# A Bayesian Mixed Logit-Probit Model for Multinomial Choice [*]

Martin Burda,[†] Matthew Harding,[‡] Jerry Hausman,[§]

July 2, 2008

## 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.

*JEL:* C11, C13, C14, C15, C23, C25

*Keywords:* Multinomial discrete choice model, Dirichlet Process prior, non-conjugate priors, hierarchical latent class models

[*] 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: www.sharcnet.ca).

[†] Department of Economics, University of Toronto, Sidney Smith Hall, 100 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-4479; Email: `martin.burda@utoronto.ca`

[‡] Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 723-4116; Fax: (650) 725-5702; Email: `mch@stanford.edu`

[§] Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142; Email: `jhausman@mit.edu`

# 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 sub-models 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), Fergusson (1973a), Fergusson (1973b), and Blackwell and MacQueen (1973), though applications were quite limited until the late 1980s. Fuelled by advances 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, Branco, and Gelfand 2002), finance (Kacperczyk, Damien, and Walker 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, Hansen, McCulloch and Rossi (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 the 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 are only slowly spilling over to other areas (Dahl 2005), (Jain and Neal 2007), and (Dahl, Mo, and Vannucci 2008).

Another important feature of our model is its hierarchical structure with respect to parameters. Hierarchical models provide a natural environment for 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, Allenby, and McCulloch 2005).

Finally, all previous studies utilizing the DPM cited above estimated non-parametrically 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, $\psi$ are treated as random variables which implies further specification of their respective probability distribution.

Consider an exchangeable sequence $z = \{z_i\}_{i=1}^n$ of realizations of a set of random variables $Z = \{Z_i\}_{i=1}^n$ defined over a measurable space $(\Phi, \mathcal{D})$ where $\mathcal{D}$ is a $\sigma$-field of subsets of $\Phi$. 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 $\sigma$-field of subsets of $\Psi$. 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[5] considered here, the joint distribution of data and parameters is defined as a mixture

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

where $G$ is the mixing distribution over $\psi$. It is useful to think of $G(d\psi)$ as the conditional distribution of $\psi$ given $G$. The distribution of the parameters, $G$, is now random which leads to a complete flexibility of the resulting mixture. The model parameters $\psi$ are no longer restricted to follow any given pre-specified 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, Fergusson (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 $\alpha$. The distribution $G_0$ may be viewed

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

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 $\alpha$ 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 $\alpha$, 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 $\Psi$ 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 $\alpha$ be a positive real number. Following Fergusson (1973a), a *Dirichlet Process* on $(\Psi, \mathcal{B})$ with a base measure $G_0$ and a concentration parameter $\alpha$, denoted by $DP(G_0, \alpha) \in \mathcal{M}(\mathcal{M}(\Psi))$, 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 $\Phi$, the random vector $(G(\Psi_1), ..., G(\Psi_J))$ is distributed as $(G(\Psi_1), ..., G(\Psi_J)) \sim Dir(\alpha G_0(\Psi_1), ..., \alpha G_0(\Psi_J))$ where $Dir(\cdot)$ denotes the Dirichlet distribution. We write $G \sim DP(G_0, \alpha)$ if $G$ is distributed according to the Dirichlet process $DP(G_0, \alpha)$. A Bayesian model with such feature is commonly referred to as a Dirichlet Process Mixture (DPM) model.[6] Since realizations of the DP are discrete with probability one, a DPM can be viewed as a countably infinite mixture (Fergusson 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 case is arguably quite limited as these need to adhere to the class of the prior. The exponential family of functions are 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 is 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

[6] 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).

Richardson (2001), Dahl (2005), Jain and Neal (2007), and Dahl, Mo, and Vannucci (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 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, $\mathrm{Dir}(\alpha/C, ..., \alpha/C)$, where

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

with $p_c \geq 0$ and $\sum p_c = 1$. The parameters $\gamma_c$ 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):

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

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

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

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:

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

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

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

which is equivalent to the conditional probabilities (5.1) 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 $\psi$ and has the DP prior. The parameter $\alpha$ of the DP prior controls the number of components in the mixture, such that a larger $\alpha$ results in a larger number of components. In contrast to the conjugate case, a different object is being integrated out than in (5.1), 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 this feature in our semiparametric model specification by handling the non-parametric 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 a 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, one customer decides to either stay at his current table or move to another table that would suit him or her "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 inference 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 wield 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 non-parametric. This step is tantamount to customers switching tables in the Chinese restaurant process.

(2) For each latent class, draw new values of parameters $\gamma_{c_i}$ 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 $F(z_i;\gamma_{c_i})$. Singleton $z_i$, on the other hand, are re-distributed among the existing latent classes with 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 non-parametric form of the density of $\psi$. 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 $\psi$ 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 $\mathbf{c} = (c_1, ..., c_n)$ and $\gamma = (\gamma_c : c \in \{c_1, ..., c_n\})$. Repeatedly sample as follows:

- For $i = 1, ..., n$, 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 $\gamma_{c^*}$ 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 $i = 1, ..., n$ : 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 $\{c_1, ..., c_n\}$ using the following probabilities:
$$P(c_i = c|c_{-i}, y_i, \gamma, c_i \in \{c_1, ..., c_n\}) = b\frac{n_{-i,c}}{n-1}L(\gamma_c|z_i)$$
where $b$ is the appropriate normalizing constant.
- For all $c \in \{c_1, ..., c_n\}$ : Draw a new value from $\gamma_c|z_i$ such that $c_i = c$, or perform some other update to $\gamma_c$ 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 Figure 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 Figure 1 we show the "target" true density from which we draw a sample of $N = 1,000$ observations. In Figure 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

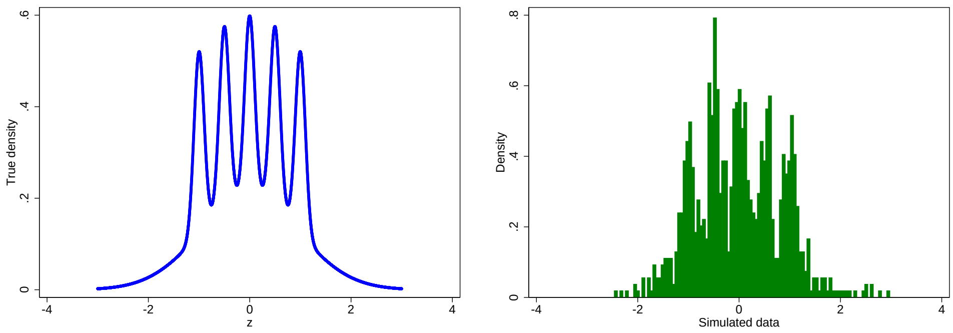

FIGURE 1. Left: trial true functional form of "the claw" posterior density of Marron and Wand (1992). Right: Histogram of a sample draw, $N$ = 1,000.

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 Figure 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 Figure 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 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 Figure 2 we average over the resulting mixture distributions with an additional component of $G_0$ with a weight $\alpha/n$. 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 Figure 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 Equation 2.6 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

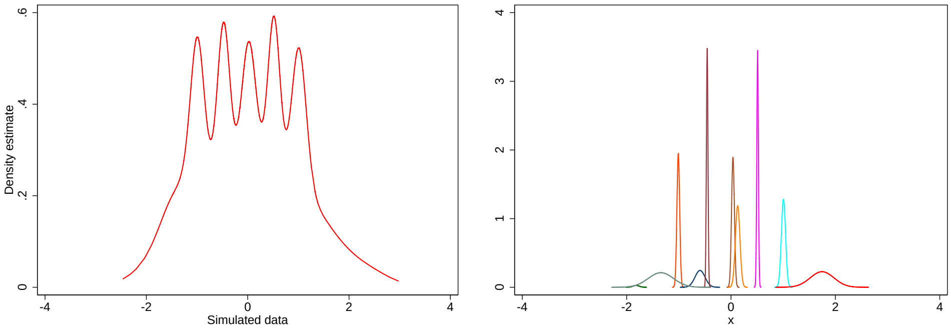

FIGURE 2. Left: DPM density estimate based on the sample in Figure 1, with 10,000 MC steps. Right: A typical snapshot of latent class positions scaled by the class membership intensity.

new clusters of their own. This property is often referred to preferential attachment or "the rich get richer property". Depending on applications this property of the random partitions induced by the Dirichlet process may prove advantageous or not. 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).

