Food Marketing Policy Center

Semiparametric Bayesian Estimation of
Random Coefficients Discrete Choice Models

by Sylvie Tchumtchoua and Dipak K. Dey

Food Marketing Policy Center
Research Report No. 102
October 2007

Research Report Series
http://www.fmpc.uconn.edu

University of Connecticut
Department of Agricultural and Resource Economics

Semiparametric Bayesian Estimation of Random Coefficients Discrete Choice Models

Sylvie Tchumtchoua
Department of Statistics, University of Connecticut, Storrs
sylvie.tchumtchoua@uconn.edu

Dipak K. Dey
Department of Statistics, University of Connecticut, Storrs
dipak.dey@uconn.edu

Abstract

Heterogeneity in choice models is typically assumed to have a normal distribution in both Bayesian and classical setups. In this paper, we propose a semiparametric Bayesian framework for the analysis of random coefficients discrete choice models that can be applied to both individual as well as aggregate data. Heterogeneity is modeled using a Dirichlet process prior which varies with consumers characteristics through covariates. We develop a Markov chain Monte Carlo algorithm for fitting such model, and illustrate the methodology using two different datasets: a household level panel dataset of peanut butter purchases, and supermarket chain level data for 31 ready-to-eat breakfast cereals brands.

Keywords: Dependent Dirichlet process, Discrete choice models, Heterogeneity, Markov chain Monte Carlo

JEL classification: C11; C14; C25

1. Introduction

Discrete choice models have been widely used in many fields (e.g., Economics, Marketing) to model instances where individuals select one alternative from a discrete set. In the general setup, a consumer i chooses alternative j from a set of J alternatives if the utility derived from alternative j, $u_{i,j}$, is the highest, i.e., $u_{i,j} > u_{i,k} \quad \forall k = 1,\ldots,J, k \neq j$. The utility, which is latent, is parameterized as $u_{i,j} = \mathrm{x}_{i,j}^{'}\beta + \varepsilon_{i,j}$, where $\mathrm{x}_{\mathrm{i,j}}$ is a vector of observed characteristics of alternative j, β reflects the marginal utility of alternative characteristics (taste parameters), and $\varepsilon_{i,j}$ is an error term commonly assumed to have an Extreme value (0,1) distribution, giving rise to the multinomial logit model. One objective of the model is to use the estimated tastes parameters to compute elasticities (percent change in the probability of choosing an alternative for a one percent change in one of the observed product characteristics (e.g., price), holding the other product characteristics constant. However, restricting the taste parameters β to be identical across individuals creates the Independence of Irrelevant Alternatives (IIA) problem in the multinomial logit model. For example, an increase in the price for one product implies a redistribution of part of the demand for that product to the other products proportionally to their original market shares and not with respect to their characteristics, as one would expect. This restricts the cross-price elasticities to be proportional to market shares. In order to avoid the IIA problem and estimate more realistic substitution pattern among the different products, heterogeneity across consumers in their tastes for the product characteristics is introduced by allowing the taste parameters β to be individual-specific (β_{i}). Since the true distribution of consumer tastes is not observed, the individual-specific parameters β_{i} are typically assumed to be drawn from a parametric distribution.

Discrete choice models can be estimated using either individual (household) level or aggregate (store, supermarket-chain, or market) level data. By individual data we mean consumers and their choices are observed over time. Aggregate-level data consist of total volume (units) sales and dollars sales of a given brand for a store, supermarket-chain, or market over time; individual choices leading to these aggregated quantities are not observed. The econometric methodology for the estimation is well documented. For individual level data see McFadden and Train (2000) for the classical approach, and Yang, Chen, and Allenby (2003), and Rossi, Allenby and McCulloch (2005) for Bayesian version. For aggregated data, see Berry, Levinsohn and Pakes (1995) and Nevo (2001) for the classical setting, and Musalem et al. (2004, 2005), and Chen and Yang (2006) for the Bayesian paradigm.

In both Bayesian and classical models, the distribution of the individual-specific parameters β_i is typically taken to be multivariate normal. The distribution of the individual-specific parameters has important effects on the quantities of interest of the model. For example in many marketing and economic applications, the individual-specific parameters are used to compute price elasticities or to predict the demand for established or new products under alternative pricing strategies. In such applications, reliable estimates of the individual-specific parameters are crucial. The assumption of normality may be too restrictive, since heterogeneity in the population is never known a priori and a normal distribution might not be a good choice; for example, there has been evidence of multimodality in the distribution of taste parameters in marketing studies (e.g., Allenby et al. 1998; Kim et al., 2004). This warrants a more flexible distribution.

There has been some work toward relaxing the normality assumption. Chintagunta et al. (1991) and Kamakura and Russell (1989) used latent class models, which do not capture variation in random coefficients within a latent class. Finite normal mixture models have been used in several studies in the marketing literature (e.g., Allenby et al. (1998); Andrew and Currim (2003) and references therein). In marketing for example, the true number of mixing components is essential since many managerial decisions on segmentation, targeting, positioning, and the marketing mix are based on it. However, determining the number of mixing components remains an unresolved issue. Dillon and Kumar (1994: 345) argued that "The challenges that lie ahead are, in our opinion, clear, falling squarely on the development of procedures for identifying the number of support points needed to characterize the components of the mixture distribution under investigation". More recently, Wedel and Kamakura (2000: 91) affirmed that "the problem of identifying the number of segments is still without a satisfactory solution." In a simulation study, Andrew and Currim (2003) showed that most commonly used mixing component retention criteria do not perform well in the context of multinomial choice data. To overcome the difficulty of choosing the number of mixing components, Kim et al. (2004) proposed the Dirichlet process prior due to Ferguson (1973). Basu and Chib (2003) also used the Dirichlet process prior in binary data regression models. However, the relationship between consumer characteristics and the unknown distribution of heterogeneity cannot be assessed using this distribution. Cifarelli and Regazzini (1978) introduced a product of Dirichlet processes that can be used to model dependence when the covariates have a finite number of levels.

In this paper, we propose a model for which heterogeneity is modeled using a nonparametric distribution which depends on consumer's continuous covariates. Instead of assuming a multivariate normal distribution on the individual-specific parameters (β_i), we use a distribution on the space of all possible distributions, and the order-based dependent Dirichlet process prior introduced by Griffin and Steel (2006) is placed on that distribution. Dependence in the Dirichlet process prior is achieved by making the weights in the Sethuraman (1994) representation of the Dirichlet Process dependent on consumer's continuous covariates.

An attractive feature of our approach is that unlike the Dirichlet process introduced by Ferguson (1973), the dependent Dirichlet process helps recover a richer variety of heterogeneity distributions while allowing the nonparametric distribution to depend on continuous consumer's characteristics. We design a Markov Chain Monte Carlo (MCMC) sampler for assessing the model parameters and apply it to a household-level panel dataset of peanut butter purchases and supermarket-chain level data for 31 ready-to-eat breakfast cereal brands.

The rest of the paper is organized as follows. Section 2 describes the Mixture of Dependent Dirichlet process models (MDDP). In section 3 we apply the MDDP model to the discrete choice model with individual data. In section 4 the model proposed allows estimation with aggregate data. Section 5 contains empirical applications of our methodology. Finally Section 6 presents conclusions.

The Matlab code to implement the method introduced in this paper is available on this website http://sylvie.tchumtchoua.googlepages.com/matlab.

2. Mixture of Dependent Dirichlet Process models

2.1. The Dirichlet Process

The Dirichlet process (Ferguson 1973) is widely used in Bayesian nonparametric applications. It is defined as follows. Let Θ be a probability space, B a σ-algebra of subsets of Θ, H a probability measure on (Θ,B), and M a positive parameter. A random probability measure G on (Θ,B) is said to have a Dirichlet Process DP (M, H) if for any finite measurable partition

$A_1,\ldots,A_p$ of the space the vector $(G(A_1),\ldots,G(A_p))$ follows a Dirichlet distribution with parameters $(M\,H(A_1),\ldots,M\,H(A_p))$.

Using the moments of the Dirichlet distribution, it follows that for $A_i \in \mathrm{B}$,

$$E[G(A_i)] = \frac{M\,H(A_i)}{\sum_{i=1}^{p} M\,H(A_i)} = H(A_i)\,, \tag{1}$$

$$Var[G(A_i)] = \frac{\left(\sum_{i=1}^{p} M\,H(A_i) - M\,H(A_i)\right) M\,H(A_i)}{\left(\sum_{i=1}^{p} M\,H(A_i)\right)^2 \left(\sum_{i=1}^{p} M\,H(A_i) + 1\right)} = \frac{H(A_i)(1 - H(A_i))}{M + 1}\,. \tag{2}$$

The role of H and M are apparent from (1) and (2); H centers the process and is often called the centering distribution or baseline measure. It is a distribution that approximates the true nonparametric shape of G. The scalar M controls the variance of the distribution and is called the precision parameter. It reflects our prior beliefs about how similar the nonparametric distribution G is to the base measure H.

As Ferguson (1973) established, realizations of G are discrete distributions and thus G is not directly used to model data. Escobar (1994) and MacEachern (1994) defined continuous nonparametric distributions by specifying the DP as prior in a hierarchical framework; the resulting model is referred to as a mixture of Dirichlet Process (MDP). It arises as follows. Suppose a random vector y_i has a parametric distribution indexed by a vector β_i which in turn has a prior distribution with known hyperparameters ψ_0. We have

Stage 1: $[y_i \mid \beta_i] \sim f(\beta_i)$,

Stage 2: $[\beta_i \mid \psi_0] \sim f(\psi_0)$,

where f(.) is a generic label for a multivariate probability distribution function. The MDP replaces the parametric prior assumption at the second stage with a general distribution G which in turn has a Dirichlet process prior, leading to the following hierarchical model

Stage 1: $[y_i \mid \beta_i] \sim f(\beta_i)$,

Stage 2: $\beta_i \overset{i.i.d}{\sim} G$,

Stage 3: $G \sim DP(M, H)$.

The above specification is a semiparametric specification because a fully parametric distribution is given in the first stage and a nonparametric distribution is given in the second and third stages.

Two representations of the Dirichlet process are frequently used in the literature. One representation widely used for practical sampling purpose is the Polya urn representation (Blackwell and MacQueen, 1973). If we assume that $\beta_1, \ldots, \beta_n \overset{i.i.d}{\sim} G$ and $G \sim DP(M, H)$, then Blackwell and MacQueen established that

$$G(\beta_n \mid \beta_1, \ldots, \beta_{n-1}) = \frac{1}{M+n-1} \sum_{r=1}^{n-1} \delta_{\beta_r} + \frac{M}{M+n-1} H . \quad (3)$$

Using this representation, $\beta_1, \ldots, \beta_n$ are sampled as follows. β_1 is drawn from the baseline distribution H. The draw of β_2 is equal to β_1 with probability $p_1 = \frac{1}{M+1}$ and is from the baseline distribution with probability $p_0 = 1 - p_1$. The process continues until β_n is sampled.

