Review of Marketing Science

Volume 2 2004 *Article* 1

Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior

Jin Gyo Kim[*] Ulrich Menzefricke[†]
Fred M. Feinberg[‡]

[*] MIT Sloan School of Management, kimjg@mit.edu

[†] Rotman School of Management, University of Toronto, menzefricke@rotman.utoronto.ca

[‡] University of Michigan Business School, feinf@umich.edu

Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior*

Jin Gyo Kim, Ulrich Menzefricke, and Fred M. Feinberg

Abstract

The finite normal mixture model has emerged as a dominant methodology for assessing heterogeneity in choice models. Although it extends the classic mixture models by allowing within component variablility, it requires that a relatively large number of models be separately estimated and fairly difficult test procedures to determine the "correct" number of mixing components. We present a very general formulation, based on Dirichlet Process Piror, which yields the number and composition of mixing components a posteriori, obviating the need for post hoc test procedures and is capable of approximating any target heterogeneity distribution. Adapting Stephens' (2000) algorithm allows the determination of 'substantively' different clusters, as well as a way to sidestep problems arising from label-switching and overlapping mixtures. These methods are illustrated both on simulated data and A.C. Nielsen scanner panel data for liquid detergents. We find that the large number of mixing components required to adequately represent the heterogeneity distribution can be reduced in practice to a far smaller number of segments of managerial relevance.

KEYWORDS: Choice Models, Heterogeneity, Dirichlet Process, Bayesian Methods, Markov chain Monte Carlo

*Jin Gyo Kim (jgkim@mit.edu), MIT Sloan School of Management, 38 Memorial Drive, E56-323, Cambridge, MA, 02142; Ulrich Menzefricke (menzefricke@rotman.utoronto.ca), Rotman School of Management, University of Toronto, 105 St. George St., Toronto, Ontario, Canada, M5S 3E6; and Fred Feinberg (feinf@umich.edu), University of Michigan Business School, 701 Tappan St., Ann Arbor, Michigan, USA, 48109. Correspondence regarding this manuscript should be addressed to the first author by e-mail to jgkim@mit.edu, or by surface mail to the above address. The authors wish to thank John D. C. Little, Greg Allenby, Michel Wedel and Thomas Otter for their suggestions. All comments welcome.

1 Introduction

One of the basic assumptions of marketing theory is that consumers may differ in their choice behavior as well as in their response to marketing-mix activities. Many studies in the choice modeling literature have demonstrated that adequate control of heterogeneity is a pre-requisite to obtain consistent estimates of the effect of independent variables on choice decisions (cf., Chamberlain 1980; Chintagunta *et al.* 1991; Gönül and Srinivasan 1993).

In the marketing literature, various approaches have been proposed to incorporate heterogeneity; Chintagunta *et al.* (1991), Wedel *et al.* (1999) and Allenby and Rossi (1999) provide reviews of previous work in the area. A number of studies have shown that household-specific regression coefficient vectors can be estimated through a random-effect specification (Rossi and Allenby 1993). These studies typically introduced a unimodal distribution for the random-effect term, usually a normal distribution. In many situations, however, the heterogeneity in the population might be multi-modal, so a normal distribution would not be an appropriate choice.

Early approaches to capture multi-modality introduced discrete point masses (e.g., Chintagunta *et al.* 1991; Kamakura and Russell 1989), the well-known latent class approach. Whereas models without heterogeneity presume that all households share a single coefficient vector, latent class models allow for separate subgroups or classes, each with its own set of coefficients. Accordingly, this approach fails to capture possible variation in regression coefficient vectors within a latent class. Households are not assigned deterministically to classes, but are considered probabilistic mixtures across them. Although 'real' heterogeneity distributions are surely not discrete, a major simplifying assumption of the latent class methodology is that one can make use of a discrete approximation. Several recent studies (Wedel and Kamakura 2001; Andrews, Ainslie and Currim 2002) have questioned whether there are any practical differences among various heterogeneity specifications, and presented empirical evidence suggesting that such differences are at best minor.

Allenby *et al.* (1998) introduced a finite normal mixture model which captures the possibility of several mixing components and variability within each (note that the latent class approach is a special case of a discrete mixture model such that parameters associated with mixing components are point masses). Although the finite normal mixture model is quite successful in recovering heterogeneity distributions of arbitrary complexity, it requires tedious and often difficult test procedures to determine the 'correct' number of mixing components (see Andrews and Currim (2001) for an in-depth discussion of this issue).

In the marketing literature, there has thus far been no systematic study concerning two major issues in mixture modeling: (1) label switching and (2) overlapping mixtures. We view these as different aspects of a single, overarching issue: determining the appropriate number and composition of 'managerially relevant' segments. The label switching problem arises because the latent indicators for mixing components are not identifiable with the likelihoods. That is, any re-labeling of mixing components yields an identical likelihood value and, when the number of such components is large, the combinatorial possibilities can be daunting. A standard response to the label switching problem is imposing an identification constraint on the latent indicators so that larger indicator values are assigned to components with larger mixing proportions. Several studies, however, have demonstrated that this standard identification restriction frequently fails to correct the label switching problem (e.g., Celeux *et al.* 2000). In addition, the overlapping mixture problem arises because there is no guarantee that parameters associated with mixing components are meaningfully separated from one another (e.g., Roeder 1994). Failing to control these two important problems, in a marketing context, entails risks of making misleading inferences regarding (1) the regression coefficients associated with mixing components, (2) the mixing proportions, and (3) the households belonging to each mixing component. Stephens (2000) provides a general framework for addressing both issues, one which can be adapted to the needs of discrete choice models as they are typically applied in marketing and economics.

In this paper, we present a nonparametric Bayesian approach to model heterogeneity with the Dirichlet process prior (cf., Antoniak 1974). This approach offers one main advantage over existing approaches – bypassing the need to determine the correct number of mixing components *post hoc* – while retaining the ability to recover a variety of heterogeneity distributions, in a unified modeling framework. In addition, we discuss how to overcome both the label-switching and overlapping mixture problems, thus estimate the cluster membership for each household, valuable input for market segmentation and targeting.

The paper is organized as follows. First, we discuss the Dirichlet Process prior, its role in heterogeneity modeling and some general issues concerning finite mixture models. Next, we consider the logit choice model in particular, devise an MCMC sampling scheme for assessing its parameters and discuss how label switching problems tend to arise. We then chart the model's performance on four simulated data sets, varying the degree of skew in the parameter heterogeneity distribution as well as the number of households; we also show how one might uncover the 'true' number of clusters, as opposed to a much larger number indicated by the posterior mode. Finally, we apply the model to A. C. Nielsen liquid detergent scanner data, and discuss its performance and implications.

2 A Heterogeneous Choice Model

In this section we describe a heterogeneous choice model where the heterogeneity across households is of rather general form. To this end, we use a Dirichlet Process prior (Antoniak 1974; Escobar and West 1998; Ferguson 1973, 1983; Neal 1998) for the regression parameters of the marketing-mix variables. Throughout, we use three generic subscripts: h denotes a household ($h = 1, ..., H$), j denotes a brand ($j = 1, ..., J$), and t_h denotes purchase occasion ($t_h = 1, ..., T_h$). Let

- $y_{ht_h} = j$ denote the event that household h chooses brand j on purchase occasion t_h,
- $y_h = (y_{h1}, ...y_{hT_h})$ denote the choices for household h at purchase occasions $1, ..., T_h$, and $y = (y_1, ..., y_H)$ denote all the choice data,
- x_{hjt_h} denote the k-dimensional vector of predictor variables for brand j and household h on purchase occasion t_h,
- $x_{ht_h} = (x_{h1t_h}, \ldots, x_{hJt_h})'$ denote the $(J \times k)$ matrix of predictor variable values for household h on purchase occasion t_h,
- β_h denote the k-dimensional vector of regression parameters for household h, and
- $p(y_{ht_h} = j|\beta_h)$ denote the probability that household h chooses brand j on purchase occasion t_h, where the notation ignores the dependence on the *known* predictor variables.

There is a variety of models for linking the choice probability to the regression parameters. One such model is the multinomial logit:

$$p(y_{ht_h} = j|\beta_h) = \frac{\exp(x'_{hjt_h}\beta_h)}{\sum_{i=1}^{J}\exp(x'_{hit_h}\beta_h)}. \tag{1}$$

This model arises from a particular latent utility structure: Letting $u_{ht_h} = (u_{h1t_h}, ..., u_{hJt_h})'$ denote the J-dimensional vector of utilities for the J brands with

$$u_{ht_h} = x_{ht_h}\beta_h + \varepsilon_{ht_h}, \tag{2}$$

the logit model assumes that the errors ε_{ht_h} have a Type I extreme value distribution. Assuming that households maximize their expected utilities, the choice probabilities in (1) result.

Other choice models assume different distributions for the error vector ε_{ht_h}. For instance, the probit model arises when the distribution of ε_{ht_h} is multivariate normal, that is, $\varepsilon_{ht_h} \sim N(0, \Sigma)$. Letting $\arg\max\{u_{h1t_h}, ..., u_{hJt_h}\}$ denote the brand with the *highest* utility, the choice probabilities for the probit model are

$$p(y_{ht_h} = j|\beta_h, \Sigma) = \Pr\left\{\arg\max\{u_{h1t_h}, ..., u_{hJt_h}\} = j|\beta_h, \Sigma\right\}. \tag{3}$$

There is no closed-form solution for the choice probabilities in the probit model, unlike the situation in the logit model, (1), and we must use multi-dimensional integration to compute the choice probabilities in the probit model. With more than a few choice alternatives, this problem is computationally difficult, and a great deal of effort has gone into fashioning efficient approximation methods (Hajivassiliou, McFadden and Ruud 1996; McCulloch, Polson and Rossi 2000).

