + e_k,\ a + \textstyle\sum_{l=k+1}^{N} e_l\Big), \quad k = 1, \ldots, N-1.$$

3. *Conditional draw for* $\mathbf{Z}$. Let $\{K_1^*, \ldots, K_{n_0}^*\}$ denote the unique set of K_i values. For each $k \notin \{K_1^*, \ldots, K_{n_0}^*\}$ draw $Z_k = (\boldsymbol{\mu}_k, \boldsymbol{\tau}_k)$ from the prior N-IW$(\mathbf{m}, \lambda^{-1}\boldsymbol{\tau}, \mathbf{S}_0, \nu_0)$. For $j = 1, \ldots, n_0$, draw $Z_{K_j^*} := (\boldsymbol{\mu}_j^*, \boldsymbol{\tau}_j^*)$ as follows: (a) conditional on $\boldsymbol{\tau}_j^*, \mathbf{K}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \mathbf{Y}$, 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 $\bar{\boldsymbol{\beta}}_j^* = (e_{K_j^*})^{-1}\sum_{\{i\,:\,K_i = K_j^*\}}\boldsymbol{\beta}_i$; (b) conditional on $\mathbf{K}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, \mathbf{Y}$, 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

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

4. *Conditional draw for* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$. For each $j = 1, \ldots, n_0$, draw independently $\boldsymbol{\beta}_i$ for $i \in \{l : K_l = K_j^*\}$ from the density proportional to $L(\mathbf{Y}_i, \boldsymbol{\beta}_i)\phi(\boldsymbol{\beta}_i|\boldsymbol{\mu}_j^*, \boldsymbol{\tau}_j^*)$ by using, for example, a standard Metropolis-Hastings procedure.

5 Simulation Study

In this section we present some empirical evidence that shows how the MMNL procedures perform overall and relative to GML models. We proceed to the estimation of the choice probabilities and the mixing distribution based on simulated data. Two different artificial datasets are generated for the simulation study: the first set (dataset 1) is produced for studying non-panel data model, while the second set (dataset 2) is designed to study models with panel data. In both cases we consider a RUM with three possible responses ($J = 3$) relative to the utilities U_1, U_2 and U_3,

$$\begin{cases} U_1 &= x_{11}\beta_1 + x_{12}\beta_2 + \varepsilon_1 \\ U_2 &= x_{21}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\ U_3 &= x_{31}\beta_1 + x_{32}\beta_2 + \varepsilon_3 \end{cases}$$

As for dataset 1, we choose $\varepsilon_1, \varepsilon_2, \varepsilon_3 \overset{\text{iid}}{\sim}$ Standard Gumbel and $\boldsymbol{\beta} = (\beta_1, \beta_2)' \overset{\text{iid}}{\sim} 0.5 \times \delta_{(-5,5)} + 0.5 \times \delta_{(5,-5)}$. For each individual i we randomly generate (componentwise) its covariates $\mathbf{x}_i = (x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32})$ independently from a Uniform $(-2, 2)$ distribution. Set $Y_i = j$ if $U_{ij} > U_{il}$, $l \neq j$, for $j = 1, 2, 3$. Repeat this procedure n times independently to get a dataset with $(Y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$. As for dataset 2, we assume there are n individuals each making $T_i = 10$ choices for $i = 1, \ldots, n$. Then, we simulate data using the same model used to generate dataset 1. The only change is that $\boldsymbol{\beta}$ is drawn from the two component mixture of bivariate normal distributions, $\boldsymbol{\beta} \overset{\text{iid}}{\sim} 0.5 \times N((-5, 5)', 2\mathbf{I}) + 0.5 \times N((5, -5)', 2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. We apply our procedures to the estimation of choice probabilities $\mathrm{P}(\{j\} \mid G, \mathbf{x})$ for $j = 1, 2, 3$ based on the set of covariates $\mathbf{x} = (1.0, -0.9, 1.0, 0.2, 1.0, 0.9)$. We also sample $\boldsymbol{\beta}$ variates from their posterior and get the estimate of the mixing distribution G.

The prior parameters for the specifications of the Bayesian MMNL models for panel and non-panel data (pertaining to the explicit models in Section 3 and 4) are set to be $a = 1$, $\nu_0 = 2$, $\mathbf{m} = (0, 0)'$ and $\mathbf{S}_0 = \mathbf{I}$. Additionally we use $N = 100$ and perform estimation for different choices of the scale parameter λ. A parametric GML model is also estimated for comparison with the same specifications for ν_0, $\mathbf{m}$, $\mathbf{S}_0$ and combinations of λ's. In all cases we use the estimator (17) based on an initial burn-in of 10000 cycles and additional 10000 Gibbs cycles ($M = 10000$) for the estimation. In addition, to measure how good of our estimates are, we define a measure, Root Mean Square (RMS) value, as

$$\mathrm{RMS} = \sqrt{\frac{1}{J} \sum_{j \in \Phi} \frac{1}{M} \sum_{m=1}^{M} \Big(\mathrm{P}(\{j\} \mid G^{(m)}, \mathbf{x}) - \mathrm{P}_{\text{true}}(\{j\} \mid \mathbf{x}) \Big)^2}.$$

Simulation results using dataset 1 ($n = 500$) and dataset 2 ($n = 100, T_i = 10$) are summarized in Table 1 for $\lambda = 0.01, 0.1, 1$, together with RMS values, for both the GML and the MMNL models. They show that the performance of the nonparametric MMNL estimators is overall better than that of the parametric GML model, as indicated by a smaller RMS, and that the improvement is more evident for the non-panel case than for the panel case. Moreover, as we decrease λ, the estimates of choice probabilities in the MMNL model remain stable for the non-panel case and are more accurate for the panel case. An interpretation of an increase of accuracy is as follows: a smaller λ corresponds to a more diffuse H, the prior predictive distribution of $\tilde{G}$. Since H is different from the distribution used to simulate the $\boldsymbol{\beta}$'s in the data generating process, we get evidence that a diffuse H helps in capturing the true form of the mixing distribution G. Note also that a smaller λ yields a smaller RMS, the latter being a measure of the combination of the accuracy and the variability of the posterior variates of $\mathrm{P}(\{j\} \mid G, \mathbf{x})$. An examination of their autocorrelation functions along the chain shows that a smaller λ causes a slower mixing of the Gibbs sampler, which increases the component of variability in the RMS. See Figure 1. The decrease in RMS then shows that such precision loss is more than balanced by an higher accuracy of the estimate, although one should also control the convergence properties of the Gibbs sampler by avoiding to take λ too small.