Three facts are worth noting about the Polya urn representation. First, the $\beta' s$ are drawn from a mixture of the baseline distribution and a discrete distribution. Second, if $\beta_r = \beta$ for all r, then β is drawn from the centering distribution with probability one, and therefore the base distribution is the prior. Finally, $\Pr ob(\beta_r = \beta_s, r \neq s) > 0$, resulting in the clustering property of the Dirichlet process (MacEachern, 1994). The n $\beta' s$ are grouped into k sets, $0 < k \leq n$, with all observations in a group sharing the same value of β, and observations in different groups have different values of β.

Another representation is the stick-breaking prior representation (Sethuraman 1994; Ishwaran and James, 2001) and is given by

$$G = \sum_{r=1}^{\infty} p_r \delta_{\theta_r} \ , \tag{4}$$

where δ_r is the Diriac measure which places measure 1 on the point t, $\theta_1, \theta_2, \ldots$ are i.i.d. realizations of H, and $p_r = V_r \prod_{l<r}(1-V_l)$ where V_r are i.i.d. Beta(1, M). Then θ_r are referred to as locations V_r as masses and p_r as the respective weights.

By the definition of the stick-breaking representation, the weights $p_r = V_r \prod_{l<r}(1-V_l)$ tend to be large for small r (recall that the masses V_r are Beta (1, M) random variables so if r is large, many of the $(1-V_l)$ will be multiplied by the weight p_r, thus making its value small

2.2. Introducing dependence in the Dirichlet Process

In many settings, one might be interested in allowing the unknown distribution G as defined above to depend on some covariate W, which could be time, space, or other known covariates. Several papers in the recent literature have extended the Dirichlet process to accommodate this dependence and are all based on the Sethuraman (1994) representation of the DP.

MacEachern (1999, 2000) introduced a dependent Dirichlet process (DDP) by replacing either the masses, V_r, or the locations, θ_r, of the stick-breaking representation by stochastic processes. MacEachern et al. (2001) focused on a model where only the locations are stochastic processes. Their model is referred to as “single p” model and has been applied to spatial modeling by Gelfand et al. (2004) and Duan et al. (2007), ANOVA-like models for densities by De Ioro et al. (2004), and quantile regression by Kottas and Krnjajic (2005).

Griffin and Steel (2006) suggested the order-based dependent Dirichlet process (πDDP) that captures nonlinear relationships between the unknown distribution G and covariate W. Dependence is introduced by making the masses, V_r, and the locations, θ_r, of the stick-breaking representation (4) depending on the covariate W. Specifically, the elements of the vectors V and θ are ranked via an ordering $\pi(W)$. At each covariate W, we still have the stick-breaking representation (4) (marginally G_W is a DP) but the order in which the masses are combined varies over the covariate domain:

$$G_W \overset{d}{=} \sum_{r=1}^{\infty} p_r(W)\,\delta_{\theta_{\pi_r}(W)}, \tag{5}$$

where δ_k denotes the Dirac measure at k, $p_r(W) = V_{\pi_r(W)} \prod_{l<r} (1 - V_{\pi_l(W)})$ with $\theta_k \overset{iid}{\sim} H$,

$V_k \overset{iid}{\sim} Beta(1, M)$, and $\sum_{k=1}^{\infty} p_k(W) = 1 \;\; a.s.$

Here $\pi(W)$ defines an ordering at the covariate value W and satisfies the following condition

$$|W - z_{\pi_1(W)}| < |W - z_{\pi_2(W)}| < |W - z_{\pi_3(W)}| < \cdots,$$

where z is the realization of a Poisson process with intensity λ. In other words, the ordering $\pi(W)$ lists the z_r in increasing order of absolute distance from W so that the most relevant z_r at W are those close to W. An index r that appears "late" in the ordering $\pi(W)$ (i.e., for which l such that $\pi_l(W) = r$, is high) would have many terms $(1 - V_{\pi_{l'}(W)})$ multiplied into its weight p_r. An infinite number of z_r appears over the infinite real line but only the z_r close to the observed covariate value would have significant weight. For practical computation, truncation of the point process similar to truncation of the stick-breaking representation is defined.

We now turn to the description of how the DP at two distinct covariate values are correlated. As mentioned previously, the marginal distribution of the πDDP at any covariate value follows a DP: $G_W \sim DP(M,H)$. Correlation of two distributions G_{W_1} and G_{W_2} depends on the order in which the masses V_r are combined at the covariate values W_1 and W_2. The intensity parameter λ controls how quickly the indexes r change. A large value of λ yields more densely packed indexes, causing the ordering $\pi(W)$ to change more quickly from one covariate to another, and consequently the G_w will be less correlated. The value of M controls the expected number of indexes with significant masses. A large value of M makes more leading terms in the stick-breaking relevant and thus implies more indexes need to change place in the ordering before the distributions decorrelate. Thus the intensity parameter λ and the precision parameter M control the correlation between the distributions G_{W_1} and G_{W_2}. Griffin and Steel defined the explicit expression of the correlation between the distributions G_{W_1} and G_{W_2} as:

$$Corr(G_{W_1}, G_{W_2}) = \left(1 + \frac{2\lambda |W_1 - W_2|}{M+2}\right) \exp\left\{\frac{-2\lambda |W_1 - W_2)}{M+1}\right\}, \tag{6}$$

where $|W_1 - W_2|$ denotes the distance between W_1 and W_2.

Like the Dirichlet process, the πDDP produces discrete realizations. To obtain continuous distributions, the πDDP is imbedded in the hierarchical model as follows:

Stage 1: $[y_i \mid \beta_i] \sim f(\beta_i)$,

Stage 2: $\beta_i \sim G_W$,

Stage 3: $G_W \sim \pi DDP(M, H, \lambda)$,

where H is the baseline distribution, M is the precision parameter, and λ is the intensity of the Poisson point process that induces the orderings.

In the following two sections, we apply the $\pi\,DDP$ model to discrete choice models. The advantage of the $\pi\,DDP$ over the "single p" DDP is that it allows dependence to be introduced on both the weights and the atoms. We derive the full conditional distributions and the MCMC sampler for fitting the models. Section 3 presents the case where the discrete choice model is estimated with individual level data. In section 4 we extend the model in 3 to account for endogeneity and allow estimation with aggregate data.

3. Dependent Dirichlet Process priors in discrete choice models with individual level data

3.1. The model

Assume we have n individuals, each making purchase decisions over T periods, and we observe the choices made by all consumers. In each period, each individual chooses one alternative from a set of J alternatives. Define the following notations:

$\mathrm{y}_{\mathrm{it}} = \mathrm{j}$ denotes the event that individual i chooses alternative j at time t,

$\mathrm{x}_{\mathrm{ijt}}$ denotes a p-dimensional vector of observed characteristics (price, brand indicator variable, and other product characteristics) of alternative j for individual i in period t,

β_{i} denotes the p-dimensional vector of parameters for individual I,

ε_{ijt} represents random variation in consumer choice behavior.

The utility individual i derives from choosing alternative j at time t is parameterized as

$$u_{ijt} = \mathrm{x}_{ijt}^{'}\beta_{\mathrm{i}} + \varepsilon_{ijt}\,. \tag{7}$$

Assuming ε_{ijt} has an Extreme value (0, 1) distribution, the probability that individual i chooses alternative j in period t is given by

$$p_{ijt} = p(y_{it} = j) = \frac{\exp(\mathrm{x}_{ijt}^{'}\beta_{\mathrm{i}})}{\sum_{k=1}^{J}\exp(\mathrm{x}_{ikt}^{'}\beta_{\mathrm{i}})}\,,\ i = 1,\ldots,n\,,\ j = 1,\ldots,J\,,\ t = 1,\ldots,T\,. \tag{8}$$

Alternatively, ε_{ijt} can be assumed to be drawn from a normal distribution, giving rise to the multinomial probit model. However, the model with logit disturbance has the advantage of yielding close form choice probabilities as in (8) and is easier to implement than the probit model. Moreover, the probit model may not accommodate a large number of products (Chintagunta, 2001). These reasons explain why the logit model is widely used.

The likelihood of individual i's choices over time is then given by

$$p(D_i \mid \beta_i) = \prod_{t=1}^{T} \prod_{j=1}^{J} p_{ijt}^{\;D_{ijt}} \,, \tag{9}$$

where $D_{ijt} = 1$ if $\mathrm{y_{it}} = \mathrm{j}$ and 0 otherwise, and $\mathrm{y_i} = (y_{it}, \ldots, \mathrm{y_{iT}})^{'}$.

Our model in (7) assumes the β_i s are heterogeneous across individuals. We want to model the β_i using a nonparametric distribution while at the same time allowing this distribution to depend on individual characteristics. To accomplish this, we use the mixture of order-based dependent Dirichlet Process model described above. The resulting model can be written in hierarchical form as:

$$p(y_i \mid \beta_i) = \prod_{t=1}^{T} \prod_{j=1}^{J} p(y_{it} = j \mid \beta_i)^{D_{ijt}} \,,$$

$$\beta_i \sim G_W \,,$$

$$G_W \sim \pi DDP(M, H, \lambda) \,.$$

There are two properties of the order-based dependent Dirichlet process G_W which give insight into heterogeneity in our model. First, like the Dirichlet process introduced by Ferguson (1973), G_W creates clusters of observations in the data. Because there is a positive probability of individuals to share regression parameters, there will be $L \le n$ distinct values of the regression parameters $\beta_1, \ldots \beta_n$. Second, because G_W varies with subject characteristics, the distribution of individuals across the L clusters depends on subjects characteristics, and this relationship is not restricted to be linear.

3.2. Prior distributions for M, H, and λ

Following Griffin and Steel (2004, 2006), we specify the prior distribution for M as an inverted Beta distribution

$$p(M) = \frac{n_0^{\eta}\Gamma(2\eta)}{\Gamma(\eta)^2}\frac{M^{\eta-1}}{(M+n_0)^{2\eta}},$$

where the hyperparameter $n_0 > 0$, the prior median of M and the prior variance of M (which exists if $\eta > 2$) is a decreasing function of η.