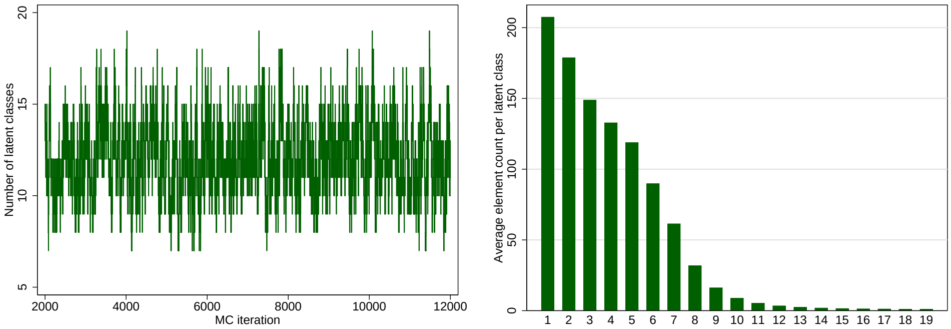

FIGURE 3. $\alpha = 1$. Left: Evolution of the number of latent classes over the MC chain. Right: Average number of latent class members, sorted by size.

The clustering property is also controlled by the parameter $\alpha$. On the one hand it measures the weight placed on the prior base distribution $G_0$. Small values of $\alpha$ correspond to more weight being placed on the prior base distribution $G_0$, while large 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 $\alpha$ also controls the relative decay in class membership as one moves between classes. A small value of $\alpha$ 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 $\alpha$ 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 Figures 3 and 4. We can see that as we increase $\alpha$ 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.

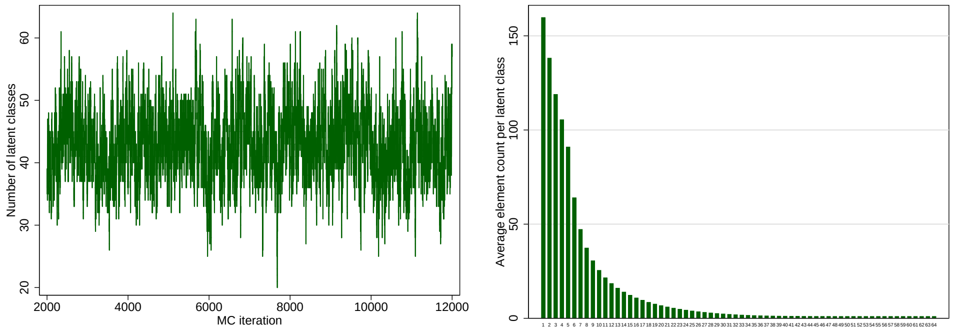

FIGURE 4. $\alpha = 10$. Left: Evolution of the number of latent classes over the MC chain. Right: Average number of latent class members, sorted by size.

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 $i = 1, ..., N$ individuals and an (unordered) set $\{1, ..., J\}$ 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 $\varepsilon_{ijt}$ 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, $\beta = (\beta_1, ..., \beta_N)'$ , $\theta = (\theta_1, ..., \theta_N)'$ are vectors of unknown coefficients. The distribution of $\beta_i$ is modeled nonparametrically while $\theta_i$ – coefficients on alternative specific indicator variables – are assumed to follow a multivariate normal distribution. We will further assume in the model implementation that

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

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 $y_{itqj} \in \{1, ..., J\}$ denote the observed choice outcome. For the logistic specification of $\varepsilon_{ijt}$, the probability of such choice is given by

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

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

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

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

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

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

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

This setup is a generalization of the multinomial mixed logit model that is obtained by setting $\#q_{itj}=1$ 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 $\phi$ represent the Normal density. The hierarchical model structure is specified as follows:

$$\begin{aligned}
\beta_i|c_i,\gamma,\theta_i,y_i,X_i &\sim F(\cdot;\gamma_{c_i})\equiv N(b_{\beta c_i},\Sigma_{\beta c_i})\\
c_i|\mathbf{p} &\sim \text{Discrete}(p_1,...,p_C)\\
\mathbf{p} &\sim \text{Dir}(\alpha/C,...,\alpha/C)\\
\gamma_c &\sim G_0\equiv BVN(b_{0\beta},\Sigma_{0\beta})IW(v_0,S_0)\\
\theta_i &\sim MVN(b_\theta,\Sigma_\theta)
\end{aligned}$$

where

$$\begin{aligned}
BVN(b_{0\beta},\Sigma_{0\beta}) &= -\frac{d_\beta}{2}\log(2\pi)-\frac{1}{2}\log(|\Sigma_{0\beta}|)-\frac{1}{2}\left(b_\beta-b_{0\beta}\right)'\Sigma_{0\beta}^{-1}\left(b_\beta-b_{0\beta}\right)\\
IW(v_0,S_0) &= \frac{|S_0|^{v_0/2}|\Sigma_\beta|^{-(v_0+d_\beta+1)/2}\exp\left(-tr\left(S_0\Sigma_\beta^{-1}\right)/2\right)}{2^{v_0d_\beta/2}\Gamma_{d_\beta}(v_0/2)}
\end{aligned}$$

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'll 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 $\theta \sim MVN(0, I_5)$. 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 $\beta_i$ from a distribution with these features. Harding and Hausman (2007b) 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 $\beta_i$ are drawn from the multivariate distribution $f$ consisting of a normal kernel $\phi$ and a logistic function $G$, such that

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

$\phi$ is the probability density of a Normal distribution with mean $b$ and covariance matrix $\Sigma$, $G$ is the cdf of a logistically distributed random variable with mean 0 and variance $\pi^2/3$ with $G(y) = \frac{1}{1+\exp(-y)}$, $\lambda$ is a $p$-dimensional vector of skewness parameters. We call $f$ the skew-Normal-logistic, SNL $(b, \Sigma, \lambda)$ distribution. Note in particular that the distribution of $\beta$ 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 Figure 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 $\beta$ and the nonparametric DPM procedure. The resulting posterior estimates are plotted in Figure 6 and the corresponding Markov chains for the DPM estimation are shown in Figure 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 closed to the original skewed preference generating process.

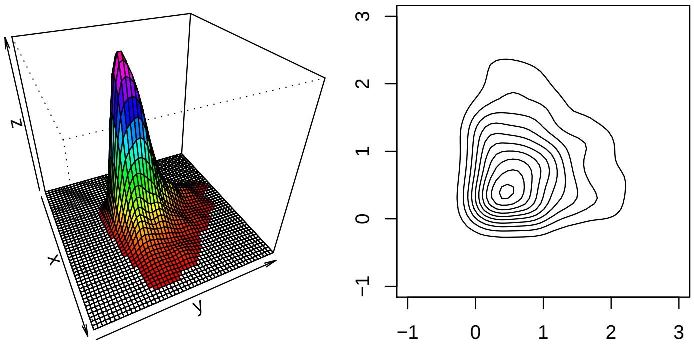

FIGURE 5. Plot of the trial density function out of which simulated $\beta_i$ were drawn.