[Table 1 and Figure 1 about here]

We also investigated the sensitivity of the results to the prior parameter ν_0, where a larger ν_0 corresponds to a more concentrated Inverse-Wishart distribution on $\mathbf{S}_0$. We did not observe sensible differences in the estimation by varying ν_0 and we decided to set $\nu_0 = 2$ and $\mathbf{S}_0 = \mathbf{I}$ as a default noninformative choice for these parameters, see Train (2003, Section 12). The nonparametric prior on G is also dependent on the total mass a, which is positively related to the number of components in the mixture distribution of the $\boldsymbol{\beta}$'s. Generally, $a = 1$ is considered a default choice for a finite mixture model with fixed but uncertain number of components. We also performed estimation for larger a, observing almost identical results: $a = 1$ was in fact sufficient for detecting the two–components mixture we used in generating the data. Although we have not done so, the blocked Gibbs procedures described in Section 3 and 4 can be easily extended to place an additional prior on a. Furthermore, the truncation level of $N = 100$ in (15) is sufficiently large as we observed almost identical estimation results from runs of the blocked Gibbs sampler with larger values of N.

Finally, we perform estimation using the MMNL model for different sample sizes for both dataset 1 and dataset 2 in order to get confirmation of the consistency results of Section 2. The prior parameters are set to be $a = 1$, $\nu_0 = 2$, $\mathbf{m} = (0,0)'$ and $\mathbf{S}_0 = \mathbf{I}$, $N = 100$ and $\lambda = 1$. Table 2 reports the results by showing, in fact, a sensible decrease of RMS for both non-panel and panel data as the number of observations increases. In addition, Figure 2 reports the histograms of samples for β_1 from its marginal posterior distribution against the mixing distribution used in the data generating process: it shows how the approximation of the true mixing distribution G improves as more and more data become available.

[Table 2 and Figure 2 about here]

6 Proof of Theorem 1

Throughout this section, we work with the family of multinomial logistic kernels:

$$k_j(\mathbf{x}, \boldsymbol{\beta}) = \frac{\exp(x_j'\boldsymbol{\beta})}{\sum_{l\in\Phi}\exp(x_l'\boldsymbol{\beta})}, \quad j = 1, \ldots, J.$$

For $q_j(\mathbf{x}; G)$ denoting the j-th element of the vector $\mathbf{q}(\mathbf{x}; G)$, we have that $q_j(\mathbf{x}; G) = \int_{\mathcal{R}^d} k_j(\mathbf{x}, \boldsymbol{\beta}) G(\mathrm{d}\boldsymbol{\beta})$. Note that $q_Y(\mathbf{x}; G_0)$ is the joint density of $(Y, \mathbf{X})$ with respect to the counting measure on the integer set Φ and the measure $M(\mathrm{d}\mathbf{x})$ on $\mathcal{X}$.

For the proof of Theorem 1 the following Lemma is essential, stating that, on the space $\mathbb{P}$, the weak topology and the topology induced by the $\mathcal{L}_1$-distance d defined in (7) are equivalent.

Lemma 1 . *Let d_w be any distance that metrizes the weak topology on $\mathbb{P}$ and $(G_n)_{n\geq 1}$ be a sequence in $\mathbb{P}$. Then $d_w(G_n, G_0) \to 0$ iff $d\big(\mathbf{q}(\cdot\,; G_n), \mathbf{q}(\cdot\,; G_0)\big) \to 0$.*

Proof. For the "if" part, it is sufficient that $d_w(G_n, G_0) \to 0$ implies $\int_{\mathcal{X}} \big|q_j(\mathbf{x}; G_n) - q_j(\mathbf{x}; G_0)\big| M(\mathrm{d}\mathbf{x}) \to 0$ for an arbitrary $j \in \Phi$. The latter is a consequence of the definition of weak convergence and an application of Scheffé's theorem, since $k_j(\mathbf{x}, \boldsymbol{\beta})$ is bounded and continuous in $\boldsymbol{\beta}$ for each $\mathbf{x} \in \mathcal{X}$. To show the converse, we prove that G distant from G_0 in the weak topology implies that $\mathbf{q}(\cdot\,; G)$ is distant from $\mathbf{q}(\cdot\,; G_0)$ in the $\mathcal{L}_1$-distance d. Define a weak neighborhood of G_0 as

$$V = \left\{G : \left| \int_{\mathcal{R}^d}\int_{\mathcal{X}} k_j(\mathbf{x}, \boldsymbol{\beta}) M(\mathrm{d}\mathbf{x})\; G(\mathrm{d}\boldsymbol{\beta}) - \int_{\mathcal{R}^d}\int_{\mathcal{X}} k_j(\mathbf{x}, \boldsymbol{\beta}) M(\mathrm{d}\mathbf{x})\; G_0(\mathrm{d}\boldsymbol{\beta})\right| < \delta,\ j \in \Phi\right\}.$$

Since $\int_{\mathcal{X}} k_j(\mathbf{x}, \boldsymbol{\beta}) M(\mathrm{d}\mathbf{x})$ is a bounded continuous function on $\mathcal{R}^d$ for each j, $G \in V^c$ implies that $d_w(G, G_0) > \delta$. Based on the following inequalities

$$\begin{aligned} d\big(\mathbf{q}(\cdot\,; G), \mathbf{q}(\cdot\,; G_0)\big) &\geq \max_{j\in\Phi} \int_{\mathcal{X}} \big|q_j(\mathbf{x}; G_n) - q_j(\mathbf{x}; G_0)\big| M(\mathrm{d}\mathbf{x}) \\ &\geq \max_{j\in\Phi} \left| \int_{\mathcal{X}}\int_{\mathcal{R}^d} k_j(\mathbf{x}, \boldsymbol{\beta}) G(\mathrm{d}\boldsymbol{\beta})\; M(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}}\int_{\mathcal{R}^d} k_j(\mathbf{x}, \boldsymbol{\beta}) G_0(\mathrm{d}\boldsymbol{\beta})\; M(\mathrm{d}\mathbf{x})\right| \end{aligned}$$

and an application of Fubini's theorem, it follows that, for any $\epsilon < \delta$ and any $G \in V^c$, $d\big(\mathbf{q}(\cdot\,; G), \mathbf{q}(\cdot\,; G_0)\big) > \epsilon$. The proof is then complete. $\square$

Remark 1 . Lemma 1 has two important consequences: (a) both $\mathscr{Q}$ and $\mathbb{P}$ are separable spaces under the metric d, (b) the statement of Theorem 1 is equivalent to saying that $\mathcal{P}_n$ accumulates all probability mass in weak neighborhood of G_0.