Other prior distributions for M have been suggested in the literature; Escobar and West (1995) suggested a gamma distribution whose parameters are elicited by considering the distribution of the number of distinct elements in the first n draws from the Dirichlet process. Walker and Mallick (1997) used the formula $M = E(\omega^2)/Var(\mu)$, where μ and ω^2 are the mean and variance of the unknown distribution. In their inverted Beta distribution, Griffin and Steel interpreted M as a “prior sample size”, because of the form of the Dirichlet process prior predictive distribution derived by Blackwell and MacQueen (1973).

The prior distribution for λ depends on the precision parameter M, the autocorrelation function, and type of construction used to induce the ordering to vary with the covariate W. Using the permutation construction and assuming a one-dimensional covariate, the distribution of λ is

$$p(\lambda) = \frac{2t^*(2t^*\lambda+1)}{(M+1)(M+2)}\exp\left\{-\frac{2t^*}{M+1}\lambda\right\},$$

where t^* is a parameter to be tuned. It is worth mentioning that for more than one covariate, the prior on λ has no closed form (see Griffin and Steel, 2006) and can only be approximated numerically.

For the centering distribution H, we specify a p-variate normal distribution with unknown mean vector μ_H and unknown covariance matrix Σ_H,

$$H \mid \mu_H, \Sigma_H = MVN(\mu_H, \Sigma_H).$$

3.3. Bayesian estimation

To complete the model specification, we assume the following prior distributions for the mean vector and covariance matrix of the baseline distribution H:

$$\mu_H \sim N_p(\mu_0, V_0), \text{ and}$$

$$\Sigma_H \sim IW_p(\upsilon_{H0}, S_{H0}),$$

where $N_p(\mu_0, V_0)$ denotes a p-dimensional Normal distribution with mean vector μ_0 and covariance matrix V_0, and $IW_p(\upsilon_{\Sigma 0}, S_{\Sigma 0})$ denotes a p-dimensional inverted Wishart distribution with parameters υ_{H0}, and S_{H0}; μ_0, V_0, υ_{H0}, and S_{H0} are known.

In addition to the parameters $\{\beta_i\}$, M, λ, Σ_H, and μ_H, the point process z needs to be sampled. The joint posterior distribution of all model parameters is

$$f(\{\beta_i\}, M, \lambda, z, \Sigma_H, \mu_H \mid Y, X) \propto \left(\prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{j=1}^{J} p_{ijt}^{\;D_{ijt}}\right) \times \left(\prod_{i=1}^{n} \pi_1(\beta_i \mid G_W)\right) \\ \times \pi_2(G_W \mid H, M, \lambda)\ \pi_3(M \mid \lambda) \pi_4(\lambda \mid M, z) \pi_5(\Sigma_H, \mu_H), \tag{10}$$

where p_{ijt} is given in (8), π_1 is the distribution of the regression parameters, π_2 is the Dirichlet process prior on this distribution, π_3 is the distribution of the precision parameter that depends on the intensity parameter λ, π_4 is the distribution of the intensity parameter that depends on the point process z and the precision parameter M, and π_5 is the prior distribution on the parameters of the baseline distribution.

Define the n-dimensional vector C such that $\beta_i = \theta_{C_i}$. The model parameters are estimated via a Markov Chain Monte Carlo algorithm that generates draws from the following sequence and conditional distributions:

(1) Update C,

(2) Update θ,

(3) Update z,

(4) Update M,

(5) Update λ,

(6) Update μ and Σ.

We discuss each of these conditional distributions in turn but before that we define some notations. Suppose $I = \{1,..,n\}$ is the set of all the n individuals; for a subset B of I, $n_l(B)$ represents the number of individuals i in B for which $C_i = l$ and

$Q_l(B) = \#\{i \in B \text{ such that there exists k} < \text{l for which } \pi_k(W_i) = l, where\ \pi_j(W_i) = C_i\}$. That is, $Q_l(B)$ is the number of observations for which l appears before C_i in the ordering at W_i.

Next we follow the following steps:

(1) Generation of C

Propose C according to the following discrete distribution

$$p(C_i = l \mid C_{-l}, M, z, W, D) \propto p(D_i \mid C_i = l, C_{-i}, D_{-i}) p(C_i = l \mid C_{-l}, M, z, W)$$
$$\propto \frac{\int f(D_i \mid \beta) \prod_{\{j \neq i \mid C_j = l\}} f(D_j \mid \beta) dH(\beta)}{\int \prod_{\{j \neq i \mid C_j = l\}} f(D_j \mid \beta) dH(\beta)} \times \frac{n_l(I_{-i}) + 1}{M + Q_l(I_{-i}) + n_l(I_{-i}) + 1} \times \prod_{j < m(l)} \frac{M + Q_{\pi_j(W)}(I_{-i}) + 1}{M + Q_{\pi_j(W)}(I_{-i}) + n_{\pi_j(W)}(I_{-i}) + 1},$$

where $\pi_{m(l)}(W) = l$.

The above expression assumes that clusters are numbered in the order they appear; this implies that for an individual to be allocated to cluster l, it must be true that she is not allocated to clusters appearing before l. Clearly, $\frac{n_l(I_{-i})+1}{M+Q_l(I_{-i})+n_l(I_{-i})+1}$ is the probability that the individual i is allocated to cluster l given that she can only be allocated to clusters l, $l+1$, … L, whereas $\frac{M+Q_{\pi_j(W)}(I_{-i})+1}{M+Q_{\pi_j(W)}(I_{-i})+n_{\pi_j(W)}(I_{-i})+1}$ is the probability that the same individual is not allocated to cluster $\pi_j(W)$.

(2) Generation of θ.

Propose θ form the distribution

$$p(\theta \mid C, D, W) \propto H(\theta_l)d\theta_l \times \prod_{\{i:C_i=l\}} prob(D_i \mid \theta_l,.) .$$

A slice sampler (Neal, 2003) can be used to sample from this distribution.

Given the draws of C and θ, the n-dimensional vector of individual specific parameters β are given by $\beta_i = \theta_{C_i}$.

(3)** Generation of z**.

To update the point process z, we use the "move a current point" update in Griffin and Steel. Assume that the current relevant elements of the Poisson process are $z = (z_1, \cdots, z_L)$.The "move a current point update" consists of choosing at random a point z_u and adding to it a random variable with zero mean and a tuning variance. The obtained moved $z_u^{'}$ is rejected if it falls outside the truncation region or is accepted with probability

$$\min\left\{1, \prod_{u=1}^{L} \frac{n_u(I)+1+Q_u^{'}(I)+M}{n_u(I)+1+Q_u{'}(I)+M}\right\}.$$

(4) Generation of λ.

The conditional distribution for the intensity parameter (λ) depends on the point process z. Sampling λ proceeds as follows for a one-dimensional W:

- For each point of the Poisson process z_u, attach a mark m_u which is uniformly distributed on (0, 1);
- Draw a proposed value $\log \lambda^{'} \sim N(\log \lambda, \sigma_{\lambda}^2)$; if $\lambda^{'} < \lambda$ the points in the data region for which $m_u > \lambda^{'} / \lambda$ are removed from the point process, otherwise $m_u^{'} = m_u \lambda / \lambda^{'}$; if $\lambda^{'} > \lambda$, a new point process with intensity $\lambda^{'} - \lambda$ is drawn in the data region.

(5) Generation of M.

Recall that the mass parameter (M) and the ordering process $\pi(W)$ determine the dependence across the covariate domain, and the number of points in the truncated domain depends on M. To update the value of M, we draw a new point M' such that $\log M^{'} \sim N(\log M, \sigma_M^2)$, where σ_M^2 is chosen to control the overall acceptance rate.

If $M^{'} > M$, the truncated region is expanded and the unobserved part of the Poisson process is sampled;

If $M^{'} < M$, the truncated region is contracted and points that fall outside the region are removed. If these points have any observations allocated to them, the new point is rejected.

Griffin and Steel define the above move as a reversible jump move where extra points are sampled from the prior distribution. The acceptance rate given by

$$\frac{M^{'}}{M}\frac{p(M^{'} \mid \lambda)}{p(M \mid \lambda)}\prod_{u=1}^{U}\frac{n_u + 1 + Q_u^{'} + M)}{n_u + 1 + Q_u + M^{'}}.$$

(6) Generation of μ *, and* Σ.

The full conditional distributions for μ_H and Σ_H reduce to

$$\mu_H \mid \beta^*, \Sigma_H \sim N_P(\mu^*, V^*) \text{ and } \Sigma_H \mid \beta^*, \mu_H \sim IW\left(L + \upsilon_H, S_H + \sum_{k=1}^{L}(\beta_k^* - \mu_H)(\beta_k^* - \mu_H)^{'}\right),$$

where $\mu^* = V^*(V_0^{-1}\mu_0 + \sum_{l=1}^{L}\Sigma_H^{-1}\beta_l^*)$ and $V^* = (V_0^{-1} + L\Sigma_H^{-1})^{-1}$;

μ_H and Σ_H are sampled using direct Gibbs sampling.

Computing marginal effects (elasticities)

Recall that probability for consumer i choosing brand j at time t is

$$p_{ijt} = p(y_{it} = j) = \frac{\exp(\mathrm{x}_{ijt}^{'}\beta_{\mathrm{i}})}{\sum_{k=1}^{J}\exp(\mathrm{x}_{ikt}^{'}\beta_{\mathrm{i}})}.$$

Assuming consumers do not make multiple purchases, the market share of brand j at time t is $s_{jt} = \frac{\sum_i s_{ijt}}{n}$.

Elasticities (percent change in the probability of choosing an alternative for a given change in one of the observed product characteristics $x_{ijt,r}$, holding the other product characteristics constant) are calculated as follows:

$$\eta_{jt,r} = \begin{cases} \frac{1}{n\, s_{jt}} \sum_i \beta_{i,r} s_{ijt} (1 - s_{ijt})\, x_{ijt,r} & if\ l = j \\ \frac{1}{n\, s_{jt}} \sum_i \beta_{i,r} s_{ijt} s_{ilt} x_{ilt,r} & if\ l \neq j, \end{cases}$$

where $\beta_{i,r}$ and $x_{ijt,r}$ are the l^{th} component of β_i and x_{ijt}, respectively.

4. Dependent Dirichlet Process priors in discrete choice models with aggregate data and/or endogeneity

Very frequently in Marketing and Economics, the utility model in (7) includes an unobserved demand shock ξ_{jt} for each brand j and time t, which is assumed to be correlated with prices, thus creating an endogeneity problem. Also discrete choice models are estimated with aggregate (store, chain, or market level) data in some product categories because individual level data are not available. In this subsection we extend the model of section 3 to account for price endogeneity and allow estimation with aggregate data.