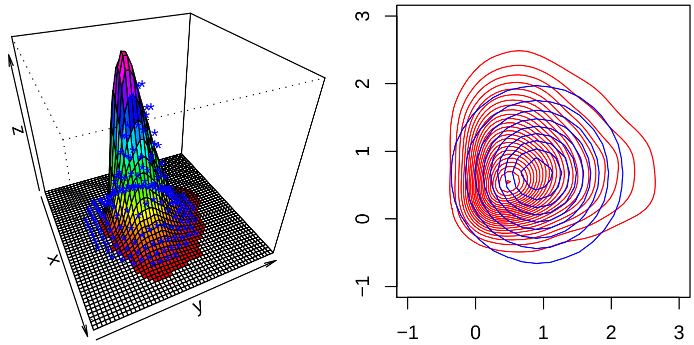

FIGURE 6. DPM estimate of the density of $\beta_i$. $N = 675, T = 24$, trips per one time period $= 7$. Overlay: parametric bivariate normal estimate.

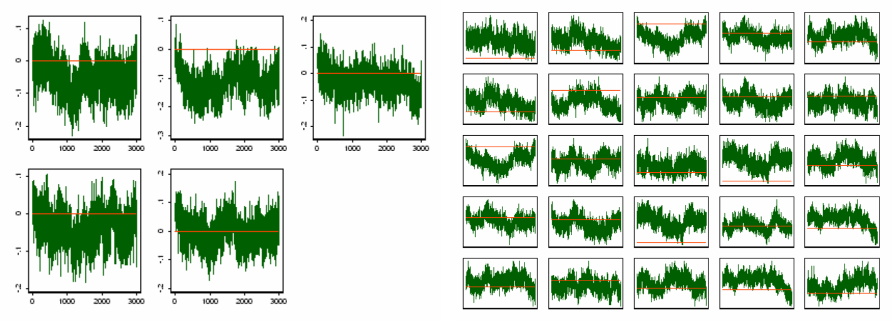

FIGURE 7. $b_\theta$ chain and $\Sigma_\theta$ chain with true values indicated by lines.

### 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, $\beta^1 \sim SNL(1, 1, 40)$ and $\beta^2 \sim SNL(-2, 1, 80)$ for 25% of the data, $\beta^1 \sim SNL(-2, 1, 70)$ and $\beta^2 \sim SNL(-2, 1, 70)$ for another 25% of the data, and $\beta^1 \sim SNL(1, 1, -50)$ and $\beta^2 \sim SNL(1, 1, -50)$ for the remaining 50% of the data. The distribution of these simulated preferences is shown in Figure 8.

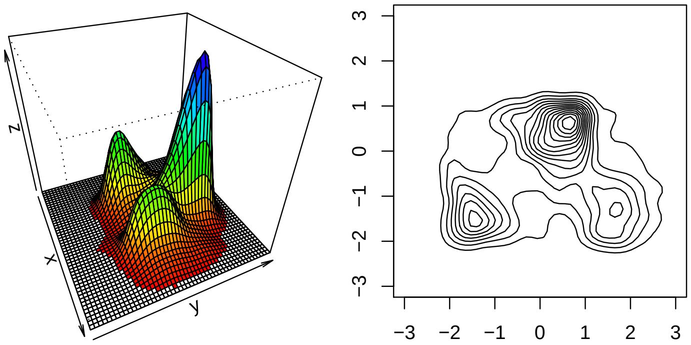

FIGURE 8. Plot of the trial density function out of which simulated $\beta_i$ were drawn.

We estimate these preferences using both the parametric Normal model and the non-parametric DPM model. The estimated preferences from the DPM model are shown in Figure 9 with an overlay of the parametric bivariate Normal model. Figure 10 shows the Markov chain parameter draws for alternative-specific indicator variable coefficients and their variance-covariance matrices with the true values indicated by red lines, which in the vast majority of cases lie within the standard errors of the estimates.

The aim of this simulation is not to show that the parametric Normal model cannot capture these complex preferences which is something we could have expected. Rather we want to exemplify the power of the non-parametric approach in capturing a complex preference structure.

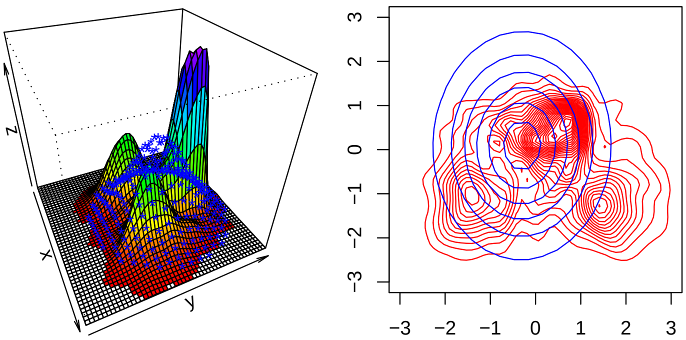

FIGURE 9. DPM estimate of the density of $\beta_i$. $N = 675, T = 24$, trips per one time period $= 7$. Overlay: parametric bivariate normal estimate.

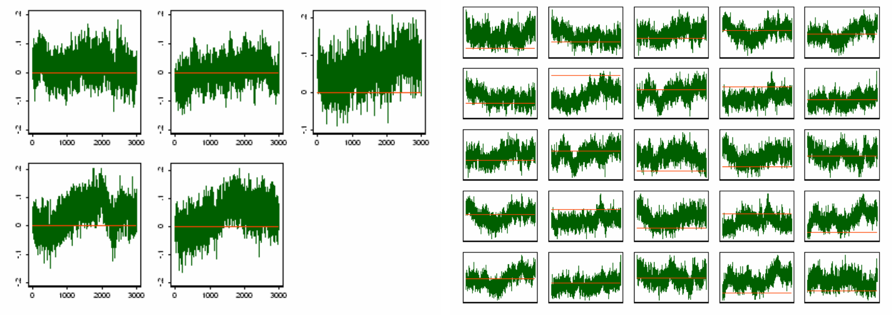

FIGURE 10. $b_\theta$ chain and $\Sigma_\theta$ chain with true values indicated by lines.

## 4. Application: Store Choice

### 4.1. Data Description

In order to explore the performance of our method in practice we now introduce a stylized yet realistic application to consumers' choice of grocery stores. Our dataset consists of observations on 675 households in the Houston area whose shopping behavior was tracked using store scanners over 24 months during the years 2004 and 2005 by AC Nielsen. We only consider each household as having a choice among 5 different stores (H.E. Butt, Kroger, Randall's, Walmart, PantryFoods[7]). We also allow for an additional category labelled "Other" which encompasses all other stores that fall under the standard grocery format, but exclude club stores or convenience stores.

Most consumers shop in at least two different stores in a given month with the average number of trips to their first choice store being approximately once a week. The mean number of trips per month conditional on shopping at a given store for the stores in the sample is: H.E. Butt (4.05), Kroger (4.60), Randall's (3.21), Walmart (4.07), PantryFoods (2.91), Other (3.91). The historgram in Figure 11 summarizes the frequency of each trip count for the households in the sample. The mode of the distribution is 7.

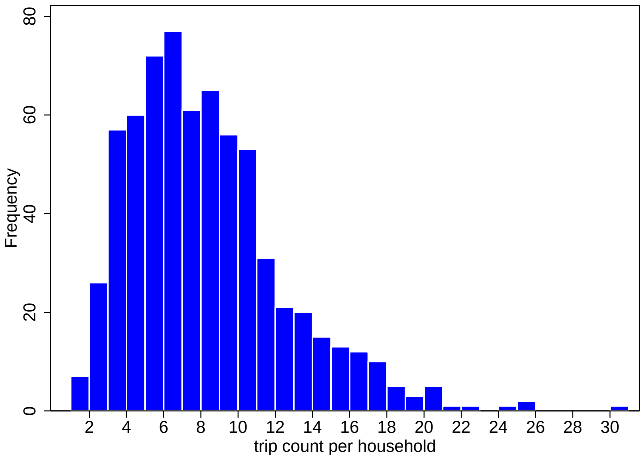

FIGURE 11. Histogram of the total number of trips to a store per month for the households in the sample.