Define $\Lambda_n(G) = \prod_{i=1}^n q_{Y_i}(\mathbf{X}_i; G)/q_{Y_i}(\mathbf{X}_i; G_0)$, such that the posterior distribution of G can be written as

$$\mathcal{P}_n(A) = \frac{\int_A \Lambda_n(G)\mathcal{P}(\mathrm{d}G)}{\int_{\mathbb{P}} \Lambda_n(G)\mathcal{P}(\mathrm{d}G)} \tag{19}$$

We now take $A = \{G :\ d\big(\mathbf{q}(\cdot\,; G), \mathbf{q}(\cdot\,; G_0)\big) > \epsilon\}$, and we will, as is usual with the Bayesian consistency literature, consider separately the numerator and the denominator of (19). To this aim, define $I_n = \int_{\mathbb{P}} \Lambda_n(G)\mathcal{P}(\mathrm{d}G)$. Relying on the separability of $\mathbb{P}$ under the topology induced by d, see Remark 1, for any $\eta > 0$ we can cover A with a countable union of disjoint sets A_j such that

$$A_j \subseteq A_j^* = \{G :\ d\big(\mathbf{q}(\cdot\,; G), \mathbf{q}(\cdot\,; G_j)\big) < \eta\}, \tag{20}$$

and $\{G_j\}_{j\geq 1}$ is a countable set in $\mathbb{P}$ such that $d\big(\mathbf{q}(\cdot\,; G_j), \mathbf{q}(\cdot\,; G_0)\big) > \epsilon$ for any j. Consider that

$$\mathcal{P}_n(A) = \sum_{j\geq 1} \mathcal{P}_n(A_j) \leq \sum_{j\geq 1} \sqrt{\mathcal{P}_n(A_j)} = \sum_{j\geq 1} \sqrt{I_n^{-1} \int_{A_j} \Lambda_n(G)\mathcal{P}(\mathrm{d}G)}.$$

Hence, Theorem 1 holds if we prove that, for all large n,

$$\forall c > 0, \qquad I_n > \exp(-nc) \quad \text{a.s.} \tag{21}$$

$$\exists b > 0 : \sum_{j\geq 1} \sqrt{\textstyle\int_{A_j} \Lambda_n(G)\mathcal{P}(\mathrm{d}G)} < \exp(-nb) \quad \text{a.s.} \tag{22}$$

As for (21), consider the Kulback-Leibler (K-L) support condition of $\mathcal{P}$ defined by

$$\mathcal{P}\left\{G :\ \int_{\mathcal{X}} K(G_0, G|\,\mathbf{x})M(\mathrm{d}\mathbf{x}) < \epsilon\right\} > 0, \quad \forall \epsilon > 0 \tag{23}$$

where $K(G_0, G|\,\mathbf{x}) = \sum_{j\in\Phi} q_j(\mathbf{x}; G_0)\log[q_j(\mathbf{x}; G_0)/q_j(\mathbf{x}; G)]$. If $\mathcal{P}$ satisfies condition (23), then (21) holds. To see this, it is sufficient to note that the K-L divergence of $q_Y(\mathbf{X}; G)$ from $q_Y(\mathbf{X}; G_0)$ with respect to the measure $M(\mathrm{d}\mathbf{x})$ on $\mathcal{X}$ and the counting measure on Φ is given by $\int K(G, G_0|\,\mathbf{x})M(\mathrm{d}\mathbf{x})$. Then, by the compactness of $\mathcal{X}$, the law of large number leads to

$$\frac{1}{n}\sum_{i=1}^n \log\frac{q_{Y_i}(\mathbf{X}_i; G_0)}{q_{Y_i}(\mathbf{X}_i; G)} \to \int_{\mathcal{X}} K(G_0, G \mid \mathbf{x})M(\mathrm{d}\mathbf{x}) \qquad \text{a.s..}$$

Result in (21) then follows from standard arguments, see, e.g., Wasserman (1998). Lemma 2 below states that (23) is satisfied under the hypotheses of Theorem 1.

LEMMA 2 . *If G_0 lies in the weak support of $\mathcal{P}$ and Condition* (i) *of Theorem 1 holds, then G_0 is in the K-L support of $\mathcal{P}$ according to* (23).

PROOF. It is sufficient to show that, for any $j \in \Phi$ and for any $\eta < 1$, there exists a δ such that $|q_j(\mathbf{x}; G)/q_j(\mathbf{x}; G_0) - 1| \leq \eta$ whenever G is in W_δ, a δ-weak neighborhood of G_0. In fact, this implies that:

$$\begin{aligned}\int_{\mathcal{X}} q_j(\mathbf{x}; G_0) \log\left[\frac{q_j(\mathbf{x}; G_0)}{q_j(\mathbf{x}; G)}\right] M(\mathrm{d}\mathbf{x}) &\leq \int_{\mathcal{X}} q_j(\mathbf{x}; G_0)\left|\frac{q_j(\mathbf{x}; G_0)}{q_j(\mathbf{x}; G)} - 1\right| M(\mathrm{d}\mathbf{x}) \\ &\leq \int_{\mathcal{X}} q_j(\mathbf{x}; G_0)\Big(\frac{\eta}{1-\eta}\Big) M(\mathrm{d}\mathbf{x}) \leq \frac{\eta}{1-\eta}\end{aligned}$$

which, in turn, leads to the thesis by the arbitrariness of j.

Let $c = \inf_{\mathbf{x} \in \mathcal{X}} q_j(\mathbf{x}; G_0)$, which is positive by Condition (i) of Theorem 1, and assume $G \in W_\delta$ for a δ that will be determined later. Note that, for any $\rho > 0$, one can set $M_\rho > 0$ such that $G_0\{\boldsymbol{\beta} : |\boldsymbol{\beta}| > M_\rho - \delta\} < \rho$. Then, using the Prokhorov's metric, $G \in W_\delta$ implies that $G\{\boldsymbol{\beta} : |\boldsymbol{\beta}| > M_\rho\} < \rho + \delta$. Note also that the family of functions $\{k_j(\mathbf{x}, \boldsymbol{\beta}),\ \mathbf{x} \in \mathcal{X}\}$, as $\boldsymbol{\beta}$ varies in the compact set $\{|\boldsymbol{\beta}| \leq M_\rho\}$, is uniformly equicontinuous. By an application of the Arzelà-Ascoli's theorem we know that, given a $\gamma > 0$, there exist finitely many points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ such that, for any $\mathbf{x} \in \mathcal{X}$, there is an index i such that

$$\sup_{|\boldsymbol{\beta}| \leq M_\rho} \big|k_j(\mathbf{x}, \boldsymbol{\beta}) - k_j(\mathbf{x}_i, \boldsymbol{\beta})\big| < \gamma. \tag{24}$$

For an arbitrary $\mathbf{x} \in \mathcal{X}$, choose the appropriate $\mathbf{x}_i$ such that (24) holds, so that

$$\begin{aligned}\left|\frac{q_j(\mathbf{x}; G)}{q_j(\mathbf{x}; G_0)} - 1\right| \leq \frac{1}{c}\Bigg(&\left|\int k_j(\mathbf{x}_i, \boldsymbol{\beta})\, G(\mathrm{d}\boldsymbol{\beta}) - \int k_j(\mathbf{x}_i, \boldsymbol{\beta})\, G_0(\mathrm{d}\boldsymbol{\beta})\right| + \\ &\int \big|k_j(\mathbf{x}, \boldsymbol{\beta}) - k_j(\mathbf{x}_i, \boldsymbol{\beta})\big|\, G(\mathrm{d}\boldsymbol{\beta}) + \int \big|k_j(\mathbf{x}, \boldsymbol{\beta}) - k_j(\mathbf{x}_i, \boldsymbol{\beta})\big|\, G_0(\mathrm{d}\boldsymbol{\beta})\Bigg) := \frac{I_1 + I_2 + I_3}{c}\end{aligned}$$

