Journal of Econometrics 147 (2008) 232–246

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A Bayesian mixed logit–probit model for multinomial choice☆

Martin Burda[a,*], Matthew Harding[b,1], Jerry Hausman[c]

[a] Department of Economics, University of Toronto, Sidney Smith Hall, 100 St. George Street, Toronto, ON M5S 3G7, Canada

[b] Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305, United States

[c] Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142, United States

ARTICLE INFO

Article history:
Available online 25 September 2008

JEL classification:
C11
C13
C14
C15
C23
C25

Keywords:
Multinomial discrete choice model
Dirichlet process prior
Non-conjugate priors
Hierarchical latent class models

ABSTRACT

In this paper, we introduce a new flexible mixed model for multinomial discrete choice where the key individual- and alternative-specific parameters of interest are allowed to follow an assumption-free nonparametric density specification, while other alternative-specific coefficients are assumed to be drawn from a multivariate Normal distribution, which eliminates the independence of irrelevant alternatives assumption at the individual level. A hierarchical specification of our model allows us to break down a complex data structure into a set of submodels with the desired features that are naturally assembled in the original system. We estimate the model, using a Bayesian Markov Chain Monte Carlo technique with a multivariate Dirichlet Process (DP) prior on the coefficients with nonparametrically estimated density. We employ a "latent class" sampling algorithm, which is applicable to a general class of models, including non-conjugate DP base priors. The model is applied to supermarket choices of a panel of Houston households whose shopping behavior was observed over a 24-month period in years 2004–2005. We estimate the nonparametric density of two key variables of interest: the price of a basket of goods based on scanner data, and driving distance to the supermarket based on their respective locations. Our semi-parametric approach allows us to identify a complex multi-modal preference distribution, which distinguishes between inframarginal consumers and consumers who strongly value either lower prices or shopping convenience.

1. Introduction

Discrete choice models are widely used in economics and the social sciences to analyze choices made by individuals among a set of alternatives. In this paper, we introduce a new flexible mixed model for multinomial discrete choice, where the key individual- and alternative-specific parameters of interest are allowed to follow an assumption-free nonparametric density specification, while other alternative-specific coefficients are assumed to be drawn from a multivariate Normal distribution.

Two advantages arise from this specification. First, we do not require the correct a priori specification of the distribution of taste parameters. Independent Normal distributions have typically been used, but Harding and Hausman (2007) demonstrate that this choice can lead to biased estimates in the presence of correlations between tastes for product attributes. Second, the use of choice specific coefficients drawn from a multivariate distribution eliminates the independence of irrelevant alternatives (IIA) that holds at the individual level in almost random coefficients logit models. Hausman and Wise (1978) employed a multivariate Probit model without the IIA assumption. However, the model proved difficult to estimate in the likelihood context, if the number of choices exceeded four.

A hierarchical specification of our model allows us to break down a complex data structure into a set of submodels with the desired features that are naturally assembled in the original system. Such model structure is directly amenable to estimation by Bayesian methods on Gibbs sampling, utilizing recent advances in Markov Chain Monte Carlo (MCMC) techniques. Estimation of the nonparametric density of a subset of the model coefficients is facilitated by specifying a multivariate Dirichlet Process (DP) prior on these coefficients.

Much of the work on nonparametric Bayesian modeling traces its origins to the seminal papers of Freedman (1963), Ferguson (1973a,b), and Blackwell and MacQueen (1973), though applications were quite limited until the late 1980s. Fueled by advances

☆ We would like to thank Peter Rossi and participants of the Harvard applied statistics workshop, and seminars at MIT, Stanford, USC, and Georgetown for their insightful comments and suggestions. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: http://www.sharcnet.ca).

* Corresponding author. Tel.: +1 416 978 4479.
E-mail addresses: martin.burda@utoronto.ca (M. Burda), mch@stanford.edu (M. Harding), jhausman@mit.edu (J. Hausman).
[1] Tel.: +1 650 723 4116; fax: +1 650 725 5702.

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.09.029

in computation, the last two decades witnessed an explosion of interest in nonparametric Bayesian models (for a recent review see e.g. Müller and Quintana, 2004).

Dirichlet Process Mixture models (DPM) (Escobar and West, 1995) form an important part of this literature. Some recent applications of univariate DPMs include epidemiology (Dunson, 2005), genetics (Medvedovic and Sivaganesan, 2002), medicine (Kottas et al., 2002), finance (Kacperczyk et al., 2003), and stochastic volatility modeling (Jensen and Maheu, 2007). In the microeconometric literature, the univariate DPM has been used in several studies. Hirano (2002) estimated a Bayesian random effects autoregressive model, with nonparametric idiosyncratic shocks. Conley et al. (2008) model the joint distribution of the error terms in a instrumental variable problem using a DPM which improves efficiency when errors are non-Normal. Chib and Hamilton (2002) analyzed the effect of a binary treatment variable on a continuous outcome in a panel data model with treatment- and outcome-specific individual random effects without distributional assumptions. Jochmann and León-González (2004) estimated the demand for health care with panel data with nonparametric random effects under the DP prior.

A multinomial discrete choice model with Bayesian estimation strategy has been analyzed recently by Athey and Imbens (2007). These authors allow for unobserved and observed individual and alternative-specific characteristics, in a fully parametric model with the key parameters of interest drawn from a multivariate Normal distribution. Although we focus on observed characteristics only, our model setup can be readily extended to include unobserved characteristics as well.