We employ two key variables, *price*, which corresponds to the price of a basket of goods in a given store-month and *distance*, which corresponds to the estimated driving distance for each household to the corresponding supermarket. Since the construction of these variables from individual level scanner data is not immediate some further details are in order to understand the meaning of these variables.

In order to construct the price variable we first normalized observations from the price paid to a dollars/unit measure, where unit corresponds to the unit in which the idem was sold. Typically,

[7] PantryFoods stores are owned by H.E. Butt and are typically limited-assortment stores with reduced surface area and facilities.

| Product Category | Weight |
|---|---|
| Bread | 0.0804 |
| Butter and Margarine | 0.0405 |
| Canned Soup | 0.0533 |
| Cereal | 0.0960 |
| Chips | 0.0741 |
| Coffee | 0.0450 |
| Cookies | 0.0528 |
| Eggs | 0.0323 |
| Ice Cream | 0.0663 |
| Milk | 0.1437 |
| Orange Juice | 0.0339 |
| Salad Mix | 0.0387 |
| Soda | 0.1724 |
| Water | 0.0326 |
| Yogurt | 0.0379 |

Table 1. Product categories and the weights used in the construction of the price index.

this is ounces or grams. For bread, butter and margarine, coffee, cookies and ice cream we drop all observations where the transaction is reported in terms of the number of units instead of a volume or mass measure. Fortunately, few observations are affected by this alternative reporting practice. We also verify that only one unit of measurement was used for a given item. Furthermore, for each produce we drop observations for which the price is reported as being outside two standard deviations of the standard deviations of the average price in the market and store over the periods in the sample.

We also compute the average price for each product in each store and month in addition to the total amount spent on each produce. Each product's weight in the basket is computed as the total amount spent on that product across all stores and months divided by the total amount spent across all stores and months. We look at a subset of the total product universe and focus on the following product categories: bread, butter and margarine, canned soup cereal, chips, coffee, cookies, eggs, ice cream, milk, orange juice, salad mix, soda, water, yogurt. The estimated weights are given in Table 1.

For a subset of the products we also have available directly comparable product weights as reported in the CPI. As shows in Table 2 the scaled CPI weights match well with the scaled produce weights derived from the data. The price of a basket for a given store and month is thus the sum across product of the average price per unit of the product in that store and month multiplied by the product weight.

| Product Category | 2006 CPI Weight | Scaled CPI Weight | Scaled Product Weight |
|---|---|---|---|
| Bread | 0.2210 | 0.1442 | 0.1102 |
| Butter and Margarine | 0.0680 | 0.0444 | 0.0555 |
| Canned Soup | 0.0860 | 0.0561 | 0.0730 |
| Cereal | 0.1990 | 0.1298 | 0.1315 |
| Coffee | 0.1000 | 0.0652 | 0.0617 |
| Eggs | 0.0990 | 0.0646 | 0.0443 |
| Ice Cream | 0.1420 | 0.0926 | 0.0909 |
| Milk | 0.2930 | 0.1911 | 0.1969 |
| Soda | 0.3250 | 0.2120 | 0.2362 |

Table 2. Comparison of estimated and CPI weights for matching product categories.

In order to construct the distance variable we employ GPS software to measure the arc distance from the centroid of the census tract in which a household lives to the centroid of the zip code in which a store is located. For stores in which a household does not shop in the sense that we don't observe a trip to this store in the sample, we take the store at which they would have shopped to be the store that has the smallest arc distance from the centroid of the census tract in which the household lives out of the set of stores at which people in the same market shopped. If a household shops at a store only intermittently, we take the store location at which they would have shopped in a given month to be the store location where we most frequently observe the household shopping when we do observe them shopping at that store. The store location they would have gone to is the mode location of the observed trips to that store. Additionally, we drop households that shop at a store more than 200 miles from their reported home census tract.

### 4.2. **Implementation Notes**

The estimation results along with auxiliary output are presented in below. In implementation of the DPM algorithm, for the univariate case, the starting parameter values for $\beta$ and $\theta$ were obtained by draws from a standard normal distribution with a random assignment to 10 initial latent classes. For the bivariate case, the individual logit estimates were taken as starting values binned to 10 initial latent classes. The maximum number of latent classes was set to 50 but this artificial ceiling was never a binding constraint: the mean number of actual latent classes was 9.19 with a standard deviation of 2.32 in the univariate case and 18.07 classes with a standard deviation of 1.96 in the bivariate case. We subjected the RW-MH updates in the second Gibbs block to scale parameters

$\rho_{beta} = 0.5$ for $\beta_i$ and $\rho_{theta} = 0.1$ (for a discussion, see e.g. p. 306 in Train 2003) which resulted desired acceptance rates of approximately 0.3.

All chains appear to be mixing well and having converged. In contrast to frequentist methods, the draws from the Markov chain converge in distribution to the true posterior distribution, not to point estimates. For assessing convergence, we use the criterion given in Allenby, Rossi, and McCulloch (2005) characterizing draws as having the same mean value and variability over iterations. Plots of individual chains are not reported here due to space limitations but will can be provided on request. The estimated DPM posterior also appeared relatively insensitive to the choice of the DPM prior scalar parameter $\alpha$. This is illustrated in Figure(14) = showing the bivariate DPM estimates for $\alpha = 1$. Following Neal (2000), we set $\alpha = 1$ throughout the analysis. In principle, one could incorporate learning about $\alpha$ using the algorithm of Escobar and West (1995).

During the MC run, for small values of $v_0$ (the shape parameter in the prior on $\Sigma_{\beta c_i}$ implying a diffuse prior when small), one latent class tended to eventually span the parameter space which resulted in significant over-smoothing of the final estimate. We have not encountered such phenomenon in any of the simulated cases. In our logit-probit model, the goal was to group individuals of similar tastes into latent classes in each MC step that can be locally well-defined without subsuming the entire parameter space. The resulting density estimate should also be capable of differentiating sufficient degree of local variation. Hence, we imposed an flexible upper bound on the variance of each latent class: if any such variance exceeded double the prior on $\Sigma_{\beta c_i}$, the strength of the prior belief expresses as $v_0$ was raised from the default minimum until the constraint was satisfied. This left the size of the latent classes to vary freely up to double the prior variance. No bound was imposed on covariances in the bivariate case. We took the composition of the innate cluster structure of the individual logit estimates as the source of prior information for determining a suitable prior on $\Sigma_{\beta c_i}$. Thus the prior was set to 0.2 for the variance of $\beta_{1c_i}$ and 0.05 for $\beta_{1c_i}$. Priors on all means were set to zero and variance of 25 to reach high diffusion in location.

Each model was subjected to 10,000 MC draws out of which the first 5,000 was discarded as a burn-in section. The overall runtime was approximately 10 minutes and 30 minutes for the univariate and bivariate DPM, respectively, on a 2.4 GHz Unix workstation using the PGI 6.1 Fortran 90 compiler.

### 4.3. **Estimation results**

In order to explore the performance of the DPM estimator in real data we estimate a series of stylized econometric models. We choose a subset of commonly employed econometric models for discrete choices in order to observe how the estimation results change as we progressively relax a series of assumptions. Thus, we estimate both fixed and random coefficients models and also compare parametric and nonparametric estimates. In order to avoid confounding effects we restrict our attention to a series of univariate and bivariate models in price and distance using the dataset described in the previous section.

### 4.3.1. ***Univariate Models***

Consumers incur a cost of time as they travel to the store. We expect this cost of time to interact with the savings in price. Thus, we define the main variable of interest as price*distance. In all specifications the employed variables are defined in logs of the original values. This removes the effect of outliers and produces a more robust specification. Throughout we have found the models to have excellent numerical properties as recorded by the convergence of the Markov chains. [8]

Additionally we include five indicator variables for each of the main stores H.E. Butt, Kroger, Randall's, Walmart and PantryFoods, leaving out the Other category of grocery stores.