We have that $G \in W_\delta$ implies $I_1 \leq \delta$. As for I_2, we have

$$\begin{aligned}I_2 &= \int_{|\boldsymbol{\beta}| \leq M_\rho} \big|k_j(\mathbf{x}, \boldsymbol{\beta}) - k_j(\mathbf{x}_i, \boldsymbol{\beta})\big| G(\mathrm{d}\boldsymbol{\beta}) + \int_{|\boldsymbol{\beta}| > M_\rho} \big|k_j(\mathbf{x}, \boldsymbol{\beta}) - k_j(\mathbf{x}_i, \boldsymbol{\beta})\big| G(\mathrm{d}\boldsymbol{\beta}) \\ &\leq \gamma + 2\, G\{\boldsymbol{\beta} :\ |\boldsymbol{\beta}| > M_\rho\} \leq \gamma + 2(\rho + \delta).\end{aligned}$$

Similar arguments lead to $I_3 \leq \gamma + 2\rho$. Finally we get

$$\left| \frac{q_j(\mathbf{x}; G)}{q_j(\mathbf{x}; G_0)} - 1 \right| \leq \frac{3\delta + 2\gamma + 4\rho}{c},$$

so that, for given $\eta < 1$, it is always possible to choose δ, ρ (by tightness of G_0) and γ (by the Arzelà-Ascoli's theorem) small enough such that the right hand side in the last inequalities is smaller than η. The proof is then complete. $\square$

We now aim at showing that (22) holds under the hypotheses of Theorem 1 by extending the method set forth by Walker (2004) for strong consistency. In order to simplify the notation, let $\Lambda_{n\,j} = \int_{A_j} \Lambda_n(G)\mathcal{P}(\mathrm{d}G)$, where $(A_j)_{j\geq 1}$ is the covering of A in (20). The following identity is the key:

$$\Lambda_{n+1\,j}/\Lambda_{n\,j} = q^{nA_j}_{Y_{n+1}}(\mathbf{X}_{n+1}) \Big/ q_{Y_{n+1}}(\mathbf{X}_{n+1}; G_0), \tag{25}$$

where $q^{nA_j}_l(\mathbf{X}_{n+1}) = \int_{\mathbb{P}} q_l(\mathbf{X}_{n+1}; G)\mathcal{P}_{nA_j}(\mathrm{d}G)$, $l \in \Phi$, and $\mathcal{P}_{nA_j}$ is the posterior distribution restricted, and normalized, to the set A_j. Note that (25) includes the case of $n = 0$ and $\Lambda_{0\,j} = \mathcal{P}(A_j)$. By using conditional expectation we have that:

$$\begin{aligned}
\mathrm{E}\big[\Lambda^{1/2}_{n+1\,j} \mid (Y_1, \mathbf{X}_1) \ldots, (Y_n, \mathbf{X}_n), \mathbf{X}_{n+1}\big] &= \Lambda^{1/2}_{n\,j} \sum_{l\in\Phi} \sqrt{q^{nA_j}_l(\mathbf{X}_{n+1}) q_l(\mathbf{X}_{n+1}; G_0)} \\
&= \Lambda^{1/2}_{n\,j} \Big(1 - h\big[\mathbf{q}^{nA_j}(\mathbf{X}_{n+1}), \mathbf{q}(\mathbf{X}_{n+1}; G_0)\big]\Big)
\end{aligned}$$

where $\mathbf{q}^{nA_j}(\mathbf{X}_{n+1}) = [q^{nA_j}_1(\mathbf{X}_{n+1}), \ldots, q^{nA_j}_J(\mathbf{X}_{n+1})]$ and, for $\mathbf{q}_1, \mathbf{q}_2 \in \Delta$,

$$h(\mathbf{q}_1, \mathbf{q}_2) = 1 - \textstyle\sum_{j\in\Phi} \sqrt{q_{1j} q_{2j}}\,.$$

Note that $h(\mathbf{q}_1, \mathbf{q}_2)$ is a variation of the Hellinger distance $\sqrt{\sum_{j\in\Phi}(q^{1/2}_{1j} - q^{1/2}_{2j})^2}$ on Δ and that $h(\mathbf{q}_1, \mathbf{q}_2) \leq 1$. By taking the conditional expectation with respect to $(Y_1, \mathbf{X}_1) \ldots, (Y_n, \mathbf{X}_n)$ only, we get the following identity:

$$\mathrm{E}\{\Lambda^{1/2}_{n+1\,j} \mid (Y_1, \mathbf{X}_1) \ldots, (Y_n, \mathbf{X}_n)\} = \Lambda^{1/2}_{n\,j} \Big(1 - \textstyle\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x})\Big). \tag{26}$$

Since the Hellinger distance and the Euclidean distance are equivalent metrics in Δ, it can be proved that, for $(\mathbf{q}_n)_{n\geq 1} \in \mathscr{Q}$ and $\mathbf{q}_0 \in \mathscr{Q}$,

$$\int_{\mathcal{X}} h\big[\mathbf{q}_n(\mathbf{x}), \mathbf{q}_0(\mathbf{x})\big] M(\mathrm{d}\mathbf{x}) \to 0 \quad \text{iff} \quad d(\mathbf{q}_n, \mathbf{q}_0) \to 0. \tag{27}$$

The equivalence in (27) can be used to show that $\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x})$ is bounded away from zero. In fact, take G_j defined in (20) and note that, by triangular inequality,

$$\begin{aligned}\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x}) \geq & \int_{\mathcal{X}} h\big[\mathbf{q}(\mathbf{x}; G_j), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x}) \\ & - \int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_j)\big] M(\mathrm{d}\mathbf{x}).\end{aligned}$$

Since $d\big(\mathbf{q}(\cdot\,; G_j), \mathbf{q}(\cdot\,; G_0)\big) > \epsilon$, (27) assures the existence of a positive constant, say ϵ_2, such that $\int_{\mathcal{X}} h\big[\mathbf{q}(\mathbf{x}; G_j), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x}) > \epsilon_2$. Now, choose η in (20) such that, for each $G \in A_j$, $\int_{\mathcal{X}} h\big[\mathbf{q}(\mathbf{x}; G), \mathbf{q}(\mathbf{x}; G_j)\big] M(\mathrm{d}$ ϵ_2, where we have used (27) again. Since $\mathbf{q}^{nA_j}(\mathbf{x})$ does not correspond exactly to a particular $G \in A_j$, we use the convexity of the distance $h\big[\mathbf{q}(\mathbf{x}; G), \mathbf{q}(\mathbf{x}; G_j)\big]$ in its first argument to show that $\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_j)\big] M(\mathrm{d}\mathbf{x}) < \epsilon_2$. Note in fact that, by Jensen's inequality,

$$\begin{aligned}\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_j)\big] M(\mathrm{d}\mathbf{x}) =& \int_{\mathcal{X}} \Big(1 - \sum_{l \in \Phi} \sqrt{\textstyle\int_{\mathbb{P}} q_l(\mathbf{X}_{n+1}; G) \mathcal{P}_{nA_j}(\mathrm{d}G)\, q_l(\mathbf{x}; G_j)}\,\Big) M(\mathrm{d}\mathbf{x}) \\ \leq & \int_{\mathbb{P}} \int_{\mathcal{X}} h\big[\mathbf{q}(\mathbf{x}; G), \mathbf{q}(\mathbf{x}; G_j)\big] M(\mathrm{d}\mathbf{x})\ \mathcal{P}_{nA_j}(\mathrm{d}G) < \epsilon_2.\end{aligned}$$