In Bayesian analysis, significant technical simplifications result from choosing a prior family of density functions that, after multiplication by the likelihood, produce a posterior distribution of the same family — a so-called conjugate prior. In such cases, only the parameters of the prior change to form the posterior with accumulation of data, not its mathematical form. Implementational simplifications also result in the case where the base DP prior is conjugate to the base distribution but such models are necessarily limited in their application. In contrast, the non-conjugate case is more involved in terms of model specification and estimation strategy, but can be applied to essentially any model with an arbitrary DP base prior. All the literature cited above have confined themselves to the relatively simple conjugate scenario.

In this paper, we venture into the realm of general (non-conjugate) models, and in one of our Gibbs blocks we employ an algorithm for non-conjugate DP priors developed recently by Neal (2000). Non-conjugate sampling methods are currently subject to active research in the statistics and machine learning literature and are only slowly spilling over to other areas (Dahl, 2005; Jain and Neal, 2007; Dahl et al., 2008).

Another important feature of our model is its hierarchical structure with respect to parameters. Hierarchical models provide a natural environment for the application of DP priors. Existing economic models of discrete choice allow for a more flexible specification, by using finite mixture models (Imai and van Dyk, 2005; Rossi et al., 2005).

Finally, all previous studies utilizing the DPM cited above estimated nonparametrically parameter densities along a single dimension, leaving out the multivariate case for a theoretical discussion. In contrast, in our paper, we implement the full multivariate DPM case allowing for arbitrary correlation among parameters of interest drawn from nonparametric densities. In our simulation and empirical studies, we consider the bivariate case for ease of graphical presentation, but given the estimation mechanism, higher dimensionality is easily accommodated by simply increasing the matrix sizes of the model parameters.

2. Dirichlet process mixture model

2.1. Parametric vs. nonparametric bayesian modelling

Econometric models are often specified by a distribution $F(\cdot;\psi)$, with associated density $f(\cdot;\psi)$, known up to a set of parameters $\psi \in \Psi \subset \mathbb{R}^d$. Under the Bayesian paradigm, ψ are treated as random variables, which implies further specification of their respective probability distribution.

Consider a sequence $z = \{z_i\}_{i=1}^n$ of realizations of a set of exchangeable random variables $Z = \{Z_i\}_{i=1}^n$, defined over a measurable space $(\Phi, \mathcal{D})$ where $\mathcal{D}$ is a σ-field of subsets of Φ. In a parametric Bayesian model, the joint distribution of z and the parameters is defined as

$$Q(\cdot;\psi, G_0) \propto F(\cdot;\psi)G_0$$

where G_0 is the (so-called prior) distribution of the parameters over a measurable space $(\Psi, \mathcal{B})$, with $\mathcal{B}$ being a σ-field of subsets of Ψ. Conditioning on the data turns $F(\cdot;\psi)$ into the likelihood function $L(\psi|\cdot)$ and $Q(\cdot;\psi, G_0)$ into the posterior density $K(\psi|G_0, \cdot)$.

In the class of nonparametric Bayesian models[2] considered here, the joint distribution of data and parameters is defined as a mixture

$$Q(\cdot;\psi, G) \propto \int F(\cdot;\psi)G(\mathrm{d}\psi)$$

where G is the mixing distribution over ψ. It is useful to think of $G(\mathrm{d}\psi)$ as the conditional distribution of ψ, given G. The distribution of the parameters, G, is now random which leads to a complete flexibility of the resulting mixture. The model parameters ψ are no longer restricted to follow any given prespecified distribution as was stipulated by G_0 in the parametric case. The parameter space now also includes the random infinite-dimensional G, with the additional need for a prior distribution for G. The Dirichlet Process prior is a popular alternative due to its numerous desirable properties; we proceed with its description in the next section.

2.2. Dirichlet process prior

In a seminal paper, Ferguson (1973a) introduced the Dirichlet process (DP) prior for random measures whose support is large enough to span the space of probability distribution functions and that leads to analytically manageable posterior distributions. Antoniak (1974) further elaborated on using the DP as the prior for the mixing proportions of a simple distribution.

A DP prior for G is determined by two parameters: a distribution G_0 that defines the "location" of the DP prior, and a positive scalar precision parameter α. The distribution G_0 may be viewed as a baseline prior that would be used in a typical parametric analysis. The flexibility of the DP prior model environment stems from allowing G – the actual prior on the model parameters – to stochastically deviate from G_0. The precision parameter α determines the concentration of the prior for G around the DP prior location G_0 and thus measures the strength of belief in G_0. For large values of α, a sampled G is very likely to be close to G_0, and vice versa.

More specifically, let $\mathcal{M}(\Psi)$ be a collection of all probability measures on Ψ endowed with the topology of weak convergence. The space $\mathcal{M}(\mathcal{M}(\Psi))$ is then the collection of all probability measures (i.e. priors) on $\mathcal{M}(\Psi)$ together with the topology of weak convergence derived from $\mathcal{M}(\Psi)$. Let $G_0 \in \mathcal{M}(\Psi)$ and let α be

[2] A commonly used technical definition of nonparametric Bayesian models are probability models with infinitely many parameters (Bernardo and Smith, 1994).