The simplest model we can run is the standard fixed coefficients logit model as implemented in numerous software packages including STATA. We report the estimated coefficients in Table 3, where $\beta_1$ corresponds to the coefficient on the price*distance variable while $\theta_1$ through $\theta_5$ correspond to the coefficients on the store indicator variables. As we would expect from economic theory, the estimated coefficient $\beta_1$ on the shopping cost variable is negative. The reported asymptotic standard error is very small. The estimated coefficients on the store indicator variables have a mixed sign pattern. The coefficients for Kroger and Walmart imply a positive store effect relative to the Other category while the coefficients for H.E. Butt, Randall's and PantryFoods are negative.

| Variable | $b_{\beta 1}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|
| Mean | -2.829 | -0.377 | 0.930 | -0.258 | 0.165 | -2.061 |
| Std. Dev. | (0.029) | (0.176) | (0.189) | (0.205) | (0.181) | (0.191) |

TABLE 3. Fixed coefficients logit point estimates. Standard errors are in brackets.

We next estimate a parametric Normal random coefficients model where we assume heterogeneity in consumer preferences. The coefficient $\beta_1$ on the price*distance variable is drawn from a univariate Normal distribution, while the coefficients $\theta$ on the store indicator variables are drawn from a multivariate Normal distribution with covariance matrix $\Sigma$. We employ a standard Bayesian MCMC estimation strategy to estimate these two Normal distributions and report the means and variance-covariance matrices with their corresponding standard errors in Tables 4 and 5. The mean of the estimated distribution for $\beta_1$ is -0.292 which is roughly 10 times smaller than the estimate for $\beta_1$ resulting from the fixed coefficients logit model. By contrast the standard error has also increased

[8] We do not report results on the convergence of the Markov chains in the paper but they are available from the authors upon request.

by a factor of 10. The means of the distributions on the store indicator variables have the same sign patterns as the estimates from the fixed coefficients model, but some of the magnitudes have changed. In particular the negative store effects for H.E. Butt, Randall's and PantryFoods increase substantially.

The parametric Normal random coefficients model also estimates the covariance matrix between the coefficients on the store indicator variables which is reported in Table 5. Since most consumers appear to buy their groceries at more than one store over the course of a month, most of the estimated covariances are positive.

| Variable | $b_{\beta 1}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|
| Mean | -0.292 | -1.616 | 0.437 | -2.196 | 0.156 | -3.799 |
| Std. Dev. | (2.066) | (0.161) | (0.130) | (0.165) | (0.132) | (0.187) |

TABLE 4. Parametric Normal random coefficients model. Means of MCMC $(b_\beta, b_\theta)$ draws. Standard errors are in brackets.

| | $\Sigma_{\theta.1}$ | $\Sigma_{\theta.2}$ | $\Sigma_{\theta.3}$ | $\Sigma_{\theta.4}$ | $\Sigma_{\theta.5}$ |
|---|---|---|---|---|---|
| $\Sigma_{\theta 1.}$ | 16.12 (0.99) | 3.11 (0.61) | 5.73 (0.71) | 6.56 (0.64) | 0.04 (0.77) |
| $\Sigma_{\theta 2.}$ | | 9.71 (0.68) | 6.04 (0.59) | 4.36 (0.51) | 2.44 (0.53) |
| $\Sigma_{\theta 3.}$ | | | 13.01 (1.03) | 2.94 (0.58) | -0.62 (0.64) |
| $\Sigma_{\theta 4.}$ | | | | 11.12 (0.74) | 2.98 (0.61) |
| $\Sigma_{\theta 5.}$ | | | | | 14.7 (1.12) |

TABLE 5. Parametric Normal random coefficients model. Means of MCMC $\Sigma_\theta$ draws. Standard errors are in brackets.

We now estimate the semiparametric Bayesian mixed logit-probit model for the univariate case discussed above. We allow the distribution on the price*distance coefficient $\beta_1$ to follow a Dirichlet Process Mixture, while modeling the store effects by a multivariate Normal distribution with a general covariance matrix $\Sigma_\theta$. We plot the estimated preference distribution for the main variable in Figure 12. In addition to a mode which is negative and close to zero the distribution estimated using DPM appears to have another mode close to -5. This corresponds to a group of consumers who are extremely sensitive to the cost of shopping as measured by our price*distance variable. We report the estimated mean coefficient estimates for $\beta_1$ and $\theta_1$ through $\theta_5$ in Table 6. As we would expect from the presence of the additional negative mode, the mean coefficient for the DPM estimate is now -0.406 which more negative than the corresponding estimate from the parametric Normal model. The estimate for the multivariate Normal distribution of store effects is very similar to that obtained using the parametric Normal model. Once again we find an alternative sign pattern for the mean store effects and positive covariances between most of the store effects.

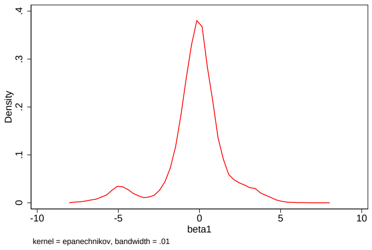

FIGURE 12. DPM estimate of the univariate density of $\beta_1$.

| Variable | $b_{\beta 1}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|
| Mean | -0.406 | -1.613 | 0.431 | -2.215 | 0.184 | -3.544 |
| Std. Dev. | (2.234) | (0.168) | (0.128) | (0.159) | (0.136) | (0.164) |

TABLE 6. Bayesian DPM Logit-Probit Model. Means of MCMC $(b_\beta, b_\theta)$ draws. Standard errors are in brackets.

| | $\Sigma_{\theta.1}$ | $\Sigma_{\theta.2}$ | $\Sigma_{\theta.3}$ | $\Sigma_{\theta.4}$ | $\Sigma_{\theta.5}$ |
|---|---|---|---|---|---|
| $\Sigma_{\theta 1.}$ | 16.33 (1.12) | 3.11 (0.61) | 6.11 (0.80) | 6.60 (0.73) | 0.53 (0.70) |
| $\Sigma_{\theta 2.}$ | | 10.01 (0.69) | 6.38 (0.68) | 4.46 (0.54) | 2.42 (0.53) |
| $\Sigma_{\theta 3.}$ | | | 13.67 (0.99) | 3.11 (0.57) | -0.79 (0.61) |
| $\Sigma_{\theta 4.}$ | | | | 11.02 (0.73) | 3.27 (0.67) |
| $\Sigma_{\theta 5.}$ | | | | | 12.83 (0.92) |

TABLE 7. Bayesian DPM Logit-Probit Model. Means of MCMC $\Sigma_\theta$ draws. Standard errors are in brackets.

### 4.3.2. ***Bi-variate Models***

We now proceed to estimate a series of bi-variate models which include both price*distance and distance as conditioning variables. In Table 8 we report coefficient estimates for the standard fixed coefficients logit model. We denote the coefficient on price*distance by $\beta_1$ and the coefficient on distance by $\beta_2$. We also include five indicator variables in order to capture the store effects for H.E. Butt, Kroger, Randall's, Walmart and PantryFoods, leaving out the Other category of grocery stores. Both main conditioning variables are significant enter with a negative sign. The store effects have an alternating sign pattern which corresponds to the results for the univariate model. In particular we find that Kroger and Walmart have a positive store effect over the excluded category of other grocery stores.

| Variable | $b_{\beta 1}$ | $b_{\beta 2}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|---|
| Mean | -1.950 | -0.202 | -0.413 | 0.858 | -0.518 | 0.283 | -1.937 |
| Std. Dev. | (0.467) | (0.009) | (0.018) | (0.019) | (0.023) | (0.019) | (0.020) |

TABLE 8. Fixed coefficients logit model. Means of MCMC $(b_\beta, b_\theta)$ draws. Standard errors are in brackets.