In general, we denote that the choice probabilities depend on the vector of regression parameters β_h and an additional parameter θ, where $\theta = \emptyset$ in the case of the logit model, and $\theta = \Sigma$ in the case of the probit model. Thus, the likelihood for household h's choices at purchase occasions $1, ..., T_h$ is

$$p(y_h|\beta_h, \theta) = \prod_{t_h=1}^{T_h} \prod_{i=1}^{J} p(y_{ht_h} = i|\beta_h, \theta)^{q_{hit_h}}, \tag{4}$$

where $q_{hit_h} = 1$, if, in the data, $y_{ht_h} = j$, and $q_{hit_h} = 0$, otherwise, and where y_h is the T_h-dimensional vector of brand choices for household h.

2.1 Heterogeneity in the Regression Parameters

Our general model in (2) assumes that the regression parameter β_h varies across households. We now specify how we model the distribution of β_h throughout the household population. We make a very general choice, the Dirichlet Process model, introduced by Ferguson (1973) and Antoniak (1974), a Bayesian nonparametric model. Our hierarchical structure for β_h thus assumes that each β_h is independently drawn from a distribution G, where we do not assume G to have a parametric form, but to have a Dirichlet Process prior, $DP(\alpha, G_0)$, with positive concentration parameter α and baseline distribution G_0,

$$\begin{aligned} \beta_h &\sim G, \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{5}$$

Escobar and West (1998) described such a Bayesian nonparametric hierarchical model in detail, and Neal (1998) discussed various algorithms for its estimation, one of which we adapt for the purposes of modeling brand choice. Let $\mathcal{X}$ denote a space of β_h and $\mathcal{A}$ be a σ-field of subsets of $\mathcal{X}$. Then, a probability measure G, the heterogeneity distribution of β_h, is a random variable from a Dirichlet Process on $(\mathcal{X},\mathcal{A})$ with parameters α and G_0. We assume the parameters for the Dirichlet Process, α and G_0, are not fixed but random variables. In this case, G becomes a mixture of the baseline distribution with Dirichlet Process mixing, as Antoniak (1974) demonstrated.

In the Dirichlet Process model, the choice of the distribution for G_0 is not critical. As Ferguson (1973) showed, whatever the true distribution function, it's Bayes estimate based on the Dirichlet Process prior model converges to it. Note that the number of mixing components is an unknown parameter determined by the data, and a sufficiently large number of mixing components ensures that the Dirichlet Process prior model approximates the target distribution well.

It is important to realize that the Dirichlet process prior for G is a probability distribution on the space of all possible heterogeneity distributions; the baseline prior distribution G_0 can be viewed as the "location" parameter of the Dirichlet Process prior. The parameter α acts as a (positive scalar) precision parameter: when α is very large, the Dirichlet Process prior G for β_h is very close to the baseline distribution G_0; and when α is small, G is not necessarily close to G_0. We note that asymptotic properties, such as tail thickness, are related to the specification for G_0, which we will later take to be normal. In typical applications, because the prior is overwhelmed by the likelihood, choice of α has little effect on substantive results.

Using the Dirichlet Process prior for β_h and the likelihood for the choice data, and conditioning on α, G_0, and θ, one can obtain the posterior distribution for β_h. After integrating out G, the resulting conditional posterior distribution for β_h is:

$$p(\beta_h|\beta_{-h}, \alpha, G_0, \theta, y) \sim q_{0h} G_b(\beta_h|\theta, y) + \sum_{i=1, i \neq h}^{H} q_{ih} \delta_{\beta_i}(\beta_h),$$

where $\beta_{-h} = (\beta_1, ..., \beta_{h-1}, \beta_{h+1}, ..., \beta_H)$, that is, the set of H values of β_i, $(i = 1, ..., H)$, but excluding β_h, and where $\delta_{\beta_i}(\beta_h) = 1$ if $\beta_h = \beta_i$, and $\delta_{\beta_i}(\beta_h) = 0$ otherwise (Escobar and West 1998). Furthermore,

- $G_b(\beta_h|\theta, y) \propto p(y_h|\beta_h, \theta) G_0(\beta_h)$, the baseline posterior distribution for β_h,
- $q_{0h} \propto \alpha \int p(y_h|\beta_h, \theta) G_0(\beta_h) d\beta_h$ is α times the marginal distribution of y_h under the baseline prior,
- $q_{ih} \propto p(y_h|\beta_i, \theta)$ is the likelihood for y_h conditional on $\beta_h = \beta_i$, and
- $1 = q_{0h} + \sum_{i=1}^{H} q_{ih}$.

As Antoniak (1974) showed, the distinct β_h's typically reduce to fewer than H due to the clustering of the β_h inherent in the Dirichlet Process. Using the superscript * to denote distinct values, the conditional posterior distribution for β_h is

$$p(\beta_h|\beta_{-h}, \alpha, G_0, \theta, y) \sim q_{0h} G_b(\beta_h|\theta, y) + \sum_{i=1}^{L_h} n_{-h,i} q^*_{ih} \delta_{\beta^*_i}(\beta_h), \tag{6}$$

where L_h denotes the number of distinct values of the regression parameter β_i for the $H-1$ households other than household h, $n_{-h,i}$ denotes the number of households other than h for whom the regression parameter equals β^*_i, and $q^*_{ih} \propto p(y_h|\beta^*_i, \theta)$, the likelihood for y_h conditional on $\beta_h = \beta^*_i$.

If we denote the number of distinct values of the regression parameters among the H households by L, it is clear that L_h equals either $L-1$ or L, depending on whether β_h is in a cluster of its own or not. Conditional on α, the expected number of mixing components (Escobar 1994) under the Dirichlet Process prior for β_h is

$$E(L|\alpha) = \sum_{h=1}^{H} \frac{\alpha}{\alpha + h - 1}. \tag{7}$$

2.2 Completion of the Model Specification

We assume that the baseline distribution for the Dirichlet Process prior for β_h is a k-variate normal distribution with unknown mean vector μ_0 and unknown covariance matrix Σ_0,

$$[G_0|\mu_0, \Sigma_0] = N(\mu_0, \Sigma_0), \tag{8}$$

and that the prior distribution for α is a gamma distribution with a shape parameter a_α and a scale parameter b_α,

$$\alpha \sim Ga(a_\alpha, b_\alpha), \tag{9}$$

that is, $p(\alpha) \propto \alpha^{a_\alpha - 1} e^{-b_\alpha \alpha}$. Escobar (1994) presented a useful discussion on the choice of prior for α.

Furthermore, the prior distributions for μ_0 and Σ_0 are assumed to be

$$\begin{aligned} p(\mu_0) &\sim N(m_0, V_0), \text{ and} \\ p(\Sigma_0) &\sim IW_k(v_{\Sigma_0}, S_{\Sigma_0}), \end{aligned} \tag{10}$$

where $IW_k(v, S)$ denotes a k-dimensional inverted Wishart distribution with parameters v and S, where $v > 0$ and S is non-singular, that is, $\Sigma_0 \sim IW(v, S)$ implies that $p(\Sigma_0) \propto |\Sigma_0|^{-(\frac{1}{2}v+k)} \exp(-\frac{1}{2}\text{tr } \Sigma_0^{-1} S)$. Note that $E(\Sigma) = S/(v-2)$.

In the above specification, the values of $a_\alpha, b_\alpha, m_0, V_0, v_{\Sigma_0}$, and S_{Σ_0} are known. The only parameter for which we have not yet specified a prior distribution is θ, the parameter characterizing the choice probability. Recall that $\theta = \emptyset$ in the case of the logit model (1), and $\theta = \Sigma$ in the case of the probit model (3). Thus θ does not arise in the logit model. The prior distribution for $\theta = \Sigma$ in the probit model is discussed in Section 7.1.

On account of the complexity of the model, it is not possible to give a closed form solution for the posterior distribution on the parameters, that is, on $\left(\{\beta_h\}_{h=1}^H, \alpha, \mu_0, \Sigma_0, \theta\right)$. It is straightforward, however, to devise a MCMC sampling scheme to sample from the posterior distribution. A useful discussion of the convergence of the MCMC sampler in the Dirichlet Process prior models is given, for example, in Escobar (1994) and Escobar and West (1995).

After obtaining MCMC samples for $\{\beta_h\}_{h=1}^H$, the heterogeneity distribution, G, is easily estimated by using standard kernel density estimation techniques. In addition, the MCMC samples can be used to shed light on other aspects of the proposed model; specifically:

1. We can estimate the number of distinct clusters, L, needed to estimate the unknown heterogeneity distribution.

2. Conditional on L, we can estimate the size, composition, and value of β_i^* for each cluster i.

The second aspect is practically important for market segmentation. From a pragmatic perspective, examination of parameters associated with L clusters is useful if there exist meaningful differences among the L clusters in terms of β_i^*. If there is overlap among the L clusters, then we may need a way to group redundant clusters, an issue discussed in detail in Section 4 and 5.

2.3 Further Discussion of Mixture Models

Estimation of mixture models typically involves determining (1) the number of mixing components and (2) the distribution in each mixing component. Many authors, including West (1992), Diebolt and Robert (1994), and Richardson and Green (1997), have shown that mixtures of normals provide a simple and effective basis for Bayesian density estimation. Therefore, it may be convenient in many mixture contexts to choose normal mixtures.