a positive real number. Following Ferguson (1973a), a *Dirichlet Process* on $(\Psi, \mathcal{B})$ with a base measure G_0 and a concentration parameter α, denoted by $\mathrm{DP}(G_0, \alpha) \in \mathcal{M}(\mathcal{M}(\Psi))$, is a distribution of a random probability measure $G \in \mathcal{M}(\Psi)$ over $(\Psi, \mathcal{B})$ such that, for any finite measurable partition $\{\Psi_i\}_{i=1}^{J}$ of the sample space Φ, the random vector $(G(\Psi_1), \ldots, G(\Psi_J))$ is distributed as $(G(\Psi_1), \ldots, G(\Psi_J)) \sim \mathrm{Dir}(\alpha G_0(\Psi_1), \ldots, \alpha G_0(\Psi_J))$ where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution. We write $G \sim \mathrm{DP}(G_0, \alpha)$, if G is distributed according to the Dirichlet process $\mathrm{DP}(G_0, \alpha)$. A Bayesian model with such a feature is commonly referred to as a Dirichlet Process Mixture (DPM) model.[3] Since realizations of the DP are discrete with probability one, a DPM can be viewed as a countably infinite mixture (Ferguson, 1983).

Having specified a flexible nonparametric prior, the subsequent estimation method crucially depends on whether the likelihood $L(\psi|\cdot)$ and the DP base prior is a conjugate pair. In general terms, a family of prior probability distributions is said to be conjugate to a family of likelihood functions if the resulting posterior distributions are in the same family as the prior distributions. The conjugate case is typically much easier to handle, since only the parameters of the prior change to create the posterior with accumulation of data, not the mathematical form of the prior. However, the class of likelihood functions that can be specified for such cases is arguably quite limited, as these need to adhere to the class of the prior. The exponential family of functions is a typical example. Since we consider the non-conjugate scenario, a brief technical description of the conjugate case has been relegated into the Appendix. In contrast, the non-conjugate case is usually more involved in terms of estimation methodology, but can be applied to essentially any DP base prior and likelihood specification. The resulting estimation strategy undertaken in this paper is thus applicable to a general class of Bayesian hierarchical models.

Sampling strategies for non-conjugate DP priors are currently an active research field. We utilize the methodology proposed recently by Neal (2000), which builds on MacEachern and Müller (1998), due to its superior efficiency properties. Other methods include Walker and Damien (1998), Green and Richardson (2001), Dahl (2005), Jain and Neal (2007), and Dahl et al. (2008). However, these methods are considerably more complex; it is not clear whether the additional benefit in terms of Markov chain convergence speed would justify their implementation for the present purpose.

In the approach suggested by Neal (2000), the key to dealing with the non-conjugacy of G_0 and L, is to bypass the need for integrating out G_0 in the first place. The DPM is obtained as a limiting case of a random "latent class" finite mixture model as the number of stochastic mixture components approaches infinity. The object that is being integrated over are the mixing proportions of these latent classes. Specifically, suppose $z = \{z_i\}_{i=1}^{n}$ are drawn independently from some unknown distribution. We can model such a distribution as a mixture of simple distributions such that

$$P(z) = \sum_{c=1}^{C} p_c f(z|\gamma_c). \tag{2.1}$$

Here, p_c are the mixing proportions, and f is a class of distributions. If we assume that the number of mixing components, C, is finite, then a typical prior for p_c is the symmetric Dirichlet distribution, $\mathrm{Dir}(\alpha/C, \ldots, \alpha/C)$, where

$$P(p_1, \ldots, p_C) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/C)^C} \prod_{c=1}^{C} p_c^{(\gamma/C)-1}$$

with $p_c \geq 0$ and $\sum p_c = 1$. The parameters γ_c are assumed to be independent with the prior distribution G_0. Using mixture identifiers c_i, the model (2.1) can be represented as follows (Neal, 2000):

$$\begin{aligned} &z_i|c_i, \boldsymbol{\gamma} \sim F(\cdot; \gamma_{c_i}) \\ &c_i|p_1, \ldots, p_C \sim \mathrm{Discrete}(p_1, \ldots, p_C) \\ &p_1, \ldots, p_C \sim \mathrm{Dir}(\alpha/C, \ldots, \alpha/C) \\ &\gamma_c \sim G_0 \end{aligned} \tag{2.2}$$

where c_i indicates which latent class is associated with z_i. For each class c, the parameters γ_c determine the distribution of observations from that class. The collection of all such γ_c is denoted by $\boldsymbol{\gamma}$. By integrating over the Dirichlet prior, the mixing proportions p_c can be eliminated to obtain the following conditional distribution for c_i:

$$P(c_i = c|c_1, \ldots, c_{i-1}) = \frac{n_{i,c} + \alpha/C}{i - 1 + \alpha} \tag{2.3}$$

where $n_{i,c}$ is the number of c_j for $j \neq i$ that are equal to c. When C goes to infinity, the conditional probabilities (2.3) reach the following limits:

$$\begin{aligned} P(c_i = c|c_1, \ldots, c_{i-1}) &\to \frac{n_{i,c}}{i - 1 + \alpha} \\ P(c_i \neq c_j|c_1, \ldots, c_{i-1}) &\to \frac{\alpha}{i - 1 + \alpha}. \end{aligned}$$

As a result, the conditional probability for ψ_i, where $\psi_i = \gamma_{c_i}$, becomes

$$\psi_i|\psi_1, \ldots, \psi_{i-1} \sim \frac{1}{i - 1 + \alpha} \sum_{j<i} \delta(\psi_j) + \frac{\alpha}{i - 1 + \alpha} G_0 \tag{2.4}$$

which is equivalent to the conditional probabilities (A.1) implied by the DPM. In other words, the limit of the finite mixture model (2.2) is equivalent to the DPM model, as the number of mixture components C goes to infinity. G is the distribution over ψ and has the DP prior. The parameter α of the DP prior controls the number of components in the mixture, such that a larger α results in a larger number of components. In contrast to the conjugate case, a different object is being integrated out than in (A.1), bypassing the need for conjugacy between the base DP prior and the likelihood function.