4.1. The model

Assume we observe aggregate market shares, prices, and product characteristics of J brands across T periods of time. We assume the observed market shares are generated by N individuals, each making choice decisions over T periods. The utility that each individual derives from choosing brand j in period t is defines as

$$u_{ijt} = x'_{jt}\beta_i - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt} \tag{11}$$

where x_{jt} and p_{jt} are respectively observed product characteristics and price of brand j at time t; they are the same for all consumers; β_i and α_i represent consumer-specific tastes for product characteristics. Further ξ_{jt} represents the effects of variables other than price and observed product characteristics contained in x_{jt} that are not included in the model and that could affect the probability of choosing brand j. It is assumed to be observed by the consumers and the manufacturers, but not by the econometrician. Here ε_{ijt} represents random variation in consumer choice behavior and is assumed to have an extreme value (0, 1) distribution.

The objective is to estimate the parameters β_{i} and α_{i} form the observed aggregate market shares, prices and product characteristics.

Denoting $\Theta_{\mathrm{i}}=(\beta_i,\alpha_i)$, $\xi_{\mathrm{t}}=(\xi_{1t},\ldots,\xi_{Jt})$, the probability that individual i chooses alternative j in period t is given by

$$p_{ijt}=\Pr ob(y_{it}=j\,|\,\xi_t,P_t,\Theta_i)=\frac{\exp(\mathrm{x}_{jt}^{'}\beta_{\mathrm{i}}-\alpha_i P_{jt}+\xi_{jt})}{\sum_{k=1}^{J}\exp(\mathrm{x}_{kt}^{'}\beta_{\mathrm{i}}-\alpha_i P_{kt}+\xi_{kt})},\ i=1,\ldots,n,\ j=1,\ldots,J,\ t=1,\ldots,T. \tag{12}$$

As previously defined, let D_{ijt} takes a value of 1 if consumer i chooses brand j in period t, and a value of 0 otherwise. We do not observe the individual choices D_{ijt}, but only aggregate share S_{jt} for each brand in period t. We want to augment observed aggregate shares S_{jt} with the latent individual choices D_{ijt} so that at the aggregate level, the sum of latent individual choices are consistent with the observed shares at each time period (i.e., $\sum_{i=1}^{n}D_{ijt}=nS_{jt}$), and at the individual level, augmented choices are consistent with utility functions across time periods.

For each consumer i, the likelihood of observing choices at purchase occasion 1,…,T is

$$p_i=prob(D_i\,|\,\Theta_i,\xi,P)=\prod_{t=1}^{T}\prod_{j=1}^{J}\Pr ob(y_{it}=j\,|\,\xi_t,P_t,\Theta_i)^{D_{ijt}}. \tag{13}$$

Thus the likelihood for observing choice sequences of all the n consumers, $\{D_i\}_{i=i}^{n}$, is then given by

$$prob(\{D_i\}_{i=i}^{n} \mid \{\Theta_i\}_{i=1}^{n}, \xi, P) = \prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{j=1}^{J} I_{\{\sum_{i=1}^{n} D_{ijt}=nS_{jt}\}} \Pr ob(y_{it} = j \mid \xi_t, P_t, \Theta_i)^{D_{ijt}} , \tag{14}$$

where the indicator function ensures that the augmented individual choices D_{ijt} are exactly consistent with the aggregate market shares.

There is a potential for correlation between prices and unobserved product characteristics ξ_{jt} because manufacturers observe the ξ_{jt}'s and demand for brand j depends on ξ_{jt}; this makes prices endogenous. We account for endogeneity by using instrumental variables techniques (Villas-Boas and Winer, 1989). We assume

$P_{jt} = \varphi\chi_{jt} + o_{jt}$, $o_t \sim MVN(0, \Sigma_o)$ and $\operatorname{cov}(\xi_t, o_t) = \Sigma = \begin{pmatrix} \Sigma_\xi & \Sigma_{\xi o} \\ \Sigma_{o\xi} & \Sigma_o \end{pmatrix}$,

where χ_{jt} represents a vector of instrumental variables.

As before, we want to model the Θ_i using a nonparametric distribution while at the same time allowing this distribution to depend on consumer characteristics. To accomplish this, we use the order-based dependent Dirichlet Process model. The hierarchical form of the model is given by

$$\Pr ob(D_{it} \mid \Theta_i, \xi_t, P_t) = \prod_{j=1}^{L} \Pr ob(y_{it} = j \mid \xi_{jt}, P_{jt}, \Theta_i)^{D_{ijt}} ,$$

$P_t \mid \{\xi_t\}, \varphi, \{\Theta_i\} \sim N(\varphi\chi_t, \Sigma_0)$,

$\xi_t \sim N(0, \Sigma_\xi)$,

$\varphi \sim N(0, \sigma_{0\varphi}^2 I)$,

$\Theta_i \sim G_W$,

$G_W \sim \pi DDP(M\, H, \lambda)$.

4.2. Identification

Since only aggregate data are available, it is important to discuss how the model parameters are identified by the aggregate data. Identification comes from examining the time patterns of the observed aggregate brand shares. The goal of the model is to estimate the distribution of consumer individual-specific parameters Θ_i, the covariance between the demand shocks and the prices, Σ, and the price equation parameter φ. By assuming that each Θ_i is drawn from a distribution that does not have a parametric form but has the order-based dependent Dirichlet process prior, $\pi DDP(M, H, \lambda)$, with precision parameter M, intensity parameter λ, and baseline distribution H assumed to be normally distributed with mean μ_H and covariance Σ_H, the goal reduces to the estimation of M, μ_H, Σ_H, λ, Σ, and φ from the aggregate brand shares. If each of these parameters induces different behavior of the aggregate brand shares through time, then the model is identified[1]. We discuss each parameter in turn.

Recall that the parameters M and λ control the correlation of the order-based dependent Dirichlet process at different values of the covariates. Larger values of M and λ cause the marginal Dirichlet processes to decorrelate faster, thus increasing the number of distinct clusters, with consumers having similar covariates sharing the same cluster. More distinct clusters mean there is heterogeneity in consumers' preferences for product characteristics (price, brand indicators, other product characteristics). For example one cluster may include consumers that have high income, are loyal to a given brand and are less price-sensitive, while another cluster is made up of low income, highly price-sensitive consumers. If many consumers are loyal to a given brand, changing the price of that brand would not decrease its market share overtime. On

[1] In addition to not being restricted to a parametric family, the Dirichlet process has another advantage over a finite mixture model (e.g., finite mixture of normals); as a random mixing distribution, it is more parsimonious than a finite mixture model which involves a large number of parameters which may not be identifiable with aggregate data.

the other hand if few consumers are loyal to that brand, its markets share would tend to decline with a price increase. There is also a situation where the negative effect of price due to price-sensitivity of some consumers compensates the positive effect due to the behavior of other consumers, thus leaving a less noticeable variation of market shares over time.

The price equation parameter φ, the off-diagonal blocks and the lower diagonal block of the covariance matrix Σ are identified by the exogenous variations of the instrumental variables over time.

The upper diagonal block of Σ, Σ_{ξ}, represents the covariance matrix of the unobserved demand shocks ξ_{jt}. Since these demand shocks capture the effect of unobserved demand factors on aggregate demand, a higher value of any of its diagonal element would indicate high volatility of the market share of the corresponding brand. An off-diagonal elements $\Sigma_{\xi}(j, j')$ measures the similarity of the utilities of brands j and j' over time with respect to demand shocks. Therefore, a high value of $\Sigma_{\xi}(j, j')$ implies an identical effect of a demand shock on the shares of brand j and j', but a different effect on the shares of the remaining brands, thus leading to different market shares patterns over time.

4.3. Bayesian estimation

Lacking observed information on individual choices D_{ijt}, a data augmentation approach (Tanner and Wong, 1987; Albert and Chib, 1993; Chen and Yang, 2004, Musalem et al., 2005) will be used. Instead of integrating out individual choices (D_{ijt}) and individual level response parameters $\Theta_{i} = (\beta_{i}, \alpha_{i})$ as in the non-likelihood based approach (Berry, Levinsohn and Pakes

(1995)), we treat them as any other unobserved model parameters and use them as conditioning arguments in generating the draws.

The prior distribution for μ_H, Σ_H, and Σ are assumed to be

$\mu_H \sim N(\mu_0, V_0)$,

$\Sigma_H \sim IW(\upsilon_{\Sigma_H 0}, S_{\Sigma_H 0})$, and

$\Sigma \sim IW(\upsilon_{\Sigma 0}, S_{\Sigma 0})$.

In the above specifications, $\sigma^2_{\varphi 0}$, μ_0, V_0, $\upsilon_{\Sigma_H 0}, S_{\Sigma_H 0}$, $\upsilon_{\Sigma 0}$, and $S_{\Sigma 0}$ are known.

The joint posterior distribution of all model parameters is

$$\begin{aligned} & f(\{\Theta_i\}, z, M, \lambda, \Sigma_H, \Sigma, \varphi, \{D_{ijt}\}, \{\xi_t\}, P \mid S, X, \chi, W) \\ & \quad \propto \left(\prod_{i=1}^{n} \prod_{t=1}^{T} \prod_{j=1}^{J} I_{\{\sum_{i=1}^{n} D_{ijt} = nS_{jt}\}} \, p_{ijt}^{\;D_{ijt}} \right) \pi_1(\{\xi_t\} \mid \Sigma) \pi_2(\{P_t\} \mid \Sigma, \{\xi_t\}) \times \left(\prod_{i=1}^{n} \pi_3(\Theta_i \mid G_W) \right) \\ & \quad \times \pi_4(G_W \mid H, M, \lambda) \pi_5(z) \pi_6(M) \pi_7(\lambda) \pi_8(\mu_H, \varphi, \Sigma, \Sigma_H), \end{aligned} \tag{15}$$

where p_{ijt} is defined in (12), and S, P, X, χ, and W are matrices of observed market shares, prices, product characteristics, instrumental variables, and consumer characteristics.

The model parameters are estimated via a Markov chain Monte Carlo algorithm that generates draws from the following sequence and conditional distributions:

(1) Sample ξ_t, $t = 1,...,T$,

(2) Sample D_t, $t = 1,...,T$,

(3) Sample C,

(4) Sample θ,

(5) Sample z,

(6) Sample M,

(7) Sample λ,

(8) Sample μ_H and Σ_H ,

(9) Sample φ and Σ.

Steps (3)-(8) are the same as in section 3; therefore us we only discuss steps (1), (2) and (9).

Generation of ξ.