Hence, there exists a $\epsilon_3 > 0$ such that $\int_{\mathcal{X}} h\big[\mathbf{q}^{nA_j}(\mathbf{x}), \mathbf{q}(\mathbf{x}; G_0)\big] M(\mathrm{d}\mathbf{x}) > \epsilon_3$.

From (26) it now follows that

$$E(\Lambda_{n+1\,j}^{1/2}) < (1 - \epsilon_3)^n \sqrt{\mathcal{P}(A_j)}.$$

and an application of Markov's inequality leads to

$$\mathrm{P}\Big\{\sum_{j \geq 1} \Lambda_{nj}^{1/2} > \exp(-nb)\Big\} < \exp(nb)(1 - \epsilon_3)^n \sum_{j \geq 1} \sqrt{\mathcal{P}(A_j)}.$$

Therefore, (21) holds for any $b < -\log(1 - \epsilon_3)$ from an application of the Borel-Cantelli's lemma, provided that the following summability condition is satisfied:

$$\sum_{j \geq 1} \sqrt{\mathcal{P}(A_j)} < +\infty. \tag{28}$$

Lemma 3 below shows that $\mathcal{P}$ satisfies condition (28) under the stated hypotheses, and, in turn, it completes the proof of Theorem 1.

Lemma 3 . *Let $H \in \mathbb{P}$ be the prior predictive distribution of $\mathcal{P}$ and assume Condition (ii) of Theorem 1 to hold. Then,* (28) *is verified.*

Proof. The proof goes along the arguments used by Lijoi, Prünster and Walker (2005). Take δ to be any positive number in $(0,1)$ and $(a_n)_{n\geq 1}$ any increasing sequence of positive numbers such that $a_n \to +\infty$. Also, let $a_0 = 0$. Define $C_n = \{\boldsymbol{\beta} : \; |\boldsymbol{\beta}| \leq a_n\}$ and consider the family of subset of $\mathbb{P}$ defined by

$$\mathbb{B}_{a_n,\delta} = \big\{G : \; G(C_n) \geq 1-\delta, \; G(C_{n-1}) < 1-\delta\big\} \tag{29}$$

for each $n \geq 1$. These sets are pairwise disjoint and $\bigcup_n \mathbb{B}_{a_n,\delta} = \mathbb{P}$. For the moment, let us assume that the metric entropy of $\mathbb{B}_{a_n,\delta}$ with respect to the distance d is uniformly bounded in n, that is the number of η-balls in the distance d that covers $\mathbb{B}_{a_n,\delta}$ is finite for any n. Then, summability in (28) is implied by:

$$\sum_{n\geq 1} \sqrt{\mathcal{P}(\mathbb{B}_{a_n,\delta})} < +\infty. \tag{30}$$

In order to prove (30), note that $\mathbb{B}_{a_n,\delta} \subset \{G : \; G(C_{n-1}^c) > \delta'\}$ for some $\delta' > \delta$. An application of Markov's inequality leads to $\mathcal{P}(\mathbb{B}_{a_n,\delta}) \leq (1/\delta')H(C_{n-1}^c)$, hence, (30) is implied by $\sum_{n\geq 1}\sqrt{H(C_{n-1}^c)} < +\infty$. Next, we have that

$$\int_{\mathcal{R}^d} |\boldsymbol{\beta}| H(\mathrm{d}\boldsymbol{\beta}) = \sum_{n\geq 1} \int_{C_{n-1}^c/C_n^c} |\boldsymbol{\beta}| H(\mathrm{d}\boldsymbol{\beta}) \geq \sum_{n\geq 1} a_{n-1}[H(C_{n-1}^c) - H(C_n^c)],$$

by a second application of Markov's inequality, so that Condition (ii) of Theorem 1 assures that $\sum_{n\geq 1} a_{n-1}[H(C_{n-1}^c) - H(C_n^c)] < +\infty$. If we now take $a_n \sim n^2$, it is easy to see that $H(C_n^c) = o(n^{-(2+r)})$ for some $r > 0$. For example,

$$\sum_{n\geq 1}(n-1)^2[H(C_{n-1}^c) - H(C_n^c)] = \sum_{n\geq 1}(2n-1)H(C_n^c).$$

This in turn ensures the convergence of $\sum_{n\geq 1} H(C_{n-1}^c)^\alpha$ for any α such that $(2+r)^{-1} < \alpha < 1$, which includes the case of $\alpha = 1/2$. Condition (30) is then verified.

In order to complete the proof, it remains to show that the metric entropy of $\mathbb{B}_{a_n,\delta}$ with respect to the distance d is uniformly bounded in n. It is actually sufficient to reason in terms of the distance over $\mathbb{P}$ induced by

$$d_j(\mathbf{q}_1, \mathbf{q}_2) = \int_{\mathcal{X}} |q_{1j}(\mathbf{x}) - q_{2j}(\mathbf{x})| M(\mathrm{d}\mathbf{x})$$

for an arbitrary $j \in \Phi$, since $\max_j d_j(\mathbf{q}_1, \mathbf{q}_2) \leq d(\mathbf{q}_1, \mathbf{q}_2) \leq J \max_j d_j(\mathbf{q}_1, \mathbf{q}_2)$. Let $\mathscr{G}$ be a set in $\mathscr{Q}$ and, for $\delta > 0$, denote by $J(\delta, \mathscr{G})$ the metric entropy of $\mathscr{G}$ with respect to d_j, that is the logarithm of the minimum of all k such that there exists $\mathbf{q}_1, \ldots, \mathbf{q}_k \in \mathscr{Q}$ with the property that $\forall \mathbf{q} \in \mathscr{G}$ there exists an i such that $d_j(\mathbf{q},\mathbf{q}_i) < \delta$. The result is then stated as follows: for $\mathscr{G}_{a_n,\delta} = \big\{\mathbf{q}(\mathbf{x};G):\ G \in \mathbb{B}_{a_n,\delta}\big\}$, there exists an $M_\delta < +\infty$ depending only on δ such that, for any n,

$$J(\delta, \mathscr{G}_{a_n,\delta}) < M_\delta \tag{31}$$

The proof of (31) consists in a sequence of three steps.