The resulting estimation procedure with the DP prior can be embedded in a wider model, and applied only to its well-defined submodel, while other procedures can be used for the remainder. We take advantage of this feature in our semiparametric model specification, by handling the nonparametric model component with the DP prior, while the parametric alternative-specific indicator variable block of parameters is estimated with a standard MCMC Gibbs sampling procedure.

2.3. Estimation strategy for non-conjugate dirichlet process prior

2.3.1. The Chinese restaurant process

In order to develop some intuition behind the estimation mechanism of the Gibbs sampling procedure based on (2.2) and (2.3), we will briefly describe the heuristics of the popular "Chinese restaurant" process that is often used to describe the behavior of estimating algorithms for models with DP priors. In general, each latent class in (2.2) can be thought of as a table in a Chinese restaurant. The table location and size represent a current draw of the parameters of interest. Consider a snapshot of time in the life of the restaurant when some tables (or clusters) have attracted many customers, while other tables may by occupied by one customer only (so-called "singletons"). At each small discrete time period,

[3] A specific subset of the literature, e.g. Antoniak (1974) and MacEachern and Müller (1998) refer to these models as "Mixture of Dirichlet Process models" (MDP).

one customer decides to either stay at his current table or move to another table that would suit him "better". This may involve the restaurant setting up new tables at customer requests, or taking away tables that have been completely vacated. After each customer has made their decision, a new state of the system is recorded by the restaurant management to make inferences about the true underlying distribution of customers' tastes. The whole customer moving decision process starts anew in the next Monte Carlo (MC) iteration. As a stylized fact, tables with more customers yield higher probability of attracting additional customers and vice versa, resulting in the clustering property of the Chinese restaurant social scene. We will keep referring to this heuristic analogy throughout the description of the estimation algorithm, to guide our intuition.

2.3.2. Estimation algorithm

Before formally stating the estimation procedure, we will describe its heuristics in general terms. The estimation algorithm is composed of two basic steps in each MC iteration:

(1) Given the state of the system, update the assignment of z_i to the latent classes c_i. New classes can be created and existing classes can vanish; the cluster structure is endogenous to the data and the likelihood function, rendering the estimation procedure nonparametric. This step is tantamount to customers switching tables in the Chinese restaurant process.
(2) For each latent class, draw new values of parameters γ_{c_i} using a Metropolis–Hastings update. In the Chinese restaurant analogy, this step enables the management to make inference about the underlying distribution of customers' tastes.

Step (1) is composed of two stages: first, the entire parameter space is being examined with positive probability for suggestions of creation of potential new classes, labeled as c_i^*. These suggestions are drawn from the base distribution G_0. Those z_i that are not "singletons", i.e. share a latent class with other $z_j, j \neq i$, change their latent class membership c_i for the newly created c_i^*, with a probability proportional to the ratio $L(\gamma_{c_i^*}|z_i)/L(\gamma_{c_i}|z_i)$ where $L(\gamma_{c_i}|z_i)$ is the likelihood of z_i being distributed as $F(z_i;\gamma_{c_i})$. Singletons z_i, on the other hand, are re-distributed among the existing latent classes with a probability proportional to their respective likelihood ratios. The analogy here is the Chinese restaurant management offering to set up tables with new menus, aiming to rescue customers from fixation on unlikely choices. This part in itself would be sufficient to produce a Markov Chain that is ergodic, i.e. convergent (in a sense) with respect to the stationary target posterior distribution. The resulting chain would sample inefficiently, though, since it can move z_i from one existing class to another, by passing through a possibly unlikely state of z_i being a singleton. Therefore, in the second stage of Step (1), partial Gibbs updates are applied only to those observations that are not singletons, and which are now allowed to change c_i directly for another existing latent class, generically denoted by c, with probability proportional to the likelihood $L(\gamma_c|z_i)$. As a result, the mixing properties of the chain improve substantially. Having (potentially) switched around the membership of observations among latent classes, in Step 2 the parameters of each latent class are updated.

The combination of these latent class densities changes in every MC step and the entire MC chain combines into the resulting stable nonparametric form of the density of ψ. Endogeneity of the number and form of these cluster-specific densities in each MC step leads to the ability of the convolution, to approximate any form of the true density of ψ to arbitrary accuracy that depends only on the number of MC draws, conditional on the dataset and model specification.

The full form of the Algorithm 7 (Neal, 2000) is as follows:

Let the state of the Markov chain consist of $\mathbf{c} = (c_1, \ldots, c_n)$ and $\boldsymbol{\gamma} = (\gamma_c : c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$, update c_i as follows: If c_i is not a singleton (i.e. $c_i = c_j$ for some $j \neq i$), let c_i^* be a newly created component, with γ_{c^*} drawn from G_0. Set the new c_i to this c_i^* with probability

$$a(c_i^*, c_i) = \min\left[1, \frac{\alpha}{n-1}\frac{L(\gamma_{c_i^*}|z_i)}{L(\gamma_{c_i}|z_i)}\right].$$

  Otherwise, when c_i is a singleton, draw c_i^* from c_{-i}, choosing $c_i^* = c$ with probability $n_{-i,c}/(n-1)$. Set the new c_i to this c_i^* with probability

$$a(c_i^*, c_i) = \min\left[1, \frac{n-1}{\alpha}\frac{L(\gamma_{c_i^*}|z_i)}{L(\gamma_{c_i}|z_i)}\right].$$