The full conditional distribution for ξ_t is given by

$$f(\xi_t \mid .) \propto \left(\prod_{i=1}^{n} \prod_{j=1}^{J} p_{ijt}^{\ D_{ijt}} \right) \pi_5(\{\xi_t\} \mid \Sigma)\pi_6(\{P_t\} \mid \Sigma, \{\xi_t\})$$

ξ_t is sampled using a random walk Metropolis-Hastings sampling.

Generation of D_t.

We sample individual choices using a multiple-block Metropolis-Hastings algorithm. Because of the large number of consumers, convergence can be very slow if the single block algorithm is used. We randomly partitioned the set of consumers into b blocks $D_{1t},\ldots D_{bt}$, each of size m. Each block is sequentially updated using the following algorithm:

- Specify an initial value $D_t^{(0)} = (D_{it}^{(0)},\ldots D_{bt}^{(0)})$,
- Repeat for $k = 1,\ldots,b$.

(i) Propose a value for the kth block, D_{kt}^{new}, conditioned on the current value of the other blocks D_{-kt} from the discrete distribution

$$q_k(D_{kt}^{new} \mid D_{-kt}) = \frac{1}{C_m^{O_{k0t}} C_{m-O_{k0t}}^{O_{k1t}} \ldots C_{m-O_{k0t}-\ldots-O_{kJ-1,t}}^{O_{kJt}}} \prod_{j=1}^{J} I_{\{\sum_{i=1}^{m} D_{kijt} = O_{kjt}\}} ,$$

where $C_A^a = \dfrac{A!}{A!(A-a)!}$ and $C_m^{O_{k0t}} C_{m-O_{k0t}}^{O_{k1t}} \ldots C_{m-O_{k0t}-\ldots-O_{kJ-1,t}}^{O_{kJt}}$ is the total number of combinations of D_{kijt} that satisfy the constraint $\sum_{i=1}^{m} D_{kijt} = \mathrm{O}_{jt}$ for all j; $O_{kjt} = O_{jt} - O_{-kjt}$, where O_{jt} is the integer approximation of nS_{jt} and O_{-kjt} is the number of consumers in the other blocks that have chosen brand j in period t.

To generate a candidate draw D_{kt}^{new} from q_k, first randomly assign O_{k0t} consumers to the no purchase alternative, then O_{k1t} consumers among the remaining $m - \mathrm{O}_{k0t}$ to brand choice 1, and so on until all consumers are allocated.

(ii) Calculate the probability of the move

$$\alpha_k(D_{kt}^{new}, D_{kt}^{old} \mid D_{-kt}) = \min\left\{1, \frac{p(D_{kt}^{new}) q_k(D_{kt}^{old} \mid D_{-kt})}{p(D_{kt}^{old}) q_k(D_{kt}^{new} \mid D_{-kt})}\right\}$$

where $p(D_{kt}) = \prod_{i=1}^{m} \prod_{j=1}^{J} I_{\{\sum_{i=1}^{m} D_{kijt} = O_{kjt}\}} \, p_{ijt}^{D_{kijt}}$.

(iii) Update the kth block with probability $\alpha_k(D_{kt}^{new}, D_{kt}^{old} \mid D_{-kt})$.

Generation of φ and Σ.

The full conditional distributions for Σ and φ reduce to

$$\Sigma \mid . \sim IW\left(T + \upsilon_{\Sigma 0}, S_{\Sigma 0} + \sum_{t=1}^{T} \begin{pmatrix} \xi_t \\ P_t - \varphi \chi_t \end{pmatrix} \begin{pmatrix} \xi_t \\ P_t - \varphi \chi_t \end{pmatrix}'\right),$$

$$\varphi \mid . \sim MVN(A, B)$$

where $A = \Psi \chi' \Delta^{-1}(P - f)$, $\Psi = (\Lambda^{-1} + \chi' \Delta^{-1} \chi)^{-1}$, $f = \Sigma_{\xi o} \Sigma_{\xi}^{-1} \xi$, $\Delta = \Sigma_o - \Sigma_{\xi o} \Sigma_{\xi}^{-1} \Sigma_{o\xi}$, and $\Lambda = \sigma_{\kappa 0}^2 I$.

Then Σ and φ are sampled using direct Gibbs sampling.

Computing marginal effects (elasticities)

Price and advertising elasticities for each chain-period are computed as follows:

The conditional probability for consumer i choosing brand j at time t is

$$s_{ijt} = \Pr ob(y_{it} = j \mid \xi_t, P_t, \Theta_i) = \frac{\exp(\mathrm{x}'_{jt}\beta_i - \alpha_i P_{jt} + \xi_{jt})}{\sum_{k=1}^{J}\exp(\mathrm{x}'_{kt}\beta_i - \alpha_i P_{kt} + \xi_{kt})}.$$

Assuming consumers do not make multiple purchases, the market share of brand j at time t is $s_{jt} = \frac{1}{n}\int \sum_i s_{ijt} f(\xi_{jt}) d\xi_{jt}$.

Price elasticities are calculated as follows:

$$\eta_{jlt} = \frac{\partial s_{jt}}{\partial p_{lt}} \frac{p_{lt}}{s_{jt}} = \begin{cases} \dfrac{p_{lt}}{s_{jt}} \dfrac{1}{n} \sum_i \alpha_i s_{ijt}(1 - s_{ijt}) & if\ l = j \\ -\dfrac{p_{lt}}{s_{jt}} \dfrac{1}{n} \sum_i \alpha_i s_{ijt} s_{ilt} & if\ l \neq j. \end{cases}$$

5. Empirical applications

5.1. Discrete choice models with individual level data

The model with individual data is estimated with an A.C. Nielson supermarket scanner dataset for peanut butter in the city of Sioux Falls, South Dakota. The objective is to assess the distribution of consumer preferences and investigate how these preferences vary with income (here our covariate W is the income).

The data was obtained from the publicly available ERIM database at the University of Chicago Graduate School of Business. We observe consumers and their choices. The number of household is 326 and the total number of purchase is 9158. There are J=4 brands of peanut butter. The product characteristics include a dummy variable for featured advertising, net price, and three dummy variables for brands 1, 2 and 3. Table 1 summarizes these variables.

TABLE 1 ABOUT HERE

The following values are chosen for the priors: $\sigma^2_{\varphi 0} = 100$, $\mu_0 = 0$, $V_0 = S_{\Sigma_H 0} = S_{\Sigma 0} = 100I$, and $\upsilon_{\Sigma_H 0} = \upsilon_{\Sigma 0} = 2$. The MCMC sampler was run for 15 000 iterations, the first 500 being discarded as burn-in period. To assess convergence, we use different starting points for the chain and examine the trace plots of the model parameters (not shown).

The nonparametric approach to modeling heterogeneity as described aims at relaxing the unimodality assumption in the distribution of the individual-specific parameters, and the linearity of the relationship between consumer-specific parameters and consumer characteristics. Figure 1 plots the posterior density function of the precision parameter M, which, recall, measures the suitability of a parametric model for the individual-specific parameters; values close to zero suggests the parametric model is inadequate. From figure 1, it appears that most of the values of M are close to 0.5, indicating that the normal centering distribution is very inadequate for the data.

FIGURE 1 ABOUT HERE

Figure 2 shows the posterior distributions of price sensitivity, advertising intensity, and brand indicators. On the left are displayed the density plot for each parameter, obtained by standard kernel density estimation with window width computed following the recommendation of Silverman (1986). On the right, the relationship between preferences and income is plotted using the Nadaraya-Watson regression estimation method. The density plots reveal that distributions of individual parameters are non-normal. The conditional density plots further show that the relationship between individual-specific parameters and income is nonlinear. It is common knowledge that high income household are less price sensitive than low income households; the conditional density plot for price shows that this is true only for income above $65,000.

FIGURE 2 ABOUT HERE

5.2. Discrete choice models with aggregate data

The model with aggregate data is applied to a ready-to-eat breakfast cereal dataset. The data were obtained from the Food Marketing Policy Center at the University of Connecticut and is of two types: dollar sales and volume sales measured every four weeks at three supermarket chains in Baltimore, Boston, and Chicago, and household income distribution in each supermarket chains trading areas.

The period of study is January 8, 1996-December 7, 1997. During this period, cereal manufacturers introduced many brands but we focus only on four major brands that were introduced between January 1996 and March 1997, so that each brand is observed for a relatively long time period. These are: Kellogg's Honey Crunch Corn Flakes, General Mill French Toast Crunch, Kellogg's Cocoa Frosted Flakes, and Post Cranberry Almond Crunch. In addition to the four new brands, the analysis includes 27 established brands. The chain-level share of these established brands varies between 35 and 80 percents of the total volume of cereal sold at each supermarket chain and quad period. Moreover, these brands are the leading established brands in the 4 cereal segments: all family, taste enhanced wholesome, simple health nutrition, and kids cereals.

The variables used in the analysis include brand's market share, price, and observed product characteristics (calories, fiber, sugar content), and household income. We do not observe consumers and their choices, but only the shares of each cereal brand at each supermarket chain in each period, and the distribution of household income in the trading area of each supermarket chain.

Market shares of the brands under consideration are defined by converting the volume sales into servings sold, and dividing by the market size. We assume that each individual has the potential to consume one serving of cereal per day; market size is then computed as the product of the total number of households in the trading area of a supermarket chain and the average household size. The market share of the outside good is defined as the difference between one and the sum of the brands under consideration.

Prices are obtained by dividing the dollar sales by the volume sales converted into number of servings.

Product characteristics were obtained from cereal boxes and include fat, sugar, and calorie contents.

The income variable was obtained by assuming that household income in the trading area of each supermarket chain has a log normal distribution, whose parameters we estimated from the distribution of income. Individual household income is then obtained by drawing a sample of 400 observations from the log normal distribution for each supermarket chain, thus given a total of 1,200 households.

Table 2 contains the list of brands included in the analysis as well as the descriptive statistics of price and within-chain market share variables. Within-chain market shares are computed by dividing the volume sales of a given brand by the supermarket chain total volume sales in a given period. Summary statistics for other variables are given in Table 3.

TABLE 2 ABOUT HERE

TABLE 3 ABOUT HERE

As instruments for prices we use a set of variables that proxy marginal costs and exogenous variations in prices over time. Over the period covered by our data, in response to low consumption of breakfast cereal, cereal manufacturers slashed cereal prices. To account for these events, we included two indicator variables for April and June 1996. As proxies for marginal production, packages, and distribution costs, we use brand and supermarket chain indicator variables.

Permutation construction is used to induce the ordering to vary with household income. The values $n_0 = 1$ and $\eta = 0.5$ are chosen in the prior of M; the following values are chosen for the other priors: $\mu_0 = 0$, $V_0 = S_{\Sigma_H 0} = S_{\Sigma 0} = 100I$, and $\upsilon_{\Sigma_H 0} = \upsilon_{\Sigma 0} = 2$. These values are chosen such that the prior variances are very large.