Step (1). Define $C_a = \{\boldsymbol{\beta} : |\boldsymbol{\beta}| \leq a\}$ and $\mathscr{F}_a = \{\mathbf{q}(\mathbf{x};G):\ G(C_a) = 1\}$. Then

$$J(2\delta, \mathscr{F}_a) \leq \left(\frac{2aK}{\delta} + 1\right)^d \left(1 + \log\frac{1+\delta}{\delta}\right), \tag{32}$$

where K is a constant that depends on the total volume of the space $\mathcal{X}$. It is easy to show that, for any $j \in \Phi$, the kernel $k_j(\mathbf{x},\boldsymbol{\beta})$ is a Lipschitz function in $\boldsymbol{\beta}$ with Lipschitz constant $K_{\mathbf{x}} = \max_{i \leq J}\{|\mathbf{x}_j - \mathbf{x}_i|\}$. Hence,

$$\int_{\mathcal{X}} |k_j(\mathbf{x},\boldsymbol{\beta}_1) - k_j(\mathbf{x},\boldsymbol{\beta}_2)| M(\mathrm{d}\mathbf{x}) \leq K|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2|,$$

where $K = \sup_{\mathbf{x}\in\mathcal{X}} K_{\mathbf{x}} < +\infty$. Given δ, let N be the smallest integer greater than $4aK/\delta$ and cover C_a with a set of balls E_i of radius $2a/N$ so that, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in E_i$, $|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| < 4a/N$. This leads to $\int_{\mathcal{X}} |k_j(\mathbf{x},\boldsymbol{\beta}_1) - k_j(\mathbf{x},\boldsymbol{\beta}_2)| M(\mathrm{d}\mathbf{x}) \leq \delta$. The number of balls necessary to cover C_a is then smaller than N^d. Using arguments similar to those used in Ghosal, Ghosh and Ramamoorthi (1999, Lemma 1), it can be shown that $J(2\delta, \mathscr{F}_a) \leq N^d\big(1 + \log[(1+\delta)/\delta]\big)$, from which (32) follows.

Step (2). Define $\mathscr{F}_{a,\delta} = \{\mathbf{q}(\mathbf{x};G):\ G(C_a) \geq 1-\delta\}$. Then

$$J(\delta, \mathscr{F}_{a,\delta}) \leq K_\delta a^d. \tag{33}$$

for a constant K_δ depending on δ. To see this, take $\mathbf{q}(\mathbf{x};G) \in \mathscr{F}_{a,\delta}$ and denote by G^* the probability measure in $\mathbb{P}$ defined by $G^*(A) = G(A \cap C_a)/G(C_a)$, so that $\mathbf{q}(\mathbf{x};G^*)$ belongs to $\mathscr{F}_a$. It is easy to verify that $d_j\big(\mathbf{q}(\cdot\,;G^*), \mathbf{q}(\cdot\,;G)\big) < 2\delta$. It follows that $J(3\delta, \mathscr{F}_{a,\delta}) \leq J(\delta, \mathscr{F}_a)$ from which (33) follows.

Step (3). We follow here a technique used by Lijoi, Prünster and Walker (2005, Section 3.2). For the sequence $(a_n)_{n\geq 1}$ introduced before, define

$$\mathscr{F}^U_{a_n,\delta} = \{\mathbf{q}(\mathbf{x};G):\ G(C_n) \geq 1-\delta\} \text{ and } \mathscr{F}^L_{a_n,\delta} = \{\mathbf{q}(\mathbf{x};G):\ G(C_n) < 1-\delta\}.$$

By construction, $\mathscr{G}_{a_n,\delta} \subset \mathscr{F}^U_{a_n,\delta}$ and $\mathscr{G}_{a_n,\delta} \subset \mathscr{F}^L_{a_{n-1},\delta}$. Moreover, $\mathscr{F}^L_{a_{n-1},\delta} \downarrow \emptyset$ as n increases to $+\infty$, thus, for any $\eta > 0$, there exists an integer n_0 such that, for any $n \geq n_0$, $J(\eta, \mathscr{F}^L_{a_n,\delta}) \leq J(\eta, \mathscr{F}^U_{a_{n_0},\delta})$. By (33) it follows that

$$J(\eta, \mathscr{G}_{a_n,\delta}) \leq K_\delta a^d_{n_0} \tag{34}$$

for any $n \geq n_0$, but, since $\mathscr{G}_{a_n,\delta} \subset \mathscr{F}^U_{a_n,\delta}$ and $\mathscr{F}^U_{a_n,\delta} \uparrow \mathscr{Q}$, (34) is true also for any $n < n_0$. Result (31) is then verified by setting $M_\delta = K_\delta a^d_{n_0}$. $\square$

References

[1] Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics*, 2, 1152–1174.

[2] Barron, A., Schervish, M.J. and Wasserman, L. (1999). The Consistency of Posterior Distributions in Nonparametric Problems. *Annals of Statistics*, 27, 536–561.

[3] Bhat, C. (1998). Accommodating Variations in Responsiveness to Level-Of-Service Variables in Travel Mode Choice Models. *Transportation Research A*, 32, 455–507.

[4] Brownstone, D. and Train, K. (1999). Forecasting New Product Penetration with Flexible Substition Patterns. *Journal of Econometrics*, 89, 109–129.

[5] Cardell, N. and Dunbar, F. (1980). Measuring the Societal Impacts of Automobile Downsizing. *Transportation Research A*, 14, 423–434.

[6] Choi, T. and Schervish, M.J. (2007). On Posterior Consistency in Nonparametric Regression Problems. *Journal of Multivariate Analysis*, 98, 1969–1987.

[7] Choudhuri, N., Ghosal, S. and Roy, A. (2005). Bayesian Methods for Function Estimation, in *Handbook of Statistics, Vol. 25*, ed. D. Dey, Elsevier, 377–418.

[8] 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.

[9] Erdem, T. (1996). A Dynamic Analysis of Market Structure Based on Panel Data. *Marketing Science*, 15, 359–378.

[10] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1, 209–230.

[11] Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior Consistency of Dirichlet Mixtures in Density Estimation. *Annals of Statistics*, 27, 143–158.

[12] Ghosal, S. and Roy, A. (2006). Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression. *Annals of Statistics*, 34, 2413–2429.

[13] Ghosal, S. and Tang, A. (2006). Bayesian Consistency for Markov Processes. *Sankhyā*, 68, 227–239.

[14] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96, 161–173.

[15] Ishwaran, H. and Zarepour, M. (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchichal Models. *Biometrika*, 87, 371–390.

[16] James, L. F., Lijoi, A. and Prünster, I. (2005). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* in press.

[17] Lijoi, A., Prünster, I. and Walker, S.G. (2005). On Consistency of Nonparametric Normal Mixtures for Bayesian Density Estimation. *Journal of the American Statistical Association*, 100, 1292–1296.