We then estimate the parametric Normal random coefficients model. We let $(\beta_1, \beta_2)$ be drawn from a bivariate Normal distribution and allow for correlations between the two random coefficients. The modeling of the correlation between taste parameters is necessary in this case given the definition of our two variables as price*distance and distance respectively. Furthermore we model the store effects $\theta$ as having a multivariate Normal distribution with full covariance matrix $\Sigma_\theta$. In Figure 13 we plot the multivariate Normal density estimate for $\beta_1$ and $\beta_2$ and the corresponding contour plot. We notice a wide dispersion of the preference parameters with a mode close to zero. Thus, the median consumer is relatively price insensitive. The presence of negative correlation between consumer attitudes to the cost of shopping and the distance traveled implies that consumers who are more price sensitive are willing to travel longer distances in order to purchase their groceries.

In Table 9 and we report mean coefficients for the $\beta$ and $\theta$ coefficients. The taste coefficients on our two main explanatory variables have very different mean values and standard errors relative to the fixed coefficients logit estimates. Similarly we see some variation in the estimates for the store effects. The signs remain the same however for both sets of coefficients. We find that standard errors have increased substantially in the random coefficients model relative to the fixed coefficients logit model.

In Table 10 we report estimates for the covariance matrix of the store effects $\Sigma_\theta$. These estimates are comparable with what we obtained for the univariate model. In particular we find most of the covariances to be positive as a result of the consumers' propensity in the sample to buy groceries at several different stores within the same time period. The only exception is the negative correlation between Randall's and PantryFoods.

We next estimate the mixed logit-probit model by specifying a Dirichlet Process Mixture for the bivariate distribution of consumer preferences on the first two variables of interest while letting the distribution for the store effects be specified parametrically by a multivariate Normal distribution. We plot the resulting nonparametric estimate of the preference distribution in Figure 14, where we also plot the corresponding contour diagram. Notice that the estimated distribution has pronounced

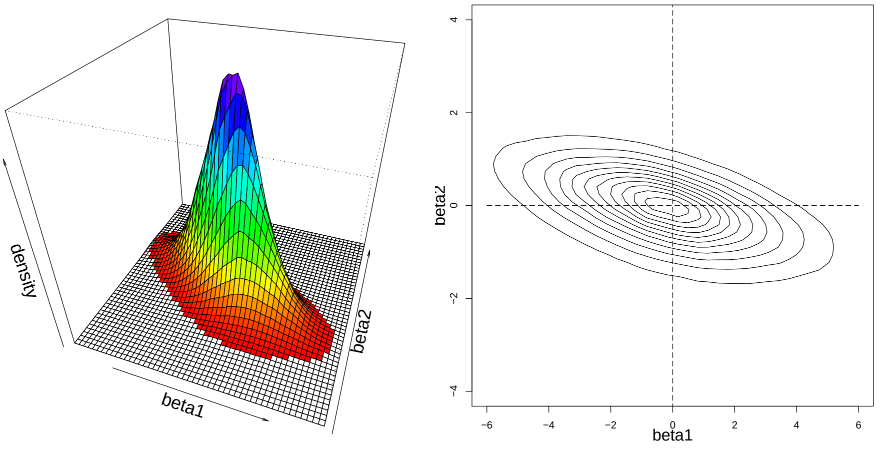

FIGURE 13. Parametric Normal random coefficients model estimate of the bivariate density of $(\beta_{i1}, \beta_{i2})$.

| Variable | $b_{\beta 1}$ | $b_{\beta 2}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|---|
| Mean | -0.296 | -0.086 | -1.641 | 0.404 | -1.982 | 0.165 | -3.372 |
| Std. Dev. | (2.556) | (0.749) | (0.163) | (0.124) | (0.231) | (0.143) | (0.153) |

TABLE 9. Parametric Normal random coefficients model. Means of MCMC $(\beta_i, \theta_i)$ draws. Standard errors are in brackets.

| | $\Sigma_{\theta \cdot 1}$ | $\Sigma_{\theta \cdot 2}$ | $\Sigma_{\theta \cdot 3}$ | $\Sigma_{\theta \cdot 4}$ | $\Sigma_{\theta \cdot 5}$ |
|---|---|---|---|---|---|
| $\Sigma_{\theta 1 \cdot}$ | 15.52 (1.67) | 2.58 (0.92) | 4.88 (2.09) | 6.52 (0.87) | 0.35 (0.72) |
| $\Sigma_{\theta 2 \cdot}$ | | 9.30 (1.13) | 4.79 (1.74) | 4.08 (0.57) | 1.94 (0.50) |
| $\Sigma_{\theta 3 \cdot}$ | | | 10.50 (3.01) | 2.53 (1.06) | -0.96 (0.60) |
| $\Sigma_{\theta 4 \cdot}$ | | | | 10.77 (0.73) | 2.70 (0.60) |
| $\Sigma_{\theta 5 \cdot}$ | | | | | 11.13 (1.14) |

TABLE 10. Parametric Normal random coefficients model. Means of MCMC $\Sigma_\theta$ draws. Standard errors are in brackets.

multi-modal features. While most consumers appear to be insensitive to the cost of shopping for groceries both in terms of price and distance, the preference distribution estimates two additional modes. The first mode corresponds to consumers who are particularly sensitive to the cost of shopping for groceries and are willing to travel a longer distance searching for a better deal. The other mode corresponds to consumers who value proximity to the store and are prepared to pay a higher price

In Table 11 we report the mean coefficient estimates. Both of our main conditioning variables have negative mean coefficients but large amounts of variation. The results are numerically different from

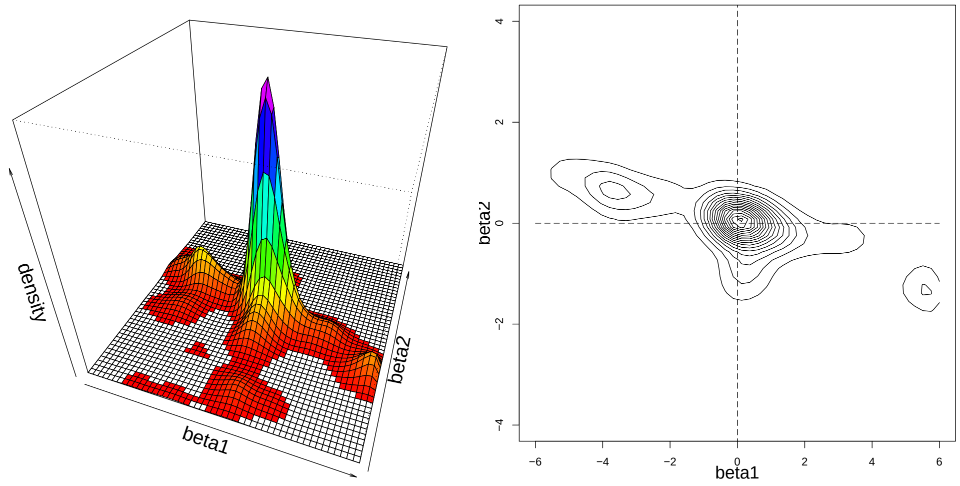

FIGURE 14. DPM estimate of the bivariate density of $(\beta_{i1}, \beta_{i2})$, $\alpha = 1$.

| Variable | $b_{\beta 1}$ | $b_{\beta 2}$ | $b_{\theta 1}$ | $b_{\theta 2}$ | $b_{\theta 3}$ | $b_{\theta 4}$ | $b_{\theta 5}$ |
|---|---|---|---|---|---|---|---|
| Mean | -0.499 | -0.174 | -1.479 | 0.587 | -1.829 | 0.134 | -3.917 |
| Std. Dev. | (2.749) | (1.242) | (0.161) | (0.123) | (0.149) | (0.138) | (0.185) |