The MCMC sampler was run for 20,000 iterations and the last 10,000 iterations were used to obtain parameter estimates. To assess convergence, we use different starting points for the chain and examine the trace plots of the model parameters (not shown).

We allowed for heterogeneity in price and cereal characteristics (sugar, fiber, and calorie contents) coefficients. Figure 3 plots the posterior density function of the precision parameter M, which, recall, measures the suitability of a parametric model for the individual specific; values close to zero suggests the parametric model is inadequate. From figure 3, it appears that most of the values of M are close to 0.1, indicating that the normal centering distribution is very inadequate for the data.

FIGURE 3 ABOUT HERE

Figure 4 shows the posterior distribution of the individual specific. For each parameter, standard kernel density estimation with window width computed following the recommendation of Silverman (1986), and the Nadaraya-Watson regression estimation are displayed. Overall, the distributions of consumer preferences are highly non-normal and the relationships between preferences and income are nonlinear. The density plots show that the distribution of price sensitivities, calorie, fiber, and sugar preferences are bimodal, thus contrasting the results of Chidmi and Lopez (2007) and Nevo (2001) who assumed a normal distribution for taste parameters. Here, the flexibility of the Dirichlet process that we used to model heterogeneity helps capture multimodality in the distribution of taste coefficients. The conditional density plots further show that the relationship between tastes parameters and income in nonlinear and high income households do not have the same preferences as low income households.

FIGURE 4 ABOUT HERE

Table 4 displays a sample of estimated own and cross price elasticities. Each entry i, j, where i indexes a row and j a column, represents the percentage change in the market share of brand i for a 1% change in the price of brand j. The values displayed are the median over the 3 supermarket chains and 25 quad-periods considered in the analysis. All own-price elasticities and most cross-price elasticities are larger than those found by Nevo (2001) and Chidmi and Lopez (2007).

6. Conclusion

In this paper, we have applied a Bayesian semiparametric technique to an important class of models, the random coefficients discrete choice demand models. We specified a Dirichlet process prior which varies with consumer's continuous covariates (Griffin and Steel, 2006) for the distribution of consumer heterogeneity. We developed an MCMC algorithm, and illustrate our methodology to estimate the extent of unobserved heterogeneity in demand for peanut butter and ready-to eat breakfast cereal. The empirical results indicate the limitations of the unimodal distribution and the linearity of the relationship between consumer preferences and demographics that are often assumed in modeling consumer heterogeneity.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* 88 (422), 669-679.

Allenby, G. M, Arora, N., and Rossi, P. E. (1998). On the Heterogeneity in demand. *Journal of Marketing Research* 35, 384-389.

Andrews, R. L. and Currim, I. S. (2003). A Comparison of Segment Retention Criteria for Finite Mixture Logit Models. *Journal of Marketing Research* 40(2), 235-243.

Antoniak, R. L. (1974). Mixtures of Dirichlet process with application to Bayesian Nonparametric Problems. *Annals of Statistics* 2, 1152-1174.

Basu, S. and Chib, S. (2003). Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of the American Statistical Association* 98 (461), 224-235.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica* 63, 841-890.

Chen, Y. and Yang, S. (2006). Estimating disaggregate model using aggregate data via data augmentation of individual choice. *Journal of Marketing Research* 4, 596-613.

Chintagunta, P. K. (2001). Endogeneity and Heterogeneity in a Probit Demand Model: Estimation Using Aggregate Data. *Marketing Science* 20(4), 442-456.

Chintagunta, P. K., Jain, D. C., and Vilcassim, N. J. (1991). Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data. Jou*rnal of Marketing Research* 28, 417-428.

Chidmi, B. and Lopez, R.A. (2007). Brand-Supermarket Demand for Breakfast Cereals and Retail Competition. *American Journal of Agricultural Economics* 89 (2), 324-337.

Cifarelli, D. M.and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The use of associative means. *Annali de l'Instituto di Matematica Finianziara dell' Universitá di Torino* Serie III (12), 1-36.

De Iorio, M., Muller, P., Rosner, G. L., and MacEachern, S.N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association* 99, 205-215.

Escobar, M.D. and West, M. (1998). Computing Bayesian Nonparametric Hierarchical Models. In Dey, D., Muller, P., and Sinha, D. (Ed.), *Practical Nonparametric and semiparametric Bayesian Statistics* (pp. 1-22). New York: Springer-Verlag.

Duan, J., Guindani, M., and Gelfand, A. (2007). Generalized Spatial Dirichlet Process Models. *Biometrika* 94 (4), 809-825.

Dillon, W. R. and Kumar, A. (1994). Latent Structure and Other Mixture Models in Marketing: An lntegrative Survey and Overview. In R.P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp. 295-351). Cambridge: Blackwell Publishers.

Ferguson, T.A. (1973). A Bayesian Analysis of some Nonparametric Nonparametric Problems. *Annals of Statistics* 1, 209-230.

Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2004). Bayesian nonparametric Spatial Modeling with Dirichlet process mixing. Technical report, Duke University.

Griffin, J. E. and Steel, M.F.J. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association* 101, 179-194.

Ishwaran, H. and James, L.F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* 96(453), 161-173.

Kamakura, W. A. and Russell G. J. (1989). A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. *Journal of Marketing Research* 26, 379-390.

Kim, J.G. and Menzefricke, U. (2004). Assessing Heterogeneity in Discrete choice models using a Dirichlet process prior. *Review of marketing Science* Volume 2, article 1.

Kottas, A., and Krnjajic, M. (2005). Bayesian Nonparametric Modeling in Quantile Regression. Technical Report, University of California, Santa Cruz.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceeding of the section on Bayesian Statistical Sciences* (pp. 50-55).

MacEachern, S.N. (2000). Dependent Dirichlet processes. Technical Report, Department of Statistics, Ohio State University, Columbus.

MacEachern, S.N., Kottas, A., and Gelfand, A. E. (2001). Spatial Nonparametric Bayesian Models. Technical report, ISDS, Duke University.

McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15,447-470.

Musalem, A., Bradlow, E.T., and Raju, J. S. (2006). Bayesian estimation of random coefficients choice models using aggregate data. Technical Report, University of Pennsylvania, Philadelphia.

Musalem, A., Bradlow, E.T., and Raju, J. S. (2005). Who's got the coupon: estimating consumer preferences and coupon usage from aggregate data. Technical Report, University of Pennsylvania, Columbus.

Neal, R. M. (2003). Slice Sampling. *The Annals of Statistics* 31(3), 705-767.

Nevo, A. (2001). Measuring market power in the ready to eat cereal industry. *Econometrica* 69(2), 307-342.

Rossi, P., Allenby, G., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley and Sons.

Sethuraman, J. (1994). A Constrictive Definition of Dirichlet Priors. *Statistica Sinica* 639-650.

Silverman, B. (1986). Density Estimation. London: Chapman and Hall.

Tanner, M. A. and Wong, W. H. (1987). The calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* 82, 528-550.

Yang, S., Chen, Y., and Allenby, G. M. (2003). Bayesian Analysis of Simultaneous Demand and Supply. *Quantitative Marketing and Economics* 1, 252-275.

Villas-Boas, J. M. and Winer, R. S. (1999). Endogeneity in brand choice models. *Management Science* 45(10), 1324-1338.

Wedel, M. and Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations, 2e Ed. Boston: Kluwer Academic Publishers.

Table 1: Descriptive statistics

	Brand 1	Brand 2	Brand 3	Brand 4
Market share	24.68	29.02	12.03	34.27
Proportion of observations with feature advertising	6.86	21.15	24.43	10.60
Average Price ($)	1.72	1.62	1.60	1.38