To illustrate this approximation, consider first a unimodal context. In Fig. 1, both the thick-tailed distribution on the left and the positively-skewed distribution on the right can be approximated by a mixture of several normal distributions. The goodness of the approximation depends on the number of normal mixing components that is used. In Fig. 1, the unimodal positively skewed distribution is approximated by a mixture of six normal distributions. This approximation may suggest

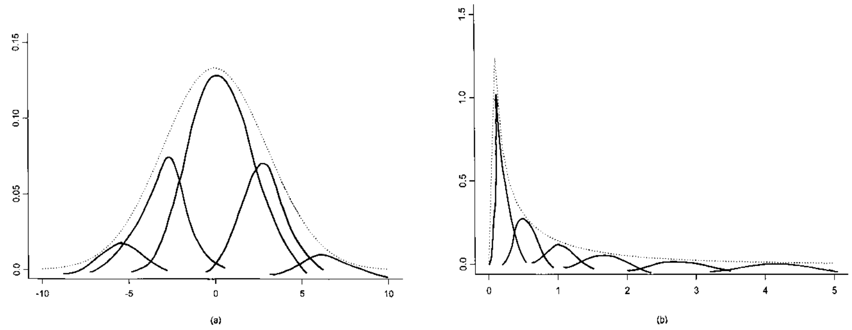


Figure 1: Approximation with the Dirichlet Process prior

that the distribution has six modes, but they are, of course, just a by-product of approximating a skewed distribution using a mixture of symmetric ones. This observation should make clear that a mixture of L normal distributions does not necessarily mean that there are L underlying mixing components each having a substantive interpretation in an applied context.

Consider now a multi-modal context with M modes where each of the modes is indicative of a cluster of observations that does have a 'substantive' interpretation. This distribution can also be approximated by a mixture of normal distributions. But since each of the underlying M distributions may be non-normal, L, the number of normal distributions required in the approximating mixture, could well be quite a bit larger than M, the actual number of 'substantive' clusters. It would thus be wrong to infer that there are L true clusters, and the approximation itself does not suggest their actual number. We denote this the "overlapping mixtures problem". Though estimation of the number of mixing components L is a by-product of the Markov chain Monte Carlo estimation method we use for the Dirichlet Process prior model, this model also runs into the overlapping mixtures problem so that L can be larger than the genuine number of mixing components. We return to this discussion in the application section.

The problem alluded to in the last paragraph is common in estimation settings featuring normal mixtures. This problem can, of course, be circumvented if the distributions of the M substantive clusters are known, since the estimation technique can then formally incorporate the form of these distributions. We suspect, though, that it is the rare application when the mixing distributions are known (while their number is not), and using normal distributions for the mixing components is a pragmatic choice. It would in any case require judicious insight into the application to determine the appropriate number of substantive clusters; see Escobar and West (1995) for one approach to determining the "genuine" number of modes.

In addition to determining the number of mixtures and the mixing distributions, another important issue when estimating mixtures is label switching, which arises because the likelihood for a mixture model is invariant to relabeling the mixing components. That is, the value of $\prod_{h=1}^{H}\{\pi_1 N(\beta_h|\beta_1^*, \Sigma_1) + \ldots + \pi_L N(\beta_h|\beta_L^*, \Sigma_L)\}$ remains the same for all permutations of the labels for the L mixing components (Redner and Walker 1984). In mixture models, this label switching problem is critical when inferences are needed for parameters associated with mixing components, component sizes, component memberships, or clustering of data since these inferences depend on the labels for the mixing components. Inferences on other parameters, for example, the estimate of the household-specific regression coefficient, β_h, are not affected by relabeling.

3 Estimation

In this section, we address estimation for the logit model, which we will use in our applications. Treatment of the probit is largely analogous – we have found little substantive difference between their inferences in the forthcoming application – and appears in the Appendix. We also discuss how to approach the label switching problem.

3.1 Estimation for the Logit Choice Model

Let $\beta = (\beta_1, ..., \beta_H)$. Then the joint posterior distribution for the parameters $(\beta, \alpha, \mu_0, \Sigma_0)$ is

$$\begin{aligned} p(\beta, \alpha, \mu_0, \Sigma_0 | y) \quad &\propto \quad \left(\prod_{h=1}^{H} p(y_h | \beta_h, \theta = \emptyset) \right) \times \left(\prod_{h=1}^{H} p(\beta_h | G) \right) \\ &\times p(G | G_0, \alpha) \times p(\alpha) \times p(\mu_0) \times p(\Sigma_0), \end{aligned} \tag{11}$$

where $p(y_h | \beta_h, \theta = \emptyset)$ is given in (4).

We implement our MCMC approach by using the following conditional distributions of the full joint posterior distribution for $p(\beta, \alpha, \mu_0, \Sigma_0 | y)$

1. $p(\beta_h | \beta_{-h}, \alpha, \mu_0, \Sigma_0; y)$ for each $h = 1, ..., H$.
2. $p(\beta_i^* | S, L, \mu_0, \Sigma_0; y)$ for each $i = 1, ..., L$. Here S denotes the cluster structure, that is, $S = (S_1, ..., S_H)$, with $S_h = i$ if $\beta_h = \beta_i^*$ $(h = 1, ..., H)$.
3. $p(\mu_0, \Sigma_0 | \beta, \alpha; y)$, and
4. $p(\alpha | \beta, \mu_0, \Sigma_0; y)$.

Let us discuss these distributions in turn, in line with the detailed discussion in Escobar and West (1998).

3.1.1 Sampling From the Conditional Distribution of β_h

To sample from the conditional distribution of β_h, we use algorithm 5 in Neal (1998). Let $c_{h,c}$ be the cluster to which household h belongs at the beginning of the current iteration, and let $\beta_{h,c}$ be the current value of β_h.

1. Draw a proposed cluster, c_p, from the integers $\{0, 1, ..., L\}$, with probabilities $\{\frac{\alpha}{H-1+\alpha}, \frac{n_{-h,1}}{H-1+\alpha}, \ldots, \frac{n_{-h,L}}{H-1+\alpha}\}$.
2. If $c_p \in \{1, ..., L\}$, let the proposed value for β_h be $\beta_{h,p} = \beta^*_{c_p}$. If $c_p = 0$, draw the proposed value, $\beta_{h,p}$, for β_h from $N(\mu_0, \Sigma_0)$.
3. Accept the proposed value with probability

$$\pi(c_{h,c}, c_p) = \min \left(1, \frac{p(y_h | \beta^*_{c_p})}{p(y_h | \beta^*_{c_{h,c}})} \right).$$

3.1.2 Sampling From the Conditional Distribution of β_i^*

Let $\mathcal{H}_i$ denote the set of households for which $\beta_h = \beta_i^*$ $(i = 1, ..., L)$. Then the posterior distribution for β_i^* is

$$p(\beta_i^*|S, L, \mu_0, \Sigma_0; y) \propto \left(\prod_{h \in \mathcal{H}_i} p(y_h|\beta_i^*, \theta = \emptyset) \right) n(\beta_i^*|\mu_0, \Sigma_0).$$

We can use the Metropolis-Hastings algorithm to sample from this posterior distribution, but it may converge slowly. Since (1) is log-concave with respect to β_h, there are several efficient algorithms for sampling from this posterior. One of them is the adaptive rejection method (Gilks and Wild 1992), another is the slice sampler (Neal 1997) which we use here. We update the k elements of $\beta_i^* = (\beta_{i1}^*, ..., \beta_{ik}^*)^{'}$ in sequence by completing the following four steps for each element β_{ij}^* $(j = 1, ..., k)$:

1. Compute $r = w(\beta_{ij}^*) = \left(\prod_{h \in \mathcal{H}_i} p(y_h|\beta_i^*, \theta = \emptyset) \right) n(\beta_i^*|\mu_0, \Sigma_0)$.
2. Define a vertical slice v by randomly sampling from $(0, r)$.
3. Find a horizontal slice $I = (D, E)$ so that $w(D) < r$ and $w(E) < r$.
4. Replace the current value of β_{ij}^* by a new value $\tilde{\beta}_{ij}^*$ that is sampled uniformly from I, if $w(\tilde{\beta}_{ij}^*) > r$.

3.1.3 Sampling From the Conditional Distribution of μ_0 and Σ_0

The conditional distribution for μ_0 and Σ_0 reduces to

$$p(\mu_0, \Sigma_0|\beta, \alpha; y) = p(\mu_0, \Sigma_0|\beta) = p(\mu_0|\Sigma_0, \beta)p(\Sigma_0|\beta),$$

where $\beta = (\beta_1, ..., \beta_H)$.

Sample μ_0 from $N(m^*, V^*)$, where $m^* = V^*(V_0^{-1}m_0 + \sum_{l=1}^{L} \Sigma_0^{-1}\beta_l^*)$, $V^* = (V_0^{-1} + L\Sigma_0^{-1})^{-1}$. Sample Σ_0 from $IW(v_\Sigma^*, S_\Sigma^*)$, where $v_\Sigma^* = v_\Sigma + L$ and $S_\Sigma^* = S_\Sigma + \sum_{l=1}^{L}(\beta_l^* - \mu_0)(\beta_l^* - \mu_0)^{'}$.

3.1.4 Sampling From the Conditional Distribution of α

Escobar and West (1998) showed that the conditional posterior distribution for $p(\alpha|\beta, \mu_0, \Sigma_0; y)$ reduces to $p(\alpha|L, y)$, and they proposed the following two-step sampling approach:

1. Sample an auxiliary variable ζ from $B(\alpha + 1, H)$, a beta distribution with mean $(\alpha + 1)/(\alpha + H + 1)$.
2. Sample α from a mixture of two gamma densities,

$$\phi Ga(a_\alpha + L, b_\alpha - \log(\zeta)) + (1 - \phi)Ga(a_\alpha + L - 1, b_\alpha - \log(\zeta)),$$

where

$$\phi = \frac{a_\alpha + L - 1}{H(b_\alpha - \log(\zeta)) + (a_\alpha + L - 1)}.$$

3.2 Label Switching

A standard response to the label switching problem is imposing an identifiability constraint on some of the parameters, for example, $(\pi_1, \ldots, \pi_L)$ or $(\beta_1^*, \ldots, \beta_L^*)$. The permutation for the labels in c could be determined so that $\pi_1 > \ldots > \pi_L$ (e.g., Allenby *et al.* 1998). Since there often exist several choices for identifiability constraints, it is unclear how to determine an appropriate one to remove the label switching problem. Indeed, Celeux *et al.* (1998), Stephens (2000), and Frühwirth-Schnatter (2001) all demonstrated that a standard identifiability constraint may fail to rectify label switching problems.