  If the new c_i is not set to c_i^*, it is the same as the old c_i.
- For $i = 1, \ldots, n$: If c_i is a singleton (i.e. $c_i \neq c_j$ for all $j \neq i$), do nothing. Otherwise, choose a new value for c_i from $\{c_1, \ldots, c_n\}$ using the following probabilities:

$$P(c_i = c|c_{-i}, y_i, \boldsymbol{\gamma}, c_i \in \{c_1, \ldots, c_n\}) = b\frac{n_{-i,c}}{n-1}L(\gamma_c|z_i)$$

  where b is the appropriate normalizing constant.
- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\gamma_c|z_i$, such that $c_i = c$, or perform some other update to γ_c that leaves this distribution invariant.

2.4. Example of multimodal density estimation

Owing to its generality that is not constrained by the requirements of conjugacy, this estimation approach can be applied to any model scenario for sampling posterior distributions of parameters of interest - univariate or multivariate. Arguably the simplest such scenario arises when the "parameter" is taken as the random variable itself in nonparametric density estimation. Before discussing our limited dependent variable model, we took the estimation strategy described above to the test of estimating highly irregular densities formulated in Marron and Wand (1992). One example is shown in Fig. 1 which is given by $0.5\ N(0,1) + \sum_{l=0}^{4} 0.1N(l/2 - 1, 0.001)$. The chosen distribution is a mixture of Normals, but as we shall see it is not the aim of this procedure to estimate the parameters of the mixture. Our procedure works rather differently, as we shall show below.

In Fig. 1 we show the "target" true density from which we draw a sample of $N = 1000$ observations. In Fig. 2 we plot the DPM density estimated as a result of our procedure, which provides a good approximation to a very difficult problem. With this opportunity, we can discuss some of the properties of this method which might be immediately apparent from our description of the estimation algorithm.

First, it is important to note that the procedure shares features with both estimation by mixtures of Normals and kernel estimation, based on the Normal kernel (Ferguson, 1983). At every step of the Markov chain, the procedure partitions the observations into n (or fewer) latent classes, which is equivalent to fitting a Normal mixture with n (or fewer) components. One such typical configuration of the mixture is shown in the right panel of Fig. 2. The aim of the procedure is not to obtain a final "optimal" configuration, but rather to let the mixtures vary over repeated Monte Carlo draws. In the left panel of Fig. 3, we show the evolution of the number of latent classes over the MC chain. The number of classes varies between 6 and 19 over repeated draws and at each step a different mixture is computed with a different number of components and corresponding parameters. Recall that in Section 2 we characterized a nonparametric Bayesian method,

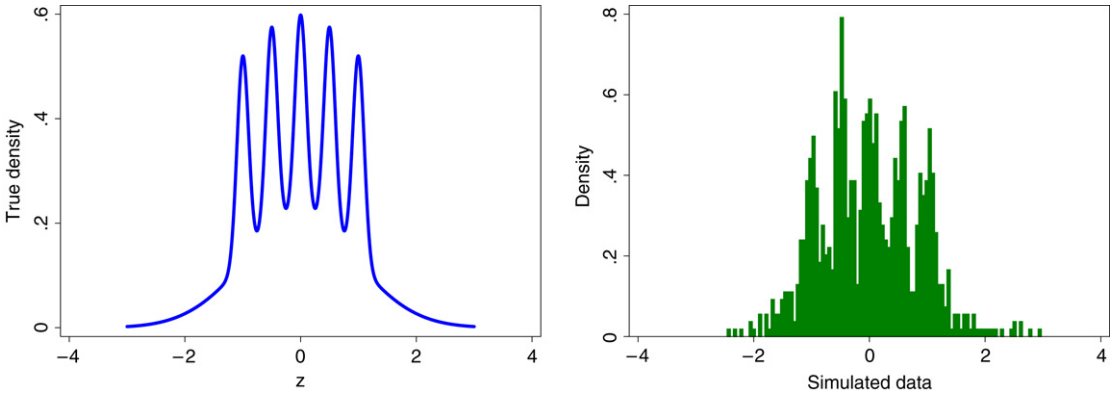

Fig. 1. Left: Trial true functional form of "the claw" posterior density of Marron and Wand (1992). Right: Histogram of a sample draw, $N = 1000$.

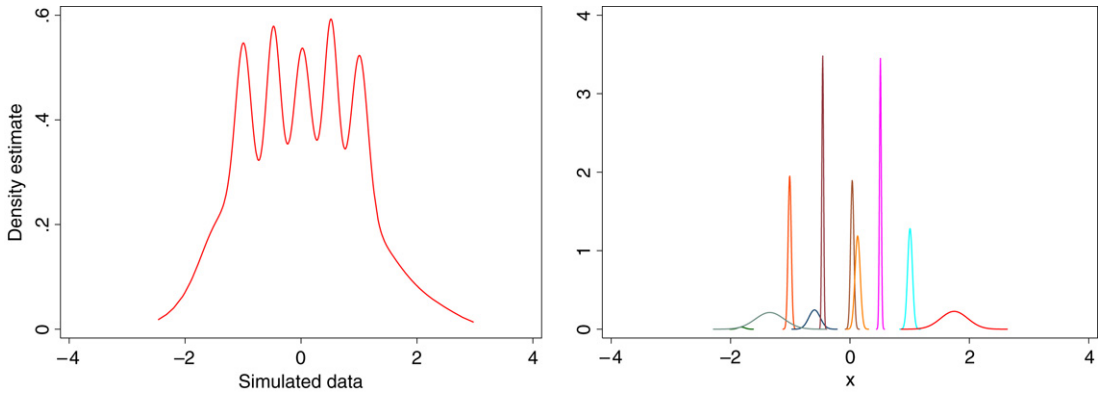

Fig. 2. Left: DPM density estimate based on the sample in Fig. 1, with 10,000 MC steps. Right: A typical snapshot of latent class positions scaled by the class membership intensity.