TABLE 11. DPM model. Means of MCMC $(b_\beta, b_\theta)$ draws. Standard errors are in brackets.

|  | $\Sigma_{\theta .1}$ | $\Sigma_{\theta .2}$ | $\Sigma_{\theta .3}$ | $\Sigma_{\theta .4}$ | $\Sigma_{\theta .5}$ |
|---|---|---|---|---|---|
| $\Sigma_{\theta 1.}$ | 14.99 (1.06) | 3.29 (0.53) | 5.58 (0.67) | 6.57 (0.63) | 0.16 (0.63) |
| $\Sigma_{\theta 2.}$ |  | 8.67 (0.57) | 4.97 (0.52) | 4.53 (0.48) | 2.08 (0.51) |
| $\Sigma_{\theta 3.}$ |  |  | 10.48 (0.95) | 2.98 (0.54) | -0.69 (0.63) |
| $\Sigma_{\theta 4.}$ |  |  |  | 11.04 (0.75) | 2.56 (0.61) |
| $\Sigma_{\theta 5.}$ |  |  |  |  | 14.05 (1.07) |

TABLE 12. DPM model. Means of MCMC $\Sigma_\theta$ draws. Standard errors are in brackets.

the results in the parametric Normal model but have comparable orders of magnitude. In particular notice that the estimates for the store effects continue to have an alternating sign patterns with Kroger and Walmart enjoying positive effects over the excluded category while H. E. Butt, Randall's and PantryFoods having large negative effects. Additionally we report the estimated covariance matrix between the store effects in Table 12. We continue to find positive correlations between these effects.

### 4.3.3. *Sensitivity to Choice of* $\alpha$

As we noted in Section 2.4 the number of latent classes in the DPM model, while endogenous to the data, also depends on the parameter $\alpha$. At the present we do not have significant experience in choosing $\alpha$. As we previously discussed however, we expect that larger values of $\alpha$ add additional latent classes, thus improving the resolution of the estimated preference distribution.

Existing practice conditions on a given value of $\alpha$, typically $\alpha = 1$ (Neal, 2000). In order to assess the importance of the choice of $\alpha$, we condition on a range of values of in order to explore the extent to which our results are sensitive to different choices of $\alpha$. The short computational time of our approach makes this approach particularly attractive. In additional output, not reported here but available from the authors, we implement a number of different models using increments of $\alpha$ from 0.5 to 100. All result in surprisingly similar estimates of the preference distributions, both visually and numerically. While the overall shape of the distribution remains the same across values of $\alpha$, higher values of $\alpha$ appear to produce an increased number of localized features and even indicate additional modes for a subset of the observations. The estimated means and variances of the model parameters are however quantitatively very similar.

Furthermore, we find little gain from using large values of $\alpha$. We conjecture that since at every step of the Markov chain the procedure implements a mixture distribution with a large number of components and then averages over the resulting mixtures at different steps, increasing the number of components provides little improvement as long as the underlying distribution has a small number of modes.

### 4.3.4. *The Importance of Controlling for Store Effects*

One potential criticism of a model that includes both choice covariates and store indicator variables is that the store indicator variables may capture most of the variation due to average differences in the shopping cost at different stores. In Figure 15 we re-estimate the mixed DPM model without including store effects in the specification. The resulting nonparametrically estimated bivariate distribution for $\beta$ has numerous modes and a complex structure. Adding store effects thus appears to make a substantial difference to the estimated consumer preferences and adds additional smoothing as it controls for some of the additional variation in outcomes across consumers. Adding store effects provides a more focused estimate of consumer preferences.

In some circumstances it is possible to estimate individual coefficients for each consumer. Beggs, Cardell and Hausman (1981) estimate individual logit coefficients and measure the dispersion in preferences in assessing the potential demand for electric cars. In our data we can also estimate some but not all of the individual coefficients by running a standard logit model individual by

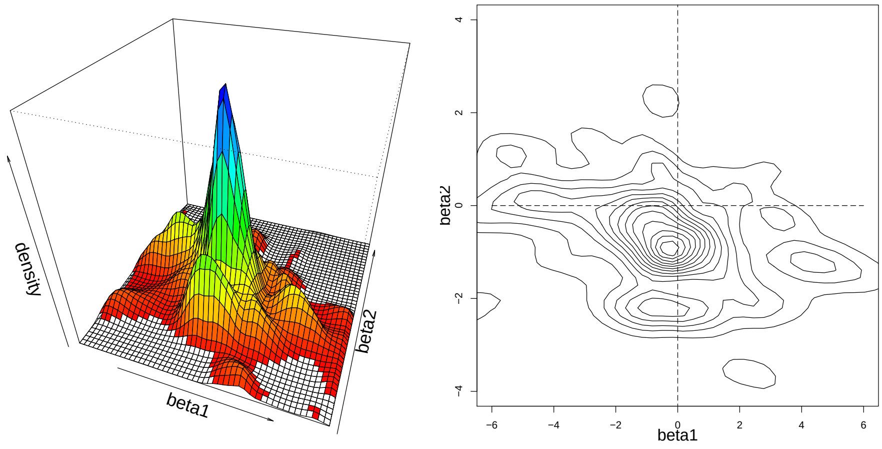

FIGURE 15. DPM estimate of the bivariate density of $(\beta_{i1}, \beta_{i2})$ without individual effects.

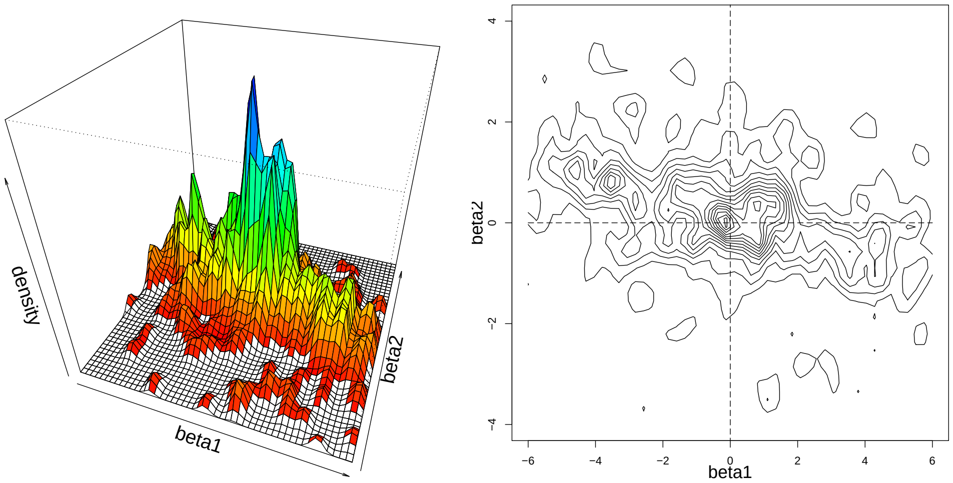

FIGURE 16. Bivariate density of individual logit estimates of $(\beta_{i1}, \beta_{i2})$.

individual. The large $T$ feature of our data helps us to identify individual coefficients but the lack of variation in shopping behavior for some consumers prevents us from estimating coefficients for some individuals.

In Figure 16 we plot the distribution of the identified coefficients using minimal smoothing. While the distribution has numerous modes it is remarkably similar to the distributions estimated by the DPM model above. Note that this distribution of individual logit coefficients was not employed in any way in the computation of the DPM estimate. In particular it appears very similar to the distribution in Figure 15. The DPM model appears to add additional smoothing and shrinkage and also allows for the estimation the distribution of store effects.