Despite its importance, few studies have addressed the label switching problem. Recently, Stephens (2000) proposed two relatively simple algorithms to find the best labeling scheme based on post-simulation examination. One of the algorithms is designed to find the best labeling scheme in order to cluster observations into L groups. However, his approach is not directly applicable to the Dirichlet process prior normal mixture since the number of mixing components varies over iterations. However, the relabeling algorithm can be used for clustering inference to post-process the MCMC output *conditional* on the estimated number of mixing components, once that number has been determined by examining the posterior distribution for L.

In order to apply Stephens' relabeling algorithm, we invoke an assumption not dictated by the theory developed thus far, that the distribution of the regression coefficients in (5) is a mixture of L normal distributions; that is,

$$f(\beta_h|\xi) = \pi_1 n(\beta_h|\beta_1^*, \Sigma_1) + \pi_2 n(\beta_h|\beta_2^*, \Sigma_2) + \ldots + \pi_L n(\beta_h|\beta_L^*, \Sigma_L), \tag{12}$$

for $h = 1, \ldots, H$, where $\xi = (\pi_1, \ldots, \pi_L, \beta_1^*, \ldots, \beta_L^*, \Sigma_1, \ldots, \Sigma_L)$, the π_l denote the mixing probabilities and $n(\beta_h|\beta^*, \Sigma)$ denotes a multivariate normal distribution with mean β^* and covariance matrix Σ.

To apply Stephens' relabeling algorithm, define the $H \times L$ matrix $R = \{r_{hl}\}$, where r_{hl} denotes the estimated probability to assign household h to mixing component l. Furthermore, define the $H \times L$ matrix of classification probabilities $B = \{b_{hl}(\xi)\}$, where $b_{hl}(\xi)$ denotes the probability based on (12) that household h belongs to mixing component l, that is,

$$b_{hl}(\xi) = \frac{\pi_l n(\beta_h|\beta_l^*, \Sigma_l)}{\sum_{i=1}^{L} \pi_i n(\beta_h|\beta_i^*, \Sigma_i)}.$$

Note that each row of both R and B sums to 1.

Let ν be a permutation, that is, a re-labeling, of $1, 2, \ldots, L$, and define the corresponding permutation of the parameter vector ξ by

$$\nu(\xi) = (\pi_{\nu(1)}, \ldots, \pi_{\nu(L)}, \beta_{\nu(1)}^*, \ldots, \beta_{\nu(L)}^*, \Sigma_{\nu(1)}, \ldots, \Sigma_{\nu(L)}).$$

Furthermore, let $\xi^{(i)}$ denote the simulated value of ξ at the i-th MCMC iteration, $i = 1, \ldots, N$, and let $\nu^{(i)}$ denote the permutation of $1, 2, \ldots, L$ at the i-th MCMC iteration.

Using a Kullback-Leibler loss function, Stephens' algorithm proceeds as follows to post-process the MCMC output in order to find a good permutation, that is, labeling, scheme. Let $\nu_1, \ldots, \nu_N$ be any set of permutations of $1, \ldots, L$ (for example, the identity permutations), and iterate the following two steps:

- Step 1: Let $\hat{r}_{hl} = \frac{1}{N}\sum_{i=1}^{N} b_{hl}\left(\nu^{(i)}(\xi^{(i)})\right)$.
- Step 2: For $i = 1, \ldots, N$, choose $\nu^{(i)}$ to minimize $\sum_{h=1}^{H}\sum_{l=1}^{L} b_{hl}\{\nu^{(i)}(\xi^{(i)})\} \log\left(\frac{b_{hl}\{\nu^{(i)}(\xi^{(i)})\}}{\hat{r}_{hl}}\right)$.

4 Simulation with Synthetically Created Data

4.1 Simulation Purpose

We conducted a simulation study to investigate the following issues related to the proposed model:

1. If the target density is a normal distribution, does the Dirichlet Process normal mixture model recover it well? If the target density is not normal, does the proposed model approximate it well?

2. Does an increase in number of households lead to an increase in L? Because the expected number of mixing components under the prior depends on the sample size as shown in (7), L may increase as the number of households becomes larger.

3. Does the proposed model recover the true number of clusters well for both the normal and the non-normal cases?

4.2 Synthetically Created Data

We generated synthetic data for our simulations based on the logit model

$$p(y_{ht_h} = j|\beta_h) = \frac{\exp(x'_{hjt_h}\beta_h)}{\sum_{i=1}^{J} \exp(x'_{hit_h}\beta_h)}, \tag{13}$$

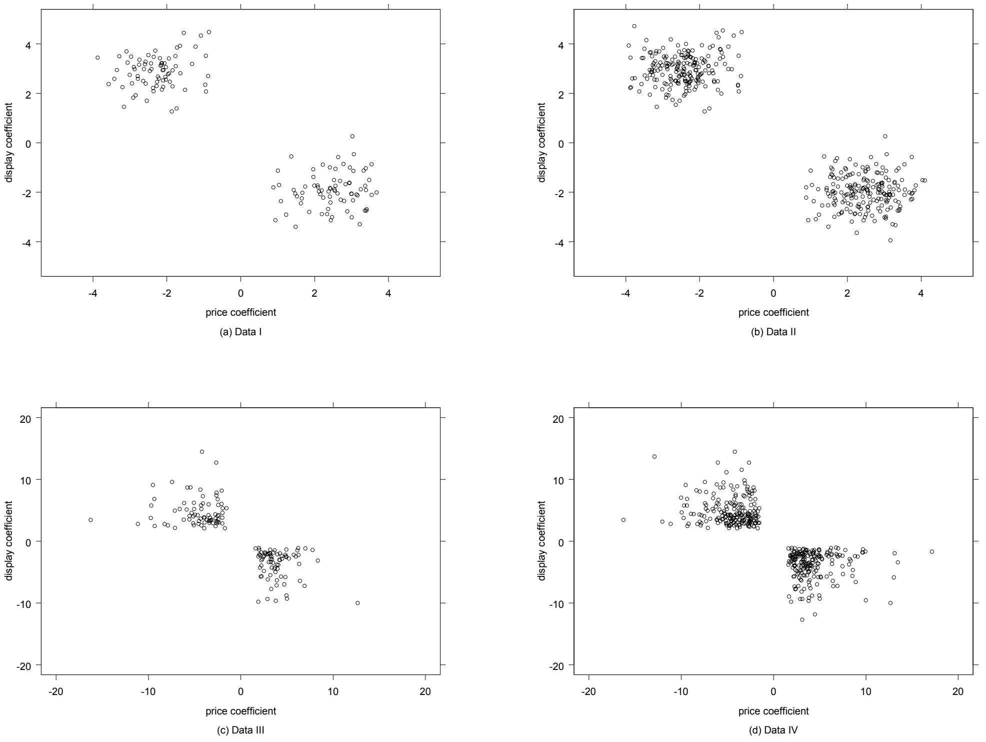
where the household-specific vector of regression coefficients comes from the distribution $p(\beta_h)$, to be discussed below. To gauge the effect of sample size, we created data sets with two values for the number of households, $H = 150$ or 400. The number of purchase occasions was fixed to 21 for all households in all data sets, in line with the forthcoming empirical application. The number of brands was $J = 3$ and there were $k = 3$ predictor variables, so that the dimension of β_h was 3. The three predictor variables were

$$x_{hjt_h} = \begin{bmatrix} \text{dummy variable for feature advertising} \\ \text{dummy variable for display} \\ \text{price} \end{bmatrix},$$

where feature advertising and display were generated from Bernoulli distributions with parameters 0.6 and 0.4, respectively, and price was generated from a normal distribution with mean 2 and variance 2, truncated to be between 1 and 3.

The distribution of feature ad, display, and price was the same for all brands, and there is no brand-specific information in x_{hjt}, that is, there are no brand dummy variables. Thus all brands can be taken to have equal baseline preference. In addition to the predictor variable values, x_{hjt_h}, we also had to generate values for the regression coefficients, β_h. In all data sets, their values were generated from a mixture of two distributions with equal mixing proportions, the distributions being normal in data sets I and II and quite skewed with heavy tails in data sets III and IV. The following table describes the mixture distributions in detail:

Data Set	Sample Size	True Density
I, II	I: 150; II: 400	$\begin{cases} N(m_1, 0.5I) \text{ with probability } 0.5 \\ N(m_2, 0.5I) \text{ with probability } 0.5 \end{cases}$
III, IV	III: 150; IV:400	$\begin{cases} m_1 - \Psi + b \text{ with probability } 0.5 \\ m_2 + \Psi - b \text{ with probability } 0.5 \end{cases}$



(a) Data I

(b) Data II

(c) Data III

(d) Data IV

Figure 2: Scatter plots of β_h in synthetic data

In this table, $m_1 = (-3, -2, 2.5)'$ and $m_2 = (2, 3, -2.5)'$. In addition, $N(m, \Sigma)$ denotes a normal distribution with mean m and covariance matrix Σ, Ψ denotes a 3-dimensional vector of independent gamma variates from $Ga(a, w)$, with shape parameter $a = 1.5$ and scale parameter $w = 0.5$, that is, $p(\theta|a, w) \propto \theta^{a-1}e^{-w/\theta}$, and $b = (1, 1, 1)'$, where 1 equals the mode of $Ga(1.5, 0.5)$. Both mixture distributions imply that the three elements of the vector β_h are independent *within* each mixture component; but they are clearly not independent overall.