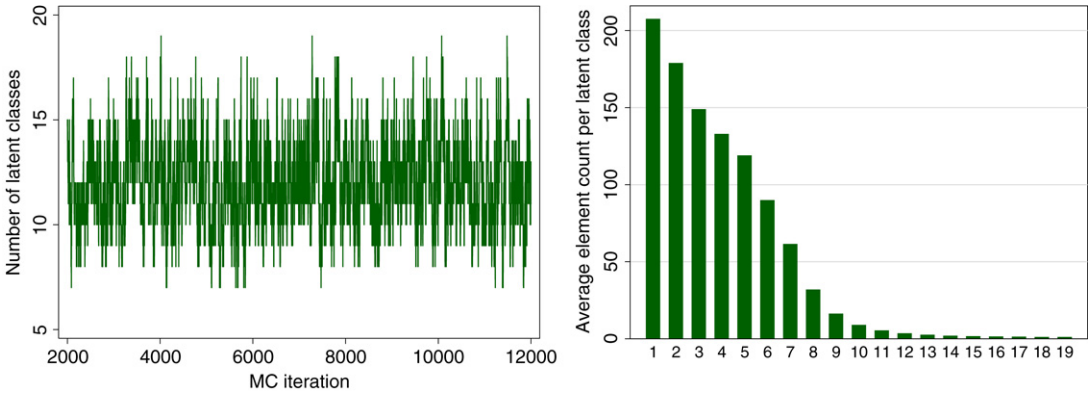

Fig. 3. $\alpha = 1$. Left: Evolution of the number of latent classes over the MC chain. Right: Average number of latent class members, sorted by size.

as one which integrates over a range of prior distributions using a distribution G over these prior distributions. We modeled this distribution G by the Dirichlet Process. Thus, in order to obtain the posterior distribution in Fig. 2 we average over the resulting mixture distributions with an additional component of G_0 with a weight α/n. Moreover, each configuration of the latent classes depends on earlier draws by virtue of the Markov chain design.

One very important feature of the procedure becomes important at this point. The number of latent classes stays small and bounded over repeated MC draws. Moreover, the right panel in Fig. 3 shows the distribution of members of the latent classes. This distribution decays very fast, and most of the observations are allocated between a few classes. This is due to two forces inherent in the construction of the Dirichlet Process. Notice from the algorithm description that the probability of allocating an observation to an existing cluster is proportional to the size of the cluster. This implies that new observations are strongly attracted by large existing clusters, and are much less likely to start new clusters of their own. This property is often referred to as preferential attachment or "the rich get richer property". Recently, alternative prior process specifications, such as the uniform process or the Pitman–Yor process have been proposed which do not have this clustering feature (Dicker and Jensen, 2008).

The clustering property is also controlled by the parameter α. On one hand it measures the weight placed on the prior base distribution G_0. Large values of α correspond to more weight being placed on the prior base distribution G_0, while small values give more weight to the empirical observations. In the context of density estimation, Ferguson (1983) shows that in the limit for $\alpha = 0$, the method fits the parametric density estimate under

the functional form given by G_0. The parameter α also controls the relative decay in class membership as one moves between classes. A small value of α corresponds to more observations being clustered in each of the first few classes. In the limit, this corresponds to all observations being attributed to a single class. As α increases, there will be many classes with few members and the class membership decays only very slowly.

To illustrate this property, let us compare the distribution of class membership in the estimation of the "claw" density under two different choices of $\alpha = 1$ and $\alpha = 10$ in Figs. 3 and 4. We can see that, as we increase α, the number of latent classes utilized also increases to between 25–65 classes. Moreover, class membership decays much slower and we have a large number of classes with only a few members.

Furthermore, it is possible to show that, in the limit, as $\alpha \to \infty$ this procedure allocates one individual per class. The distribution becomes a mixture of n Normal distributions, where each distribution is the Bayesian density estimate based on a single observation with prior G_0. Ferguson (1983) shows that this yields a variable kernel estimator with constant window size but centered at a point between the observation and the prior hypothesized mean. Thus, even in the limiting kernel case the procedure maintains a certain degree of shrinkage towards the prior.