## 5. Conclusion

In this paper we have introduced a new flexible mixed model for multinomial discrete choice where key individual and alternative-specific parameters are allowed to follow arbitrary distributions. We also allow for some parameters to follow a more traditional multivariate Normal distribution.

We estimate the model using a Bayesian Markov Chain Monte Carlo Gibbs sampling technique with a multivariate Dirichlet Process prior on the coefficients with nonparametric density. In implementation of the DP prior, we employed a latent class sampling algorithm that is applicable to a general class of models including non-conjugate DP base priors.

We apply our model to the estimation of supermarket choices for a panel of Houston households whose shopping behavior was tracked over a 24-month period in the 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, calculated using GPS software. Our model also allows for supermarket indicator variables.

The semi-parametric model captures a much richer preference structure than a parametric model and estimates a multi-modal preference distribution. While most consumers appear to be inframarginal and relatively insensitive to changes in price and travel distance, some consumers are willing to travel longer distances searching for bargains while others are prepared to pay higher prices for the convenience of shopping at the nearest store.

The Dirichlet Process prior utilizes a sensitivity parameter which controls the number of mixtures estimated at each step of the Markov Chain. While we have currently found the model to be relatively insensitive to the choice of this parameter, future work will address the choice of a prior for this parameter.

## Appendix: The Case of the Conjugate Dirichlet Process Prior

In the conjugate case, the following model (Escobar and West 1995) has often been used as a departure point for estimation:

$$\begin{aligned} y_i|\psi_i &\sim F(\cdot;\psi_i) \\ \psi_i|G &\sim G \\ G &\sim DP(G_0,\alpha) \end{aligned}$$

For this model, Blackwell and MacQueen (1973) have characterized the prior distribution for $\psi_i$ given $\psi_j$, $i \neq j$, by integrating out $G$ as

$$\psi_i|\psi_1,...,\psi_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\psi_j) + \frac{\alpha}{i-1+\alpha}G_0 \tag{5.1}$$

where $\delta(\psi_j)$ is the Dirac measure at $\psi_j$. Combining this prior with the likelihood for $\psi_i$ that results from $z_i$ having distribution $F(\cdot;\psi_i)$, denoted as $L(\psi_i|y_i)$, leads to the following posterior:

$$\psi_i|\psi_{-i}, y_i \sim \sum_{j\neq i} q_{i,j}\delta(\psi_j) + r_i H_i \tag{5.2}$$

where $H_i$ is the posterior distribution for $\psi_i$ based on the prior $G_0$ and $z_i$. The values $q_{i,j}$ and $r_i$ are defined by

$$\begin{aligned} q_{i,j} &= bL(\psi_j|y_i) \\ r_i &= b\alpha\int L(\psi|y_i)dG_0(\psi) \end{aligned} \tag{5.3}$$

where $b$ is a normalizing constant such that $\sum_{j\neq i} q_{i,j} + r_i = 1$. If the DP prior is conjugate to the likelihood, then the integral defining $r_i$ can be derived analytically and the posterior (5.2) is directly amenable to Gibbs sampling. Since the $z_i$ are exchangeable, one can treat each $z_i$ in turn as the last member of a Markov Chain. This estimation method, often referred to as the Polya urn scheme, was used by Escobar (1994), Escobar and West (1995) and subsequently by many researchers who benefited from the conjugacy of the DP prior with respect to their likelihoods.

The Blackwell and MacQueen (1973) sampling scheme (or its variants) cannot easily be applied to models where $G_0$ is not the conjugate prior to $L$ as the integral in (5.3) will usually not be analytically tractable. Sampling from $H_i$ in the posterior (5.2) may also be hard when the prior is not conjugate.

## References

Allenby, G. M., P. E. Rossi, and R. E. McCulloch (2005): "Hierarchical Bayes Models: A Practitioners Guide," Ssrn working paper, Ohio State University, University of Chicago.

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

Athey, S., and G. Imbens (2007): "Discrete Choice Models with Multiple Unobserved Choice Characteristics," working paper, Harvard University.

Bernardo, J. M., and A. F. M. Smith (1994): *Bayesian Theory.* Wiley, New York.

Blackwell, D., and J. B. MacQueen (1973): "Fergusson Distribution via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Chib, S., and B. Hamilton (2002): "Semiparametric bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67–89.

Dahl, D. B. (2005): "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," under invited revision, Journal of Computational and Graphical Statistics.

Dahl, D. B., Q. Mo, and M. Vannucci (2008): "Simultaneous inference for multiple testing and clustering via a Dirichlet process mixture model," *Statistical Modelling*, 8(1), 23–39.

Dicker, L., and S. T. Jensen (2008): "Prior Distributions for Partitions in Bayesian Nonparametrics," mimeo, Harvard University.

Dunson, D. (2005): "Bayesian semiparametric isotonic regression for count data," *Journal of the American Statistical Association*, 100, 618–627.

Escobar, M. D. (1994): "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and M. West (1995): "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fergusson, T. S. (1973a): "A Bayesian Analysis of some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

——— (1973b): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 1, 615–629.

——— (1983): "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, ed. by H. Rizvi, and J. Rustagi. Academic Press, New York.

Freedman, D. (1963): "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case," *Annals of Mathematical Statistics*, 34, 1386–1403.

Green, P., and S. Richardson (2001): "Modelling Heterogeneity with and without the Dirichlet Process," *Scandinavian Journal of Statistics*, 124(28), 355–375.

HARDING, M. C., AND J. A. HAUSMAN (2007): "Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Nonlinear Least Squares," *International Economic Review*, 48(4), 1311–1328.

HAUSMAN, J. A., AND D. A. WISE (1978): "A Conditional Probit Model for Qualitative Discrete-Choice Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46(2), 403–426.

HIRANO, K. (2002): "Semiparametric bayesian inference in autoregressive panel data models," *Econometrica*, 70, 781–799.

IMAI, K., AND D. A. VAN DYK (2005): "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of Econometrics*, 124(2), 311–334.

JAIN, S., AND R. M. NEAL (2007): "Splitting and merging components of a nonconjugate Dirichlet process mixture model," *Bayesian Analysis*, 2(3), 445–472.

JENSEN, M., AND J. MAHEU (2007): "Bayesian semiparametric stochastic volatility modeling," Manuscript, Federal Reserve Bank of Atlanta and University of Toronto.

JOCHMANN, M., AND R. LEÓN-GONZÁLEZ (2004): "Estimating the demand for health care with panel data: a semiparametric Bayesian approach," *Health Economics*, 13, 1003–1014.

KACPERCZYK, M., P. DAMIEN, AND S. G. WALKER (2003): "A new class of bayesian semiparametric models with applications to option pricing," technical report, University of Michigan Bussiness School.

KOTTAS, A., M. D. BRANCO, AND A. E. GELFAND (2002): "A nonparametric bayesian modeling approach for cytogenetic dosimetry," *Biometrics*, 58, 593–600.

MACEACHERN, S. N., AND P. MÜLLER (1998): "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7(2), 223–238.

MARRON, J. S., AND M. P. WAND (1992): "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20(2), 712–736.

MEDVEDOVIC, M., AND S. SIVAGANESAN (2002): "Bayesian infinite mixture model-based clustering of gene expression profiles," *Bioinformatics*, 18, 1194–1206.

MÜLLER, P., AND F. A. QUINTANA (2004): "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19(1), 95–110.

NEAL, R. (2000): "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

ROSSI, P. E., G. M. ALLENBY, AND R. MCCULLOCH (2005): *Bayesian Statistics and Marketing.* Wiley series in Probability and Statistics.

TRAIN, K. (2003): *Discrete Choice Methods with Simulation.* Cambridge University Press.

WALKER, S., AND P. DAMIEN (1998): "Sampling Methods for Bayesian Nonparametric Inference Involving Stochastic Processes," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ed. by D. Dey, P. Müller, and D. Sinha. Springer-Verlag, New York.