We first generated data sets II and IV, with data sets I and III random samples from them. Fig. 2 displays the scatter plots for two components of β_h for each data set. The sample variance of each component of β_h is $(6.8, 6.5, 6.6)'$ in data set II and $(24.2, 24.6, 26.15)'$ in data set IV. Having generated values for the predictor variables, x_{hjt_h}, and the regression coefficients, β_h, we then used equation (1) to generate 21 purchases for each household, seven purchases for each of the three brands.

Note that the two mixing components of the distribution of β_h have nothing to do with the $J = 3$ brands: we have two sub-populations of households that are quite distinct with respect to their regression coefficients β_h. In all data sets, both sub-populations have the same market shares for the $J = 3$ brands since we generated seven purchase observations per brand in each household.

Data Set I	Data Set II	Data Set III	Data Set IV
0.9608	1.5668	1.0606	2.0569
$(0.4127^{*}; 0.0076^{**})$	$(0.5085; 0.0068)$	$(0.4315; 0.0081)$	$(0.5839; 0.0077)$
$[0.3944, 1.7221]^{***}$	$[0.8450, 2.4939]$	$[0.4557, 1.8614]$	$[1.2008, 3.1007]$

Note: *: standard deviation; **: MC error; *** : [5 percentile, 95 percentile]

Table 1: Estimated α with synthetic data

4.3 Simulation Results

We fitted the proposed model for the logit case with these four synthetic data sets. The chosen prior values, needed in (9) and (10), are $a_\alpha = 0.5, b_\alpha = 4, m_0 = 0_k, v_{\Sigma_0} = 2, V_0 = S_{\Sigma_0} = 20I_k$ (for data sets I and II) and $V_0 = S_{\Sigma_0} = 100I_k$ (for data sets III and IV). Note that data sets III and IV have larger variance for β_h than data sets I and II, and we therefore chose a larger value for the prior parameter S_{Σ_0}.

The prior distribution for α has mean $a_\alpha/b_\alpha = 0.125$, standard deviation 0.18, and $\Pr(\alpha < 1) = 0.995$. For given value of α, the *a priori* expected value for L, the number of mixing components, can be obtained from (7). When $\alpha = 0.125$, $\mathrm{E}(L|\alpha = 0.125) = 1.67$ for data sets I and III. When $\alpha = 1$, $\mathrm{E}(L|\alpha = 1) = 5.6$ for data sets I and III. The corresponding values for data sets II and IV are very slightly higher. Thus, we believe *a priori* that there is at most a moderate number of mixing components.

For all data sets, we ran 40,000 iterations with burn-in period 20,000. After 20,000, all MCMC chains seem to reach stable states. We used the Geweke convergence diagnostic for μ_0, Σ_0, and α. All of these parameters passed the convergence diagnostic in all data sets.

4.3.1 Estimated Number of Mixing Components

Estimate of α First, we examined α, the positive concentration parameter – see Eq. (5). Table 1 gives the estimates of α across these four data sets. A larger sample size tended to increase α, which affects the number of mixing components. Between data sets I and II and between data sets III and IV, the 90% posterior intervals overlap, suggesting that the difference in α may not be significant. For example, the difference in posterior means of α between data sets III and IV is 0.99, and the standard error for this difference is $\sqrt{(.43)^2 + (.58)^2} = 0.73$. Thus the value 0.99 is at best marginally significant. In our simulations, we observed that a larger sample size led to a somewhat larger posterior mean for α. We also saw that a larger sample size had a somewhat more pronounced effect on the posterior mean for α when the true distribution is a mixture of non-normal distributions than when it is a mixture of normal distributions.

Estimate of L Fig. 3 depicts histograms of L for these four data sets. The modes of L were 8, 14, 9, and 18 for data sets I, II, III, and IV, respectively. Other descriptive summary measures are given in Table 2. Again, we observed that larger sample size led to larger L. Furthermore, with posterior estimates and posterior standard deviations, there was a statistically meaningful difference in L both between data sets I and II and between data sets III and IV.

Escobar and West (1995) discussed the effects of sample size, H, with a fixed α, on the expected value of L. When the number of mixing components is likely to be relatively smaller than the sample size, the prior probabilities for L do not vary dramatically with H and tend to decay rapidly as L increases. However, our simulations suggest that the posterior distribution for L can vary quite a lot with H. Furthermore, the simulations clearly show that the posterior modes of L are not the same

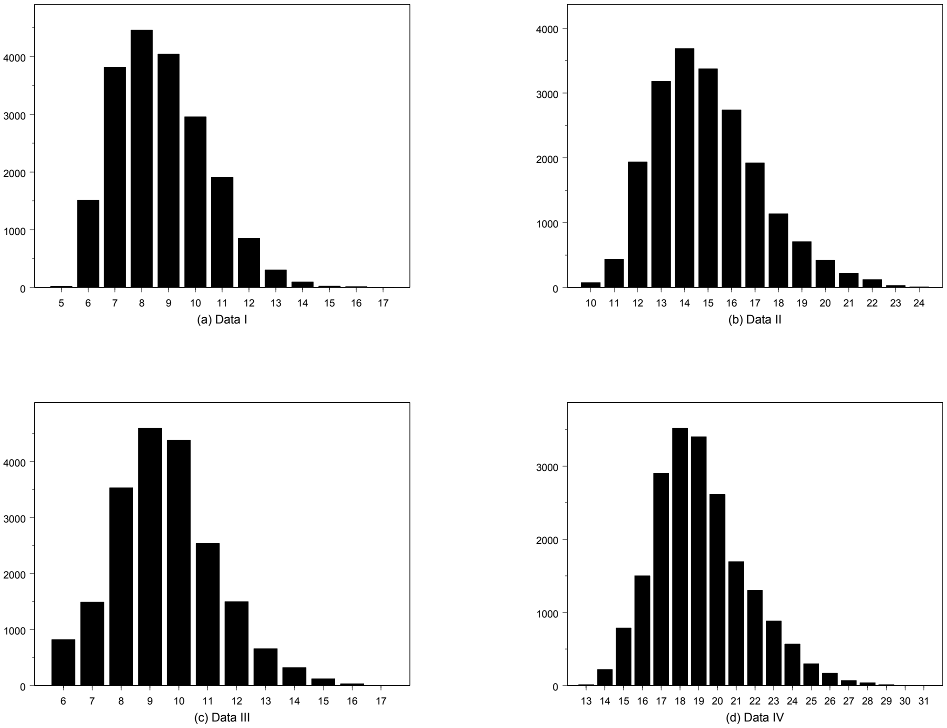


Figure 3: Histogram of simulated L in synthetic data

as the true number of clusters, 2, for any of the data sets. Considering the posterior means and the standard deviations, the true number of mixing components is very far away from the posterior modes of L.

As discussed in Section 2.3, the number of mixing components, L, obtained from our Dirichlet process model is not guaranteed, nor even anticipated, to equal the "true" number of mixing components; even a unimodal distribution can be approximated well by a mixture consisting of several components. As such, the discrepancy between the "correct" number of mixing components (here, 2) and those in Table 2 should be treated as an issue of proper interpretation. For example, clustering can depend on managerial ability to divide up markets and use second-degree price discrimination to optimize relative to that division. For first-degree price-discrimination, by contrast, explicit household allocation to clusters is necessary. We turn our attention to these issues more fully in Section 4.3.4.

	Mode	Mean	StdDev	[Min,Max]
Data Set I	8	8.73	1.73	$[5, 17]$
Data Set II	14	14.97	2.26	$[10, 24]$
Data Set III	9	9.51	1.79	$[6, 17]$
Data Set IV	18	19.11	2.50	$[13, 31]$