3. Semi-parametric Bayesian logit–probit model

3.1. Model environment

There are $i = 1, \ldots, N$ individuals and an (unordered) set $\{1, \ldots, J\}$ of alternative choices, indexed by j, that each individual is facing. During a time period t, each individual chooses one or more alternatives. The occasions on which an individual i made a choice during time t are indexed by q. These choice occasions total to $Q_{it} \geq 1$. Each alternative j has associated with it a K-dimensional column vector X_{itqj} of observed attributes (these may or may not be constant over q). Let $B = \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it}$. Consider the random utility model

$$U_{itqj} = g\left(X_{itqj}, \beta_i, \theta_i\right) + \varepsilon_{itqj}$$

where ε_{ijt} is iid extreme value type I and U_{itqj} denotes the (unobserved) utility for an individual i associated with choice j on occasion q during time t. Furthermore, $\beta = (\beta_1, \ldots, \beta_N)'$, $\theta = (\theta_1, \ldots, \theta_N)'$ are vectors of unknown coefficients. The distribution of β_i is modeled nonparametrically, while θ_i – coefficients on alternative-specific indicator variables – are assumed to follow a multivariate Normal distribution. We will further assume in the model implementation that

$$g\left(X_{itqj}, \beta_i, \theta_i\right) = X'_{1itqj}\beta_i + X'_{2j}\theta_i$$

where $X_{itqj} = (X_{1itqj}, X_{2j})$. Since our estimation methodology is applicable to any nonlinear parametrization of $g(\cdot)$, we will preserve the generic notation in this section.

The inclusion of these choice-specific random Normal variables forms the "probit" element of the model. We introduce this extension of the standard logit model, in order to eliminate the IIA assumption at the individual level. In typical random coefficients logit models used to date, for a given individual the IIA property still holds since the error term is independent extreme value. With the inclusion of choice specific correlated random variables, the IIA property no longer holds, since a given individual who has a positive realization for one choice is more likely to have a positive realization for another positively correlated choice specific variable. Choices are no longer independent conditional on attributes and hence the IIA property no longer holds. Thus, the logit part of the model allows for ease of computation while the probit part of the model allows an unrestricted covariance matrix of the stochastic terms in the choice specification.

An individual chooses the alternative j, if the associated utility U_{itqj} is higher than that associated with any of the alternatives. Let $y_{itqj} \in \{1, \ldots, J\}$ denote the observed choice outcome. For the logistic specification of ε_{ijt}, the probability of such choice is given by

$$P(y_{itqj} = j) = \frac{\exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)}{\sum_{j=1}^{J} \exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)}. \tag{3.1}$$

(see e.g. Train, 2003). The probability of an individual i choosing a set $\left\{y_{itqj} = j\right\}_{q=1}^{Q_{it}}$ at time t can be expressed by the *iid* property of ε_{itqj} as

$$\prod_{q=1}^{Q_{it}} P(y_{itqj} = j). \tag{3.2}$$

Using (3.1), (3.2) and the *iid* property of ε_{itqj}, the joint probability of observing the complete set of y_{itqj} is

$$\begin{aligned} P(y, X|\beta, \theta) &= \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{q=1}^{Q_{it}} \prod_{j=1}^{J} P(y_{itqj} = j)^{y_{itqj}} \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{q=1}^{Q_{it}} \prod_{j=1}^{J} \left(\frac{\exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)}{\sum_{j=1}^{J} \exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)} \right)^{y_{itqj}}. \end{aligned} \tag{3.3}$$

Denote by $\#q_{itj}$ the number of choices j that an individual i made during period t. The joint likelihood obtained from (3.3) takes the form

$$L(\beta, \theta|y, X) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=1}^{J} \left(\frac{\exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)}{\sum_{j=1}^{J} \exp\left(g\left(X_{itqj}, \beta_i, \theta_i\right)\right)} \right)^{\#q_{itj}}. \tag{3.4}$$

This setup is a generalization of the multinomial mixed logit model that is obtained by setting $\#q_{itj} = 1$ or 0 for each j, i, t. Mixed logit is a flexible discrete choice model that allows for random coefficients and/or error components that induce correlation over alternatives and time.

Recalling the notation of the latent class DPM model (2.2), let $z_i = \beta_i$, $\psi = \{b_\beta, \Sigma_\beta\}$, and $\gamma_c = \{b_{\beta_c}, \Sigma_{\beta_c}\}$. Let ϕ represent the Normal density. The hierarchical model structure is specified as follows:

$$\begin{aligned} &\beta_i|c_i, \gamma, \theta_i, y_i, X_i \sim F(\cdot; \gamma_{c_i}) \equiv N(b_{\beta c_i}, \Sigma_{\beta c_i}) \\ &c_i|\mathbf{p} \sim \text{Discrete}(p_1, \ldots, p_C) \\ &\mathbf{p} \sim \text{Dir}(\alpha/C, \ldots, \alpha/C) \\ &\gamma_c \sim G_0 \equiv \text{BVN}(b_{0\beta}, \Sigma_{0\beta})\text{IW}(\upsilon_0, S_0) \\ &\theta_i \sim \text{MVN}(b_\theta, \Sigma_\theta) \end{aligned}$$

where

$$\begin{aligned} \text{BVN}(b_{0\beta}, \Sigma_{0\beta}) &= -\frac{d_\beta}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_{0\beta}|) \\ &\quad - \frac{1}{2}\left(b_\beta - b_{0\beta}\right)' \Sigma_{0\beta}^{-1}\left(b_\beta - b_{0\beta}\right) \\ \text{IW}(\upsilon_0, S_0) &= \frac{|S_0|^{\upsilon_0/2}|\Sigma_\beta|^{-(\upsilon_0+d_\beta+1)/2} \exp\left(-tr\left(S_0\Sigma_\beta^{-1}\right)\Big/2\right)}{2^{\upsilon_0 d_\beta/2}\Gamma_{d_\beta}(\upsilon_0/2)} \end{aligned}$$

Fig. 4. $\alpha = 10$. Left: Evolution of the number of latent classes over the MC chain. Right: Average number of latent class members, sorted by size.