Figure 1

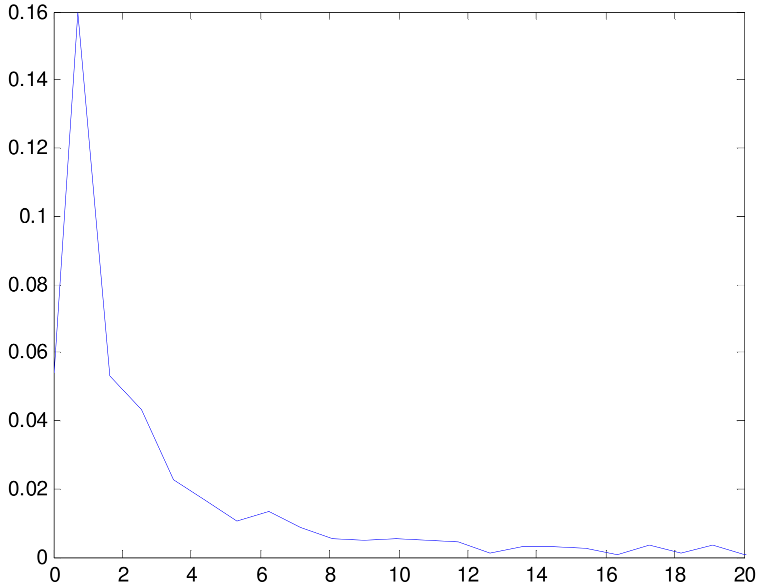
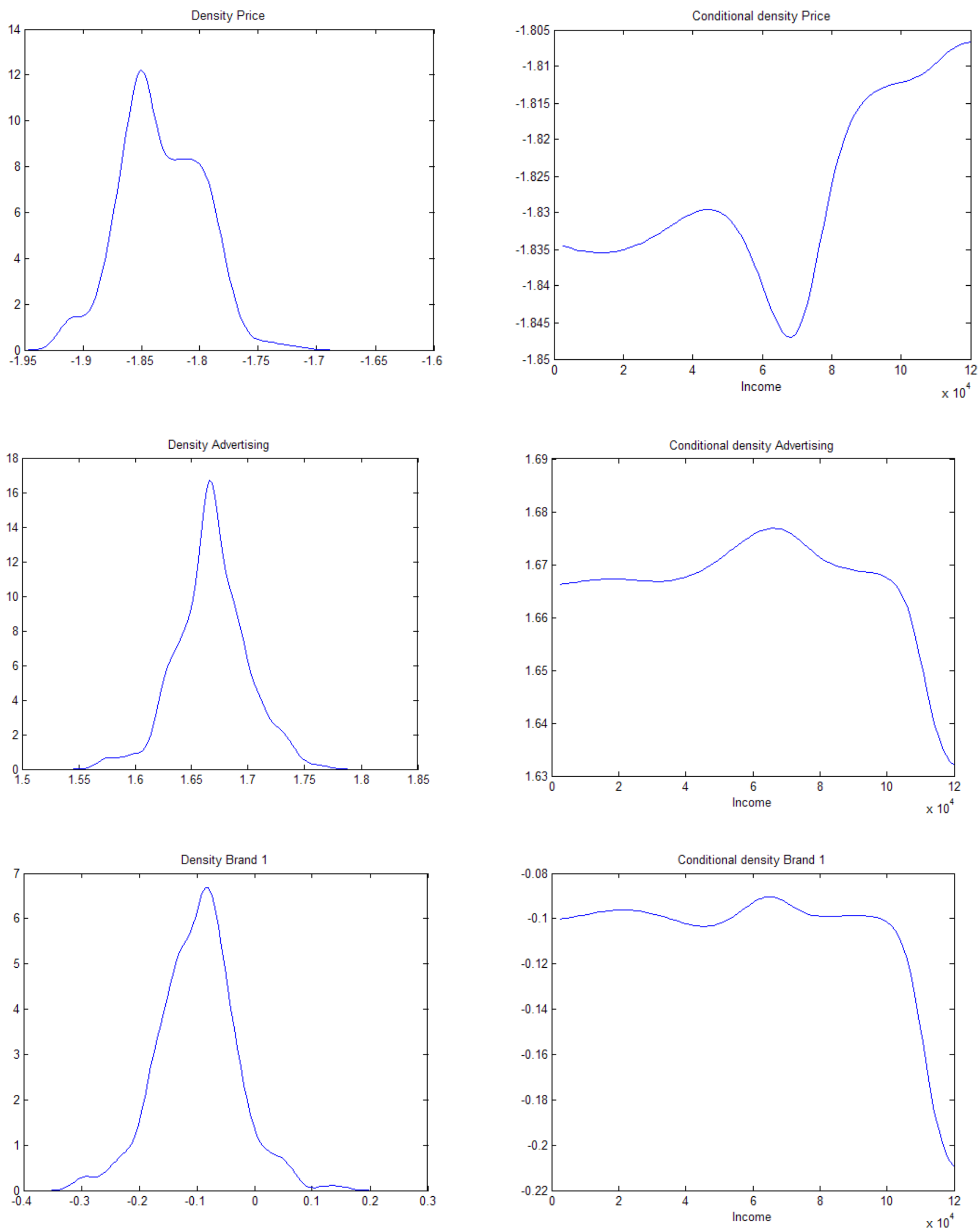
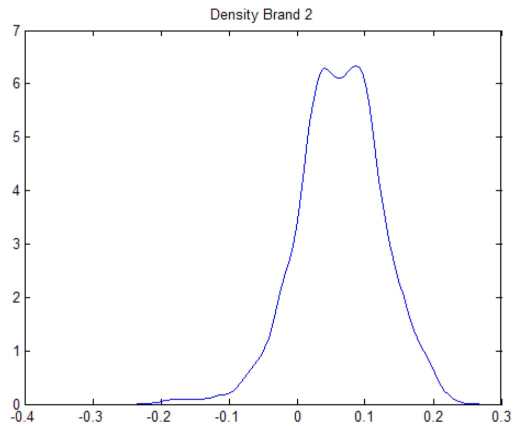
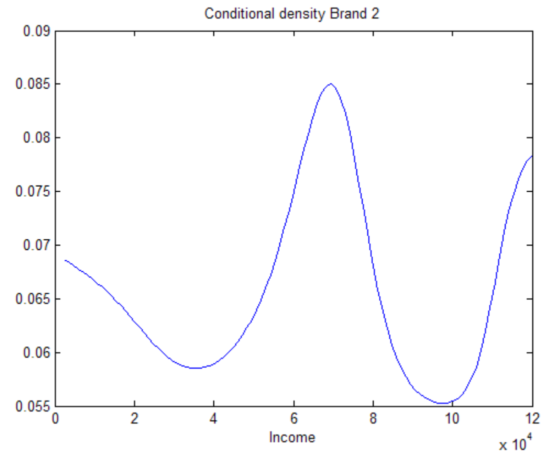
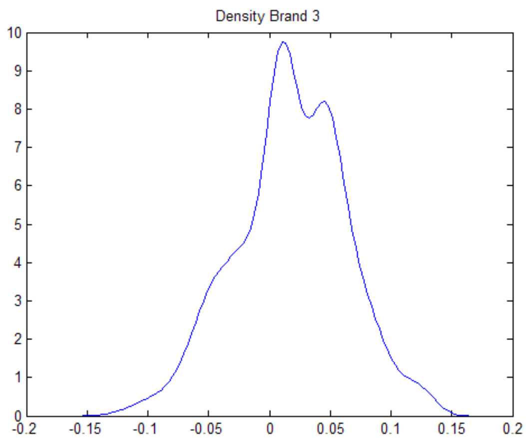


Figure 2: Density for the individual-specific parameters




Density Brand 2


Conditional density Brand 2
Income
x 10^4


Density Brand 3

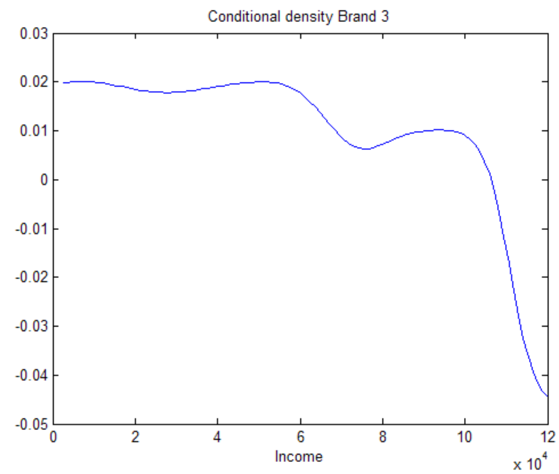

Conditional density Brand 3
Income
x 10^4

Table 2 Price and market share of brands included in the analysis

Brand	Price ($/serving)		Within chain market share (%)	
	Mean	S. D.	Mean	S. D.
K Frosted Flakes	0.482	0.061	4.03	1.99
K Corn Flakes	0.3563	0.0590	3.92	1.89
K Frosted Mini Wheat	0.8164	0.1180	3.47	1.45
K Raisin Bran	0.8076	0.1276	3.56	1.73
K Froot Loops	0.5763	0.0955	1.91	1.20
K Rice Krispies	0.6455	0.0819	2.05	0.99
K Corn Pop	0.6200	0.0959	1.80	1.09
K Special K	0.7074	0.0862	2.02	1.12
K Apple Jacks	0.6094	0.0962	1.26	0.93
K Crispix	0.6629	0.0941	1.14	0.59
K Honey Crunch Corn Flakes*	0.4869	0.0816	1.42	0.82
K Cocoa Frosted Flakes*	0.5147	0.0793	0.90	0.81
GM Cheerios	0.5700	0.0821	3.97	1.48
GM Honey Nuts Cheerios	0.5041	0.0555	3.14	1.37
GM Lucky Charms	0.6268	0.0889	1.86	1.15
GM Cinnamon Toasted Crunch	0.6241	0.0862	1.48	0.77
GM Weathies	0.5083	0.0740	1.27	0.76
GM Kix	0.7296	0.0926	1.33	0.61
GM Frosted Cheerios	0.5188	0.0727	1.36	1.09
GM Total	0.7171	0.0783	1.16	0.64
GM Golden Graham	0.6486	0.0638	0.93	0.62
GM French Toast Crunch*	0.6232	0.1591	0.73	0.72
P Grape nuts	0.7513	0.1446	1.82	0.92
P Raisin Bran	0.7761	0.1208	1.89	1.29
P Honey Bunch of Oats	0.5133	0.0759	1.65	0.98
P Fruity Peeple	0.5359	0.0756	1.03	0.62
P Honey Comb	0.5618	0.0910	0.84	0.60
Post Shredded Wheat	0.7733	0.1016	1.22	0.71
P Cranberry Almonds Crunch*	1.1752	0.1621	0.67	0.45
Q Cap N Crunch	0.4705	0.0847	1.73	1.26
Q Cap N Crunch Crunch Berrs	0.4576	0.0809	1.24	0.94

Source: Authors computation
*=New brands

Table 3. Sample statistics

	Mean	Std	Min	Max
Calories	130.8	32.8	101	220
Fiber	1.9677	1.9754	0	7.0000
Sugar	9.4516	5.0022	0	20.0000
Household Income ($)	53,761	28,117	6,997	216,260