Table 2: Summaries for L with synthetic data

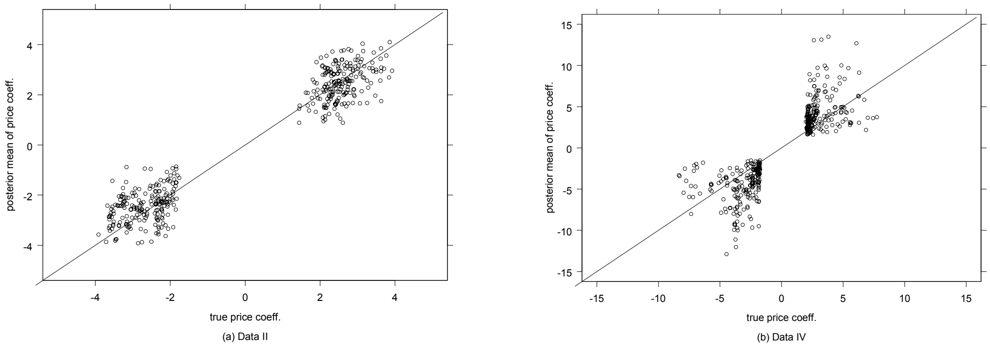


Figure 4: Comparison between estimate and corresponding true value

4.3.2 Estimated Heterogeneity Distribution

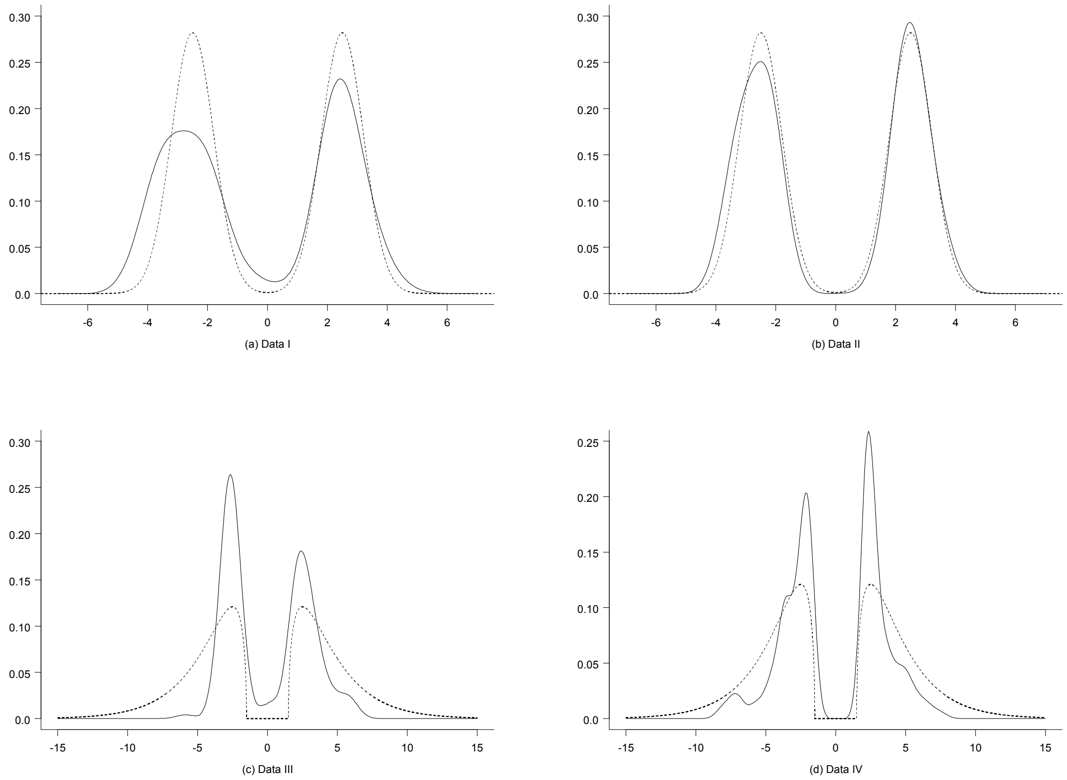
Though our Dirichlet process model does not well estimate the number of mixing components for the distribution of β_h, let us examine how well it estimates the distribution of β_h. For each of our four data sets, this distribution is described in Section 4.2, and a graph is given in Fig. 2. Each MCMC iteration generates a value for the tri-variate vector β_h $(h = 1, ..., H)$ for all H $(= 150$ or $400)$ households. We can thus estimate the posterior density for each tri-variate vector β_h $(h = 1, ..., H)$, and obtain the posterior mean of each β_h as an estimate of the "true" β_h that was used to generate the synthetic data.

Consider the third component of β_h. For data sets II and IV, Fig. 4 gives a plot of the H posterior means of β_h $(h = 1, ..., H)$ for this third component vs. the corresponding true value, and it suggests that the true values are reasonably well recovered by the posterior means in data set II. For data set IV, there are many cases for which the posterior means do not seem to be close to the corresponding true values.

Fig. 5 contrasts the distribution of the H posterior means for the third component of β_h with its true distribution. Not surprisingly, a larger sample size tends to produce better recovery of the true distribution. Note that a larger sample size also tends to produce larger L, so that a sufficiently large number of mixing components, L, may produce better recovery of the true distribution.

For data sets I and II, which were based on a mixture of normal distributions, recovery of the true distribution of the third component of β_h is quite good. This is not the case for data sets III and IV, which were based on a mixture of very skewed distributions. Though, broadly speaking, the fact that the mixture consists of two components is well recovered and the estimated location of the two major modes is generally good, the shape of the skewed distributions making up each of the two mixture components is poorly recovered, especially in data set III. However, in data IV, both mixture components are reasonably well recovered. These observations are also supported by a comparison of both the bivariate contour plots in Fig. 6 and the surface plots in Fig. 7 with the plots in Fig. 2.

We now see that the results in Section 4.3.1 regarding poor estimation of the number of mixing components are not as worrisome as they might at first have appeared. Some of these estimated components are very close and virtually indistinguishable – they are, in short, overlapping – so that the effective number of recovered components is in fact equal to the true number.

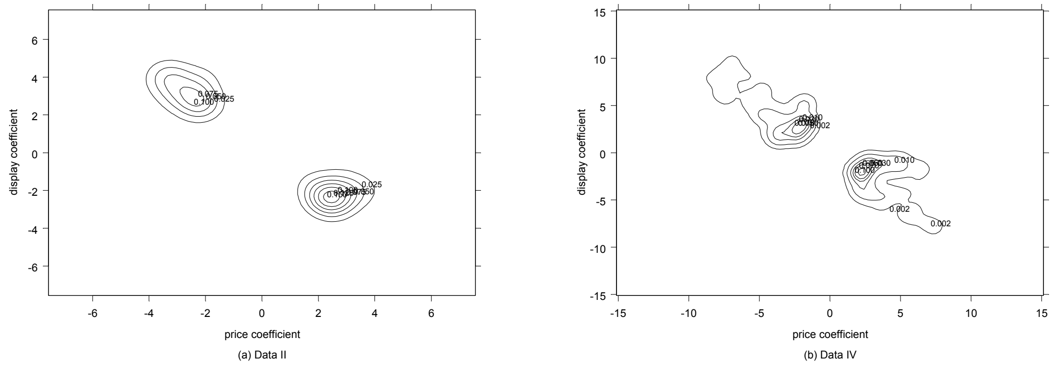

(a) Data I

(b) Data II

(c) Data III

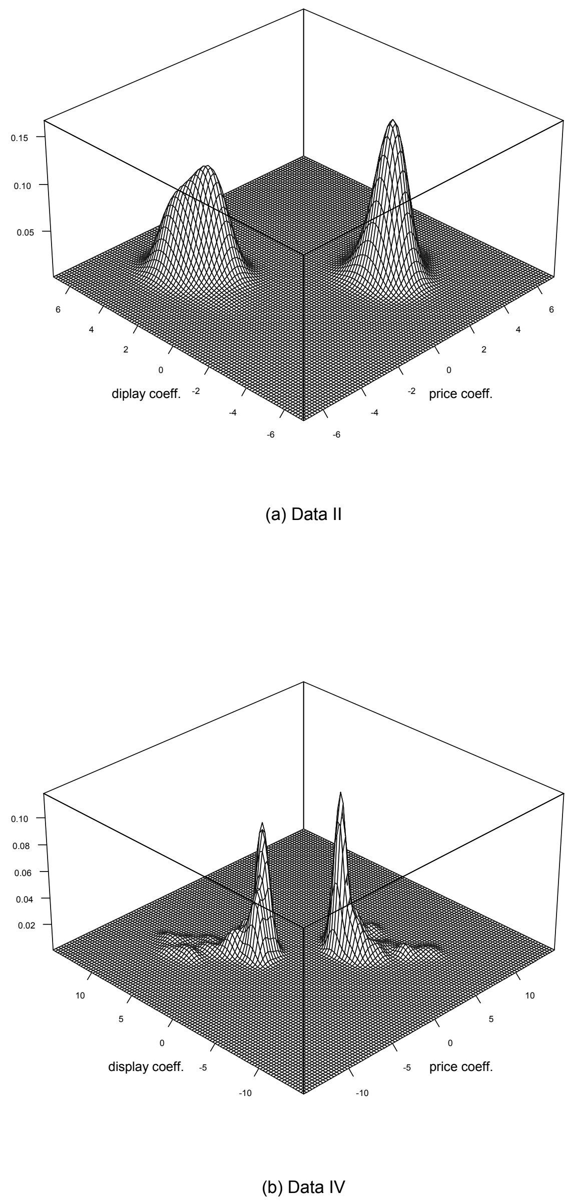
(d) Data IV

Figure 5: Densities for the price coefficient in the synthetic data sets (Estimate: solid line, True: dashed line)


(a) Data II

(b) Data IV

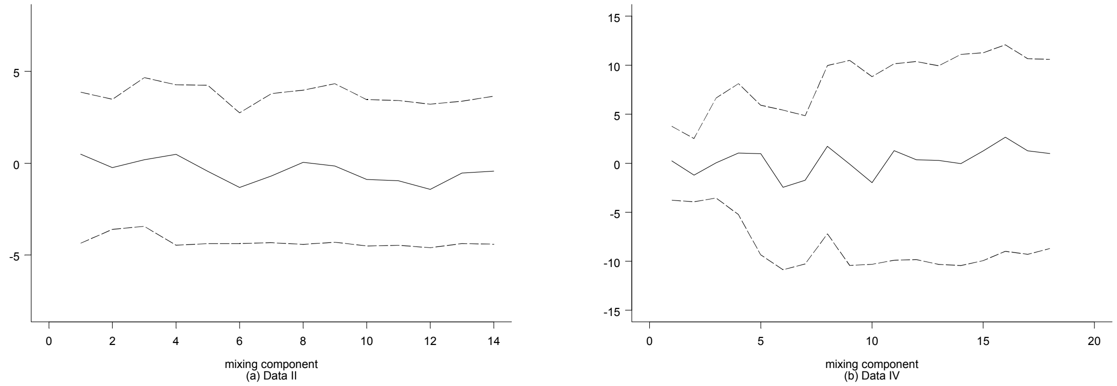
Figure 6: Countour plots of estimated densities in the synthetic data sets



(a) Data II

(b) Data IV

Figure 7: Surface plots of estimated densities in the synthetic data sets



Note: The solid line denotes estimates, the dotted lines 5th and 95th percentiles.