[18] Lo, A. Y. (1984). On a Class of Bayesian Nonparamertic Estimates: I. Density Estimates. *Annals of Statistics*, 12, 351–257.

[19] MacEachern, S. N. and Muller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7, 223–238.

[20] McFadden, D. (1974). Conditional Logit Anaylsis of Qualitative Choice Behavior, in *Frontiers of Econometrics* ed. P. Zarembka, New York, Academic Press, 105–142.

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

[22] Pitman, J. and Yor, M. (1997). The Two–Parameter Poisson–Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability*, 25, 855–900.

[23] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional Results for Means of Random Measures with Independent Increments. *The Annals of Statistics*, 31, 560–585.

[24] Srinivasan, K. and Mahmassani, H. (2005). Dynamic Kernel Logit Model for the Analysis of Longitude Discrete Choice Data: Properties and Computational Assessment. *Transportation Science*, 39, 160–181.

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

[26] Walker, J., Ben-Akiva, M. and Bolduc (2007). Identification of Parameters in Normal Error Component Logit-Mixture (NECLM) Models. *Journal of Applied Econometrics*, 22, 1095–1125.

[27] Walker, S. G. (2003a). On Sufficient Conditions for Bayesian Consistency. *Biometrika* 90, 482–488.

[28] Walker, S. G. (2003b). Bayesian Consistency for a Class of Regression Problems. *South African Statististical Journal* 37, 151–169.

[29] Walker, S. G. (2004). New Approaches to Bayesian Consistency. *Annals of Statistics* 32, 2028–2043.

[30] Walker, S.G., Lijoi, A. and Prünster, I. (2005). Data Tracking and the Understanding of Bayesian Consistency. *Biometrika* 92, 765–778.

[31] Wasserman, L. (1998). Asymptotic Properties of Nonparametric Bayesian Procedures, in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Muller, and D. Sinha, New York, Springer-Verlag, 293–304.

dataset 1 (non-panel case)

			$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$
		True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
GML ($n = 500$)	$P(\{1\} \mid G, \mathbf{x})$	.4980	.3203 (.2907,.3501)	.3201 (.2908,.3509)	.3201 (.2908,.3507)
	$P(\{2\} \mid G, \mathbf{x})$	.0167	.3348 (.3308,.3377)	.3348 (.3307,.3377)	.3348 (.3307,.3377)
	$P(\{3\} \mid G, \mathbf{x})$	.4853	.3449 (.3191,.3715)	.3451 (.3185,.3715)	.3450 (.3186,.3715)
	RMS		.2258	.2258	.2258
		True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
MMNL ($n = 500$)	$P(\{1\} \mid G, \mathbf{x})$	.4980	.4856 (.4748,.4945)	.4857 (.4750,.4946)	.4857 (.4752,.4948)
	$P(\{2\} \mid G, \mathbf{x})$	.0167	.0257 (.0069,.0551)	.0259 (.0073,.0552)	.0258 (.0070,.0553)
	$P(\{3\} \mid G, \mathbf{x})$	.4853	.4886 (.4615,.5057)	.4884 (.4618,.5059)	.4885 (.4609,.5057)
	RMS		.0137	.0137	.0136

dataset 2 (panel case)

			$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$
		True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
GML ($n = 100$, $T_i = 10$)	$P(\{1\} \mid G, \mathbf{x})$	.4939	.4585 (.4476,.4685)	.4585 (.4477,.4684)	.4585 (.4477,.4683)
	$P(\{2\} \mid G, \mathbf{x})$	.0279	.0521 (.0378,.0675)	.0522 (.0379,.0678)	.0524 (.0381,.0679)
	$P(\{3\} \mid G, \mathbf{x})$	.4782	.4894 (.4717,.5061)	.4893 (.4712,.5056)	.4891 (.4710,.5056)
	RMS		.0266	.0266	.0267
		True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
MMNL ($n = 100$, $T_i = 10$)	$P(\{1\} \mid G, \mathbf{x})$	.4939	.4586 (.4495,.4670)	.4597 (.4522,.4675)	.4596 (.4530,.4666)
	$P(\{2\} \mid G, \mathbf{x})$	.0279	.0494 (.0329,.0679)	.0479 (.0309,.0678)	.0471 (.0296,.0669)
	$P(\{3\} \mid G, \mathbf{x})$	.4782	.4920 (.4705,.5107)	.4924 (.4694,.5117)	.4933 (.4706,.5126)
	RMS		.0265	.0257	.0257

Table 1: *Simulation results for dataset 1 (top) and for dataset 2 (bottom) with* $\mathbf{x} = (1.0, -0.9, 1.0, 0.2, 1.0, 0.9)$ *and different* λ *values. The estimates (Est), the credible intervals (C.I.) and the RMS values are presented.*

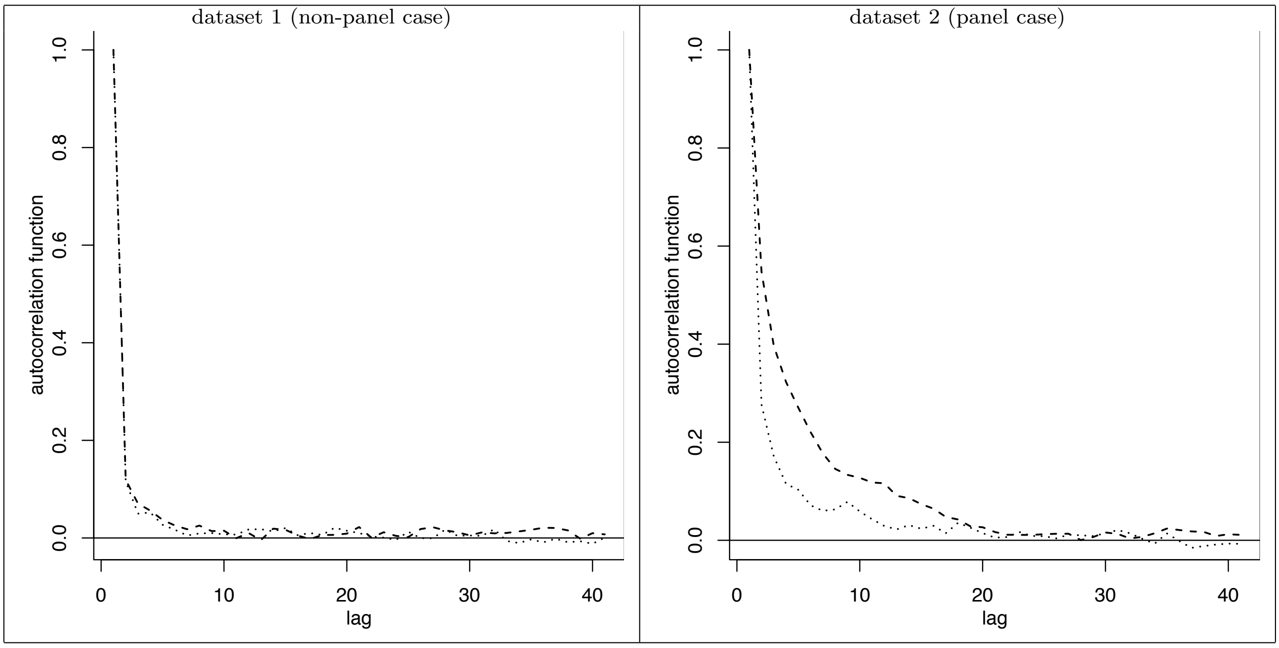

Figure 1: *Autocorrelation functions for the choice probability* $P(\{1\} \mid G, \mathbf{x})$ *for dataset 1 (left) and dataset 2 (right), obtained from the posterior sample of the* $\boldsymbol{\beta}$*'s for the MMNL model with prior hyper parameter* $\lambda = 0.01$ *(dashed) and* $\lambda = 1$ *(dotted).*

dataset 1 (non-panel case)