Source: Cereal boxes and samples from the log-normal distributions

Figure 3 Posterior density of the precision parameter M

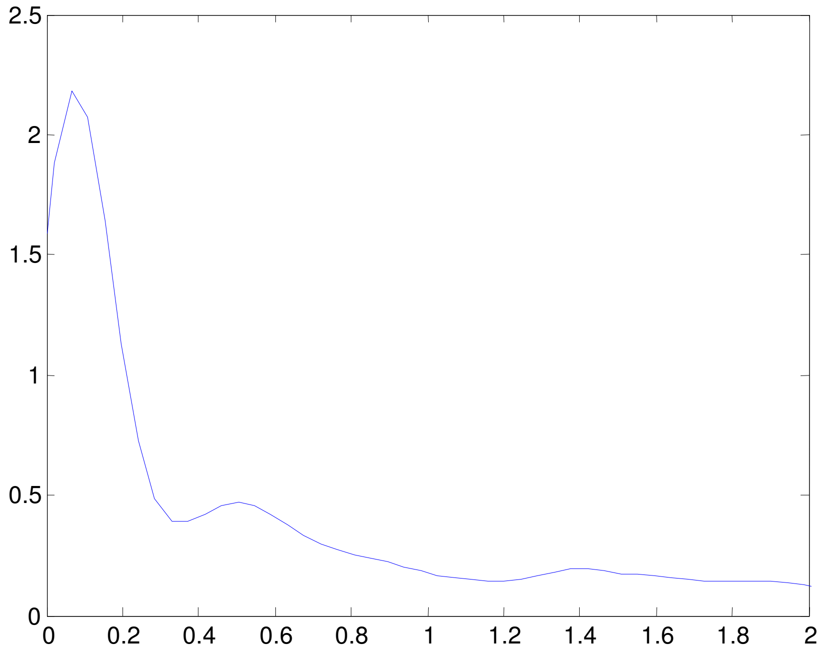
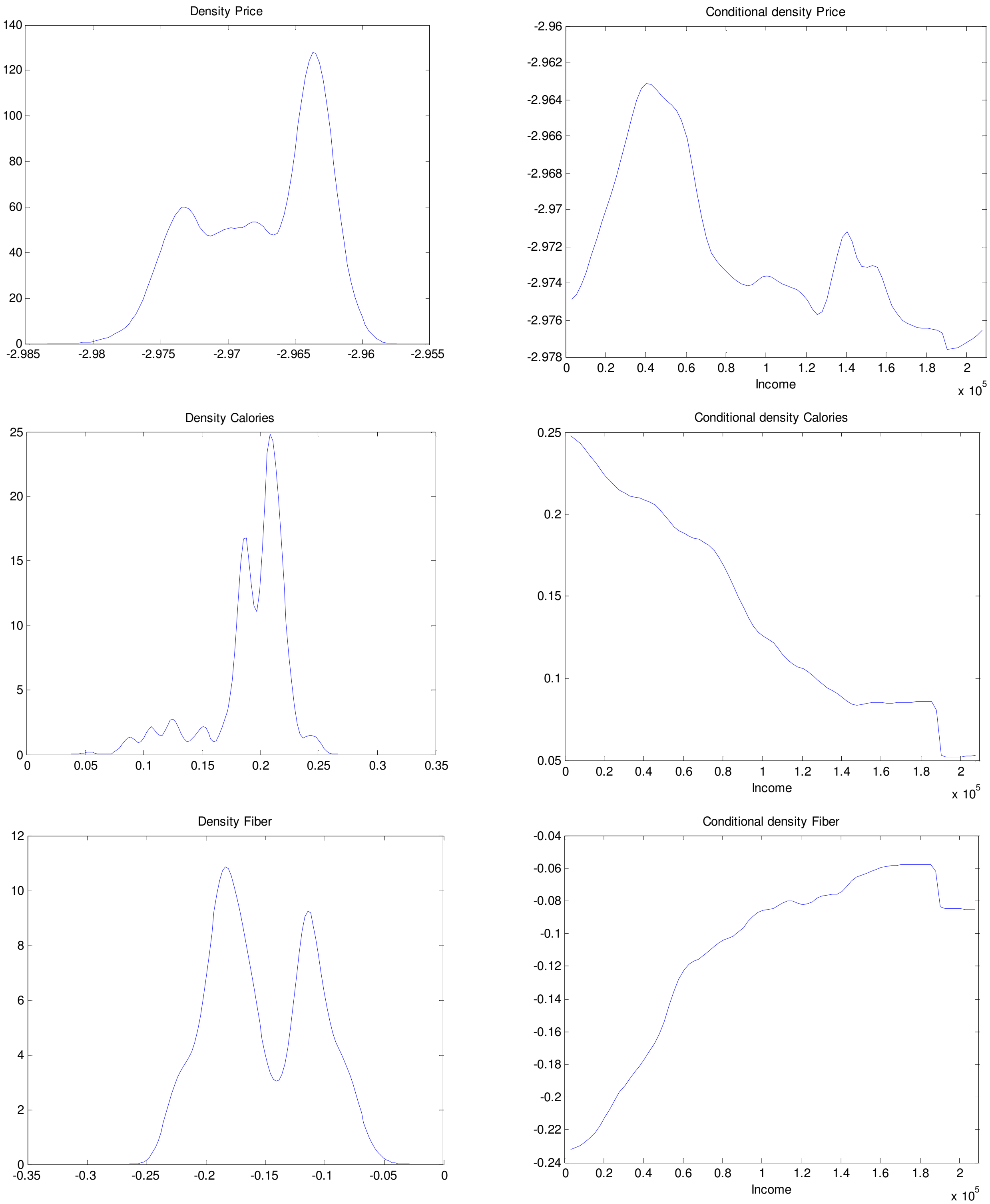
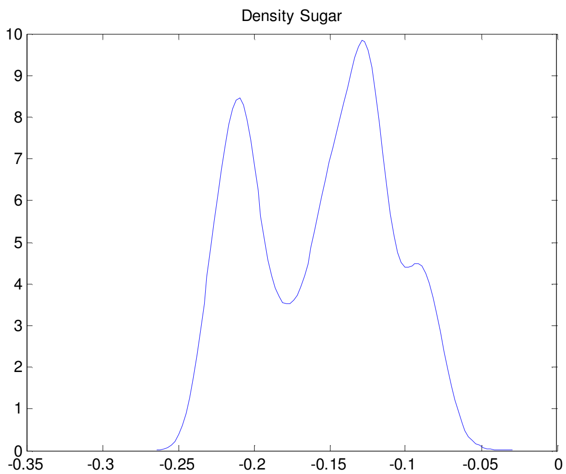


Figure 4: Density for the individual-specific parameters




Density Sugar
10
9
8
7
6
5
4
3
2
1
0
-0.35
-0.3
-0.25
-0.2
-0.15
-0.1
-0.05
0

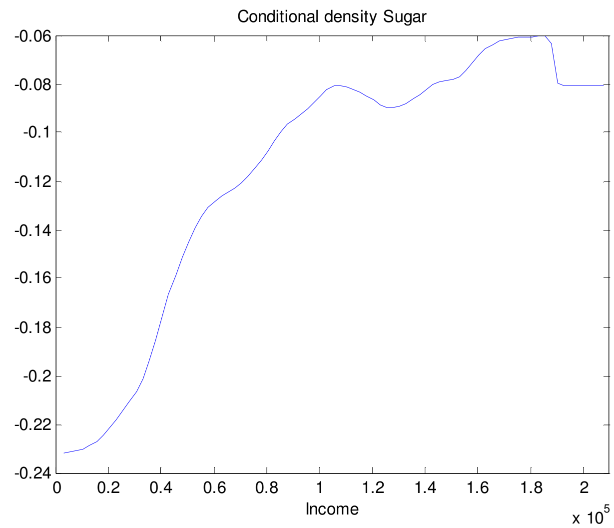

Conditional density Sugar
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
-0.2
-0.22
-0.24
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Income
x 10^5

Table 4: Median own and cross-price price elasticities

	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Brand 7	Brand 8	Brand 9	Brand 10
Brand 1	-1.782	0.063	0.070	0.046	0.095	0.030	0.081	0.037	0.062	0.013
Brand 2	0.066	-1.753	0.049	0.019	0.039	0.115	0.021	0.113	0.016	0.041
Brand 3	0.116	0.066	-1.776	0.091	0.062	0.024	0.035	0.038	0.037	0.011
Brand 4	0.122	0.041	0.149	-1.696	0.063	0.018	0.087	0.023	0.080	0.008
Brand 5	0.177	0.063	0.070	0.046	-1.839	0.030	0.081	0.037	0.062	0.013
Brand 6	0.072	0.263	0.041	0.018	0.042	-1.755	0.023	0.115	0.017	0.047
Brand 7	0.170	0.049	0.051	0.081	0.090	0.025	-1.760	0.028	0.108	0.011
Brand 8	0.073	0.245	0.051	0.019	0.042	0.109	0.022	-1.685	0.017	0.039
Brand 9	0.168	0.046	0.066	0.100	0.089	0.022	0.149	0.026	-1.704	0.010
Brand 10	0.072	0.263	0.041	0.018	0.042	0.133	0.023	0.114	0.017	-1.794
Brand 11	0.056	0.225	0.070	0.021	0.034	0.083	0.019	0.095	0.015	0.032
Brand 12	0.127	0.108	0.105	0.029	0.070	0.043	0.033	0.062	0.028	0.018
Brand 13	0.180	0.054	0.068	0.060	0.096	0.026	0.105	0.032	0.079	0.012
Brand 14	0.152	0.096	0.069	0.028	0.082	0.044	0.049	0.055	0.037	0.018
Brand 15	0.066	0.195	0.092	0.022	0.039	0.068	0.020	0.092	0.016	0.027
Brand 16	0.066	0.259	0.049	0.019	0.039	0.115	0.021	0.113	0.016	0.041
Brand 17	0.178	0.063	0.070	0.046	0.095	0.030	0.082	0.037	0.062	0.013
Brand 18	0.073	0.178	0.103	0.023	0.043	0.062	0.021	0.088	0.017	0.025
Brand 19	0.152	0.096	0.069	0.028	0.082	0.044	0.049	0.055	0.037	0.018
Brand 20	0.085	0.115	0.186	0.043	0.048	0.038	0.023	0.062	0.021	0.016
Brand 21	0.120	0.044	0.186	0.243	0.063	0.018	0.067	0.025	0.068	0.008
Brand 22	0.093	0.199	0.057	0.019	0.052	0.089	0.025	0.101	0.019	0.032
Brand 23	0.178	0.069	0.048	0.037	0.096	0.037	0.102	0.040	0.066	0.016
Brand 24	0.167	0.077	0.070	0.035	0.090	0.036	0.063	0.045	0.048	0.016
Brand 25	0.177	0.063	0.070	0.046	0.095	0.030	0.082	0.037	0.062	0.013
Brand 26	0.178	0.063	0.070	0.046	0.095	0.030	0.082	0.037	0.062	0.013
Brand 27	0.056	0.182	0.115	0.028	0.034	0.061	0.019	0.080	0.015	0.025
Brand 28	0.120	0.086	0.058	0.028	0.096	0.042	0.038	0.053	0.035	0.019
Brand 29	0.136	0.071	0.060	0.034	0.110	0.035	0.051	0.044	0.047	0.016
Brand 30	0.120	0.037	0.102	0.127	0.077	0.019	0.083	0.022	0.074	0.008
Brand 31	0.124	0.038	0.052	0.078	0.101	0.022	0.118	0.027	0.089	0.008

Each entry i , j , where i indexes a row and j a column, represent the median over supermarket chains and time of the percent change in market share of brand i with respect to one percent a change in the price of brand j The 95% credible intervals are not reported.

FOOD MARKETING POLICY CENTER
RESEARCH REPORT SERIES

This series includes final reports for contract research conducted by Policy Center Staff. The series also contains research direction and policy analysis papers. Some of these reports have been commissioned by the Center and are authored by especially qualified individuals from other institutions. (A list of previous reports in the series is available on our web site.) Other publications distributed by the Policy Center are the Working Paper Series, Journal Reprint Series for Regional Research Project NE-165: *Private Strategies, Public Policies, and Food System Performance*, and the Food Marketing Issue Paper Series. Food Marketing Policy Center staff contribute to these series. Individuals may receive a list of publications in these series and paper copies of older Research Reports are available for $20.00 each, $5.00 for students. Call or mail your request at the number or address below. Please make all checks payable to the University of Connecticut. Research Reports can be downloaded free of charge from our web site given below.

Food Marketing Policy Center
1376 Storrs Road, Unit 4021
University of Connecticut
Storrs, CT 06269-4021

Tel: (860) 486-1927
FAX: (860) 486-2461
email: *fmpc@uconn.edu*
http://www.fmpc.uconn.edu