Figure 8: Estimated price coefficient, by mixing component, conditional on the mode of L

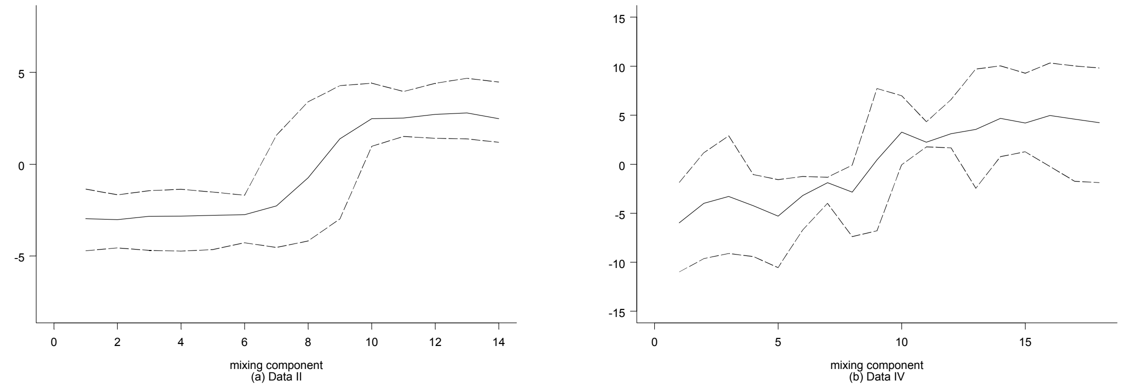
4.3.3 Parameters Associated with Mixing Components

As discussed in Section 2.3, a mixture model can suffer from the label switching problem. To investigate this possibility, we examine parameters associated with the mixing components, that is, the mixing proportions and β^*_l. We condition our analysis on the mode of L, denoted $\hat{L}$. In particular, we describe and estimate the following quantities based on only the MCMC iterations when $L = \hat{L}$,

1. the proportion of the H households in mixing component l $(l = 1, ..., \hat{L})$,
2. the regression coefficient for each mixing component, β^*_l $(l = 1, ..., \hat{L})$, and
3. the probability that household h belongs to mixing component l $(l = 1, ..., \hat{L})$, based on each household's component membership, c_h.

As shown in Fig. 3, the mode of L, $\hat{L}$, is 14 and 18 in data sets II and IV, respectively. Fig. 8 plots estimates of the price coefficients for each mixing component in data sets II and IV, that is, $\beta^*_{l,3}$, $l = 1, ..., \hat{L}$. The 90% posterior intervals for all mixing components strongly overlap, suggesting that the posterior means for all mixing components given $\hat{L}$ appear to be the same. The posterior means of all mixing components essentially equal 0 – the mean of the mixture distribution used to generate β_h (see Section 4.2) – and this suggests the existence of the label switching problem. Note that the middle point between two true clusters is zero.

In order to correct the label switching problem, we post-processed the MCMC output by Stephens' (2000) relabeling algorithm. Fig. 9 plots estimates of $\beta^*_{l,3}$, $l = 1, ..., \hat{L}$ after post-processing. The estimates for some mixing components concentrate around 2.5, for others around -2.5. Recall from Section 3.2 that these were the values of the means of the two mixing components used to generate β_h. Thus, our adaptation of Stephens' algorithm rectified the label switching problem quite well. However, there are strong overlaps: clearly, some mixing components are redundant, in the sense that the 90% posterior intervals overlap. Note that our Dirichlet process model does not well estimate the number of mixing components for the distribution of β_h but it estimates the distribution of β_h reasonably well in data sets II and IV. In other words, the Dirichlet process model approximated the true distribution, with several redundant mixing components, as suggested by Fig. 5.



Note: The solid line denotes estimates, the dotted lines 5th and 95th percentiles.

Figure 9: Estimated price coefficient associated with mixing components conditional on the mode of L after post-processing the MCMC output

For selected households (h = 1, 2, 3, 398, 399, 400), Fig. 10 plots estimates of the probability that the household belongs to mixing component l ($l = 1, ..., \hat{L}$). These estimates were based on each household's component membership, c_h, which is generated at each MCMC iteration. Households 1, 2, and 3 were from one cluster, and households 398, 399, and 400 were from another. Notice, however, that the estimated probabilities in Fig. 10 do not show similar patterns for households in the same true cluster. This may be due to a label switching problem, as depicted in Fig. 8.

Fig. 11 is a plot of $\hat{R}$, the estimated classification probability matrix obtained through the relabeling scheme. A comparison between Figures 10 and 11 shows that the algorithm effectively removed the label switching problem. Fig. 11 further suggests that households coming from the same 'true' cluster tend to exhibit similar patterns in terms of probabilities of households belonging to mixing component l ($l = 1, ..., \hat{L}$). Note as well that the overlapping mixtures problem is also in evidence, as per Fig. 11, and that households from the same true cluster have high probabilities for some mixing components.

It is well-known that the finite normal mixture model (e.g., Allenby *et al.* 1998) also suffers from overlapping mixture problems. Even after estimating the finite mixture model with fixed L after correcting the label switching problem, there is no guarantee that all estimated parameters associated with mixing components are meaningfully separated from one another (e.g., Roeder 1994). Many previous studies using latent class and finite mixture models appear to rely on the presumption that the estimated parameters associated with mixing components (or latent classes) are distinct, in the sense that there exists a finite number of distinct clusters of households in the data set. However, as implied by previous studies in density estimation with mixture models in the statistics literature, this presumption may not always be supported (e.g., Celeux *et al.* 2000).

4.3.4 Clustering Inference

Conditioning on the estimated number of mixing components, $\hat{L}$, we saw that some of the $\hat{L}$ mixing components are strongly overlapping, so that the effective number of mixing components is substantially less than $\hat{L}$. The relabeling algorithm of the last section can be used to identify the effective number of mixing components, to which we next turn our attention.

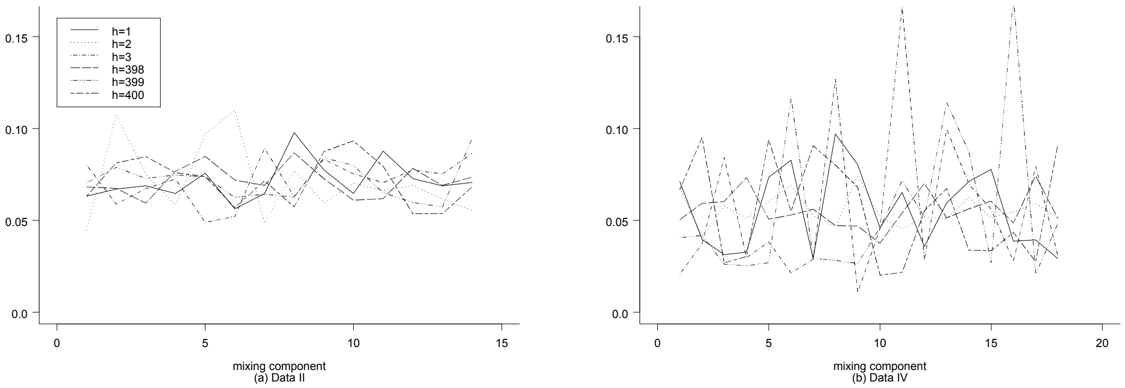


Figure 10: Estimated probabilities of households belonging to mixing component l $(l = 1, ..., L)$ with synthetic data

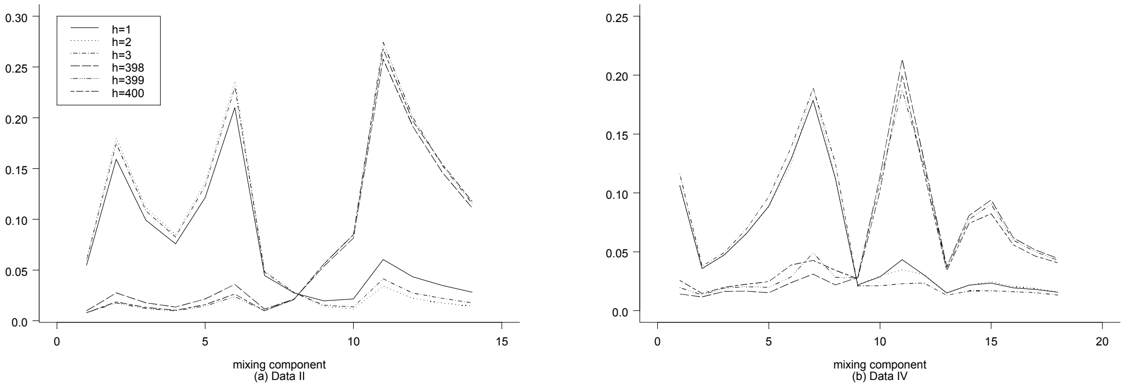


Figure 11: Adjusted probabilities of households belonging to mixing component l $(l = 1, ..., L)$ with Stephens' (2000) relabeling algorithm for synthetic data

	Feature Ad	Display	Price
Data II			
cluster I	-3.0822(0.8726)	-2.1829(0.7631)	2.5278(0.7635)
	[-4.76,-2.09]{-3}	[-3.41,-0.84]{-2}	[1.47,4.08]{2.5}
cluster II	2.1858(0.7754)	3.2651(1.0864)	-2.6636(0.9334)
	[1.09,3.60]{2}	[1.87,4.88]{3}	[-4.43,-1.55]{-2.5}
Data IV			
cluster I	-3.2519(1.7465)	-2.4951(2.0436)	3.1399(1.8741)
	[-6.06,-1.32]{-5}	[-6.96,-0.79]{-4}	[1.84,7.08]{4.5}
cluster II	2.5591(1.8768)	3.3754(2.447)	-3.2170(2.1152)
	[1.06,6.72]{4}	[1.44,9.01]{5}	[-8.17,-1.58]{-4.5}

Note: *: estimate; (): std. dev; ;[]: 90% posterior interval;{}: true mean

Table 3: Estimated regression paramters for two clusters using synthetic data

Recall from Section 3.2 that the $(H \times \hat{L})$ matrix $R = \{r_{hl}\}$ yields the estimated probability that household h belongs to mixing component l, so that the rows of R sum to 1. We first obtained $\tilde{c}_h$, the mixing component corresponding to the largest value of r_{hl} for household h, that is, $\tilde{c}_h = \arg\max\{r_{h1}, ...r_{h\hat{L}}\}$. We then defined the effective number of mixing components to equal $\tilde{L}$, the distinct number of values in $\tilde{c} = \{\tilde{c}_h, h = 1, ..., H\}$. The value of $\tilde{L}$ was 2 for data sets II and IV, that is, it equaled the true number of clusters for these synthetically generated data sets. Then, we checked the hit ratios of $\tilde{c}$ against the true cluster classification. This approach recovered the true cluster composition perfectly for both data sets II and IV. Cluster membership given $\tilde{c}_h$ was identical to true cluster membership. Therefore, the algorithm seemed to be successful in correcting uncertainty about cluster membership that resulted from *both* the overlapping mixtures and label switching.