		$n = 50$	$n = 100$	$n = 500$
	True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
$\mathrm{P}(\{1\} \mid G, \mathbf{x})$	.4980	.4927 (.3739,.5685)	.5145 (.4674,.5421)	.4856 (.4748,.4945)
$\mathrm{P}(\{2\} \mid G, \mathbf{x})$	.0167	.1046 (.0180,.2384)	.0489 (.0067,.1242)	.0257 (.0069,.0551)
$\mathrm{P}(\{3\} \mid G, \mathbf{x})$	.4853	.4027 (.2970,.4794)	.4366 (.3683,.4767)	.4886 (.4615,.5057)
RMS		.0867	.0440	.0137

dataset 2 (panel case)

		$n = 10, T_i = 10$	$n = 50, T_i = 10$	$n = 100, T_i = 10$
	True	Est (95% C. I.)	Est (95% C. I.)	Est (95% C. I.)
$\mathrm{P}(\{1\} \mid G, \mathbf{x})$	.4939	.5956 (.5491,.6273)	.4176 (.4018,.4308)	.4586 (.4495,.4670)
$\mathrm{P}(\{2\} \mid G, \mathbf{x})$	.0279	.0527 (.0125,.1042)	.0562 (.0359,.0807)	.0494 (.0329,.0679)
$\mathrm{P}(\{3\} \mid G, \mathbf{x})$	.4782	.3517 (.2892,.3953)	.5261 (.4973,.5492)	.4920 (.4705,.5107)
RMS		.0977	.0556	.0265

Table 2: *MMNL model: simulation results for dataset 1 (top) and for dataset 2 (bottom) with* $\mathbf{x} = (1.0, -0.9, 1.0, 0.2, 1.0, 0.9)$ *and different number of observations. The estimates (Est), the credible intervals (C.I.) and the RMS values are presented.*

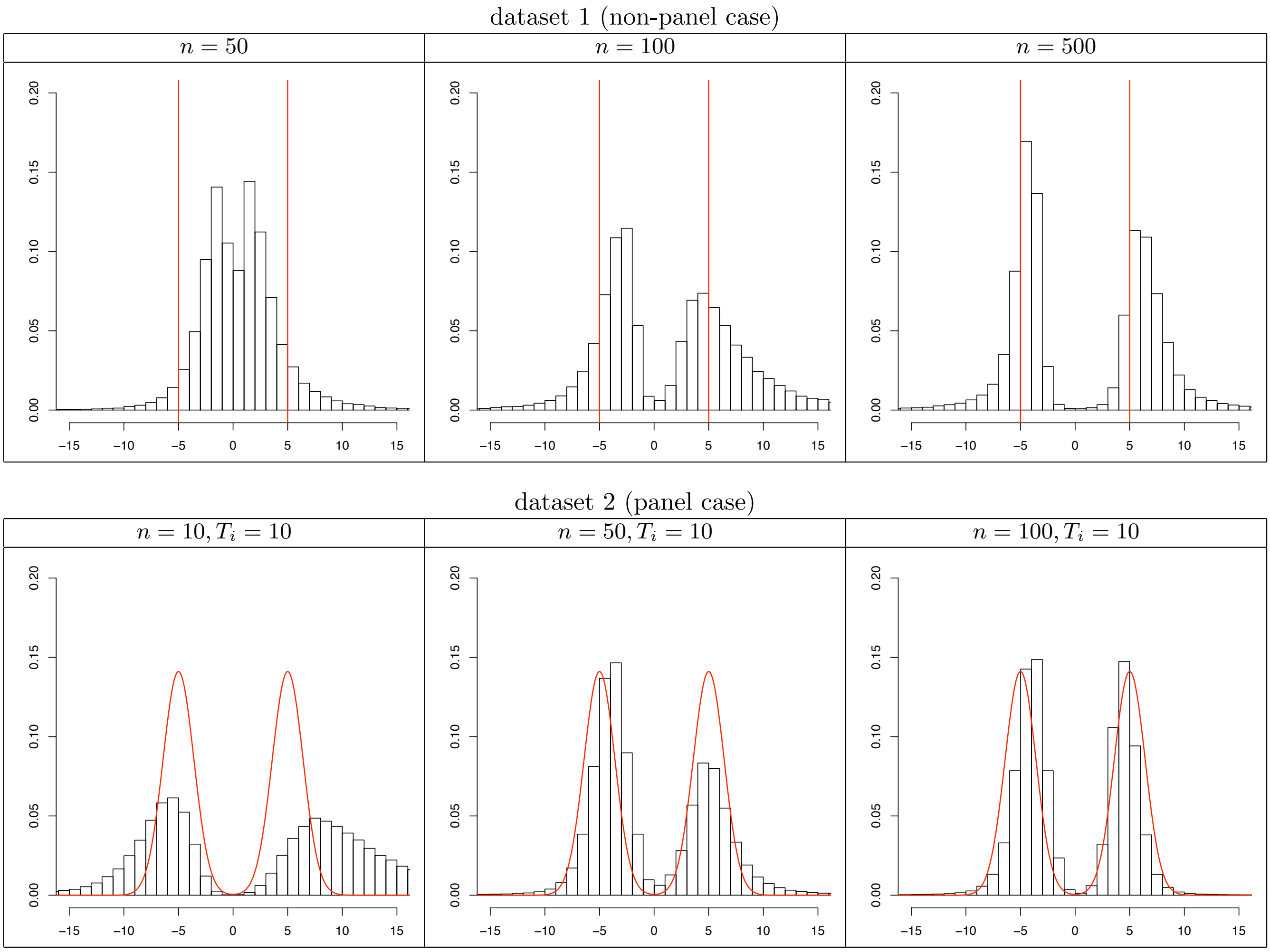

Figure 2: *MMNL model: histogram of sampled* β_1*'s from their posterior distribution for dataset 1 (top) and for dataset 2 (bottom) with* $\mathbf{x} = (1.0, -0.9, 1.0, 0.2, 1.0, 0.9)$ *and different number of observations. The solid lines represent the true mixing distribution.*