with the multivariate gamma function specified as

$$\Gamma_{d_\beta}(v_0/2) = \pi^{d_\beta(d_\beta-1)/4} \prod_{j=1}^{d_\beta} \Gamma(v_0/2 + (1-j)/2)$$

and a diffuse prior on $\{b_\theta, \Sigma_\theta\}$. Consequently, our estimation is based on the following Gibbs blocks:

(1) Given the state of the system:
  (a) Update latent classes c_i using the scheme described in Algorithm 7 of Neal (2000)
  (b) $b_{\beta c_i}|\beta_i, \Sigma_{\beta c_i}, \theta_i, b_\theta, \Sigma_\theta \forall i$ s.t. $c_i = c$
  (c) $\Sigma_{\beta c_i}|\beta_i, b_{\beta c_i}, \theta_i, b_\theta, \Sigma_\theta \forall i$ s.t. $c_i = c$
(2) $\beta_i, \theta_i|b_{\beta c_i}, \Sigma_{\beta c_i}, \theta_i, b_\theta, \Sigma_\theta \forall i$
(3) $b_\theta|\beta_i, b_{\beta c_i}, \Sigma_{\beta c_i}, \theta_i, \Sigma_\theta$
(4) $\Sigma_\theta|\beta_i, b_{\beta c_i}, \Sigma_{\beta c_i}, \theta_i, b_\theta$.

The Gibbs updates in blocks 1b and 3 are implemented using result A in Train (2003), p. 298, and the updates in blocks 1c and 4 are implemented using result B in Train (2003), p. 300. The updates in block 2 are performed using the likelihood (3.4) with random walk Metropolis–Hastings steps (see e.g. Train, 2003). Further details on the implementation of the DPM estimation procedure for our model are discussed further below.

3.2. Example 1: Estimation of skewed preferences with and without DPM

Before applying the estimation procedure to real data, we estimate two challenging models with preference heterogeneity, using simulated data. We fix the number of observations to $N = 675$ and $T = 24$ since these are the dimensions of the data we will employ in the next section. We simulate the data, using the model specification described in Section 3.1. We generate two variables X_1 and X_2 as random draws from Uniform $[-5, 5]$ distribution. Consumers can choose between six different alternatives, and we also generate choice specific effects $\theta \sim \text{MVN}(0, I_5)$. We also assume that each consumer undertakes seven shopping trips in each period.

In the first example, we want to capture the intuition that in some models it is important to account for skewed preferences. Consumers may feel very strongly about a particular product characteristic, and hence their preferences will be skewed on one side of the real line with almost no probability mass in the opposite tail. These distributions cannot be modeled as Normal distributions and thus we would expect a parametric model to fail to capture them. In order to simulate preferences which have this skewness property, we need to draw β_i from a distribution with these features. Harding and Hausman (2008) show that a flexible parametric form which allows for skewness and which can be easily implemented numerically can be constructed from the convolution of a Normal kernel with a skewing function. Thus, in order to simulate the data, consumer taste parameters β_i are drawn from the multivariate distribution f consisting of a Normal kernel ϕ and a logistic function G, such that

$$f(\beta; b, \Sigma, \lambda) = 2\phi(\beta; b, \Sigma)G(\lambda'(\beta - b)), \tag{3.5}$$

ϕ is the probability density of a Normal distribution with mean b and covariance matrix Σ, G is the cdf of a logistically distributed random variable with mean 0 and variance $\pi^2/3$ with $G(y) = \frac{1}{1+\exp(-y)}$, λ is a p-dimensional vector of skewness parameters. We call f the skew-Normal-logistic, SNL(b, Σ, λ) distribution. Note, in particular, that the distribution of β approaches that of the Half-Normal distribution $|\beta|$ as $\lambda \to \infty$. In the first example, preferences are assumed to follow a SNL$(0, I_2, [50, 50])$ distribution. We plot these preferences in Fig. 5, where the left panel shows the 3D density while the right panel shows the corresponding contour plot. It is easy to see that these preferences are skewed towards the first quadrant.

We apply both a parametric Bayesian estimation procedure that imposes a Normal prior on the preference distribution of β and the nonparametric DPM procedure. The resulting posterior estimates are plotted in Fig. 6, and the corresponding Markov chains for the DPM estimation are shown in Fig. 7. Notice how estimating the model under the Normality assumption fails to capture the skewness which characterizes the underlying preferences. The DPM on the other hand, recovers a posterior which is close to the original skewed preference generating process.

3.3. Example 2: Estimation of multimodal preferences with and without DPM

For our second example, we choose an even more challenging example, which allows for multimodal preferences. This is a plausible assumption in cases where consumers have extremely polarized preferences over a given set of product attributes. It is possible to imagine situations where different segments of the consumer population feel very differently about a certain characteristic. Consider a product which contains nuts. Some consumers may love the extra crunchiness, many will feel indifferent and some may be allergic to them. Additionally each segment may have different degrees of skewness. In this example, $\beta^1 \sim \text{SNL}(1, 1, 40)$ and $\beta^2 \sim \text{SNL}(-2, 1, 80)$ for 25% of the data, $\beta^1 \sim \text{SNL}(-2, 1, 70)$ and $\beta^2 \sim \text{SNL}(-2, 1, 70)$ for another 25% of the data, and $\beta^1 \sim \text{SNL}(1, 1, -50)$ and $\beta^2 \sim \text{SNL}(1, 1, -50)$ for the remaining 50% of the data. The distribution of these simulated preferences is shown in Fig. 8.

We estimate these preferences, using both the parametric Normal model and the nonparametric DPM model. The estimated