Then, given the classification scheme $\tilde{c}$, we obtained the posterior distribution of parameters associated with these two clusters, $\tilde{\beta}_l^*$ $(l = 1, 2)$, for data sets II and IV by randomly sampling β_h from $\tilde{\beta}_l = \{\beta_h : \tilde{c}_h = l, h = 1, \ldots, H\}$ at each MCMC iteration. Table 3 presents the summary of the posterior mean of $\tilde{\beta}_l^*$ $(l = 1, 2)$. Broadly speaking, the true means of both clusters were reasonably well recovered for data sets II and IV.

To further check whether the above clustering scheme did work, we plotted the posterior mean of β_h for each household in the synthetic data. In Fig. 12, the number w $(= 1, 2)$ indicates a household assigned to cluster w. As shown in Fig. 12, the clustering scheme appears to work very well indeed.

5 Empirical Application

5.1 Description of the Data and the Prior Distributions

The proposed model is estimated with A. C. Nielsen liquid detergent scanner data. The number of households is $H = 429$, the number of purchase observations is 6682, and the number of brands is $J = 4$. There are three marketing-mix variables: a dummy variable for feature advertising, a dummy variable for display, and net paid price. We also introduce three dummy variables for brands A, B, and C. The market shares of brands A, B, C, and D are 29.7%, 28.3%, 18.8%, and 23.1%, respectively. Further summaries of the marketing-mix variables are given in Tables 4 and 5.

The chosen prior values, needed in (9) and (10), are $a_\alpha = 0.5, b_\alpha = 4, m_0 = 0_k, V_0 = S_{\Sigma_0} = 50I_k$ and $v_{\Sigma_0} = 2$.

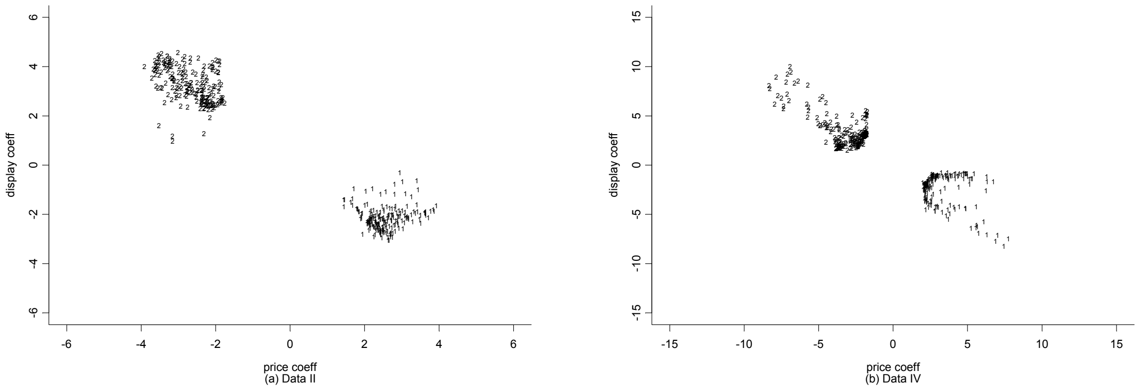


Figure 12: Scatter plot of estimates for household regression parameters in synthetic data sets

Marketing-Mix Variable	Mean (Std. Dev.)
Feature advertising	0.16(0.37)
Display	0.11(0.31)
Net paid price	5.60(0.71)

Table 4: Marketing activities in the liquid detergent data

	Brand A	Brand B	Brand C	Brand D
Market share	29.7%	28.3%	18.8%	23.1%
Proportion of observations with feature advertising	23.3%	22.2%	6.3%	12.2%
Proportion of observations with display	14.6%	14.1%	6.6%	9.9%
Average and standard deviation of net paid price	\$5.09	\$5.95	\$5.93	\$5.42
	(0.58)	(0.64)	(0.51)	(0.68)

Table 5: Marketing activities of brands in the liquid detergent data

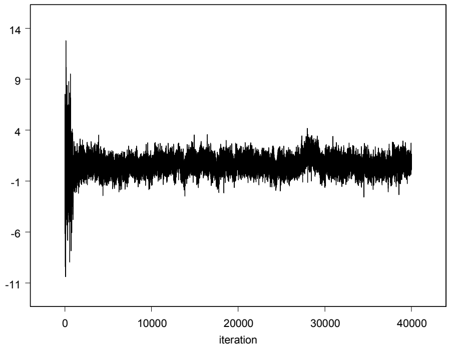


Figure 13: Trace plot of $\mu_{0,3}$

The prior distribution for alpha has mean $a_\alpha/b_\alpha = 0.125$, standard deviation 0.18, and $\Pr(\alpha < 1) = 0.995$. For given value of α, the *a priori* expected value for L, the number of mixing components, can be obtained from (7). For various values of α, Table 6 lists the corresponding prior expected value for L. In particular, when $\alpha = 0.125$, $E(L|\alpha = 0.125) = 1.81$, and when $\alpha = 1$, $E(L|\alpha = 1) = 6.64$. Escobar and West (1995) examined the sensitivity of the number of mixing components to varying values of α. As α increased, the posterior mean for L became somewhat larger, but its distribution was only slightly altered.

To demonstrate the present methodology, owing to its greater transparency, we focus on the logit model. Estimation of the probit model is straightforward, and the present discussion applies to it as well largely unchanged (results are available from the authors). The burn-in period was 20,000 iterations, and convergence of the MCMC sampler was assessed by using Geweke's (1992) convergence diagnostics. Among the parameters α, μ_0, and Σ_0, only a single element in μ_0, $\mu_{0,3}$, failed to pass the diagnostics. Fig. 13 is the trace plot of $\mu_{0,3}$: even though there was a slight increase around iteration 28,000, the sampled values for $\mu_{0,3}$ seem to have converged after 6000 iterations. Since μ_0 is the hyperparameter for G_0, we expect the small degree of fluctuation around iteration 28,000 to have a minimal effect on the estimation results. The results in the remainder of this section were based on the last 20,000 MCMC iterations.

5.2 Estimation Results for the Logit Model

5.2.1 Estimation of α and L

The trace plot in Fig. 14 suggests that α converged well before iteration 20,000. Fig. 15 shows that the posterior distribution of α is quite different from its prior distribution. The posterior mean and standard deviation of α are 4.91 and 0.95, respectively, and the posterior interval for α that extends from the 5th to the 95th percentiles (for short, the "90% posterior interval" for α) is $[3.42, 6.55]$. Based on Table 6 and Eq. (7), such prior values for α would have implied a far larger *a priori* expected value for L. For example, for $\alpha = 4.91$, $E(L|\alpha = 4.91) = 23$. Let us now examine the posterior distribution for L which incorporates both this prior information and the data.

The trace plot for L, the number of mixing components, in Fig. 16 suggests that L converges quickly. The posterior distribution for L (as usual, a histogram of L across all MCMC iterations) in

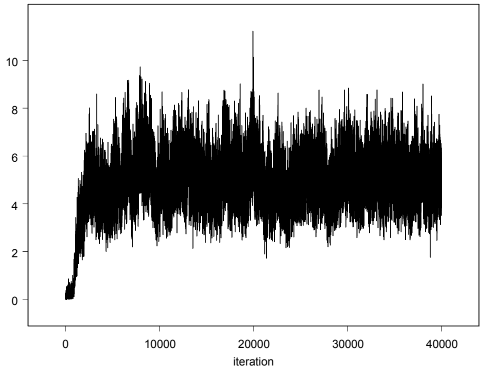


Figure 14: Trace plot of α

Fig. 17 shows that the posterior mode of L is 42 and that the 90% posterior interval for L is $[35, 48]$. The number of mixing components is clearly quite large. As discussed in Section 4 with synthetic data sets, this, however, does not imply that there is a large number of distinct substantive clusters, because mixing components may overlap with one another.

5.2.2 Distribution of the Regression Parameters

Fig. 18 presents plots for the posterior distributions of the six components of β. These plots were produced by a standard kernel density estimation techniques, for which we used Sheather and Jones' (1991) derivative-based method to determine window width. In Fig. 18, the simulated β values are indicated by ticks. All plots, except that for feature advertising, suggest at least partial departures from unimodality.

Figures 19 and 20 give contour plots and surface plots based on the posterior distribution for β, and also indicate several modes.

Tables 7 and 8 give estimates for μ_0 and Σ_0, the hyper-parameters for G_0. The coefficients of the dummy variable of brand B, display, and price are meaningfully different from zero since their 90% posterior intervals do not contain zero. The other three coefficients seem to be essentially zero. All diagonal elements of Σ_0 seem to be different from 0 since the posterior means are at least twice as large as the posterior standard deviations. Among the off-diagonal elements of Σ_0, only the covariance between price and Brand B, and possibly that between price and Brand C, are different from 0.

5.2.3 Analysis on Mixing Components

In Section 5.2.1 we examined the posterior distribution for L, the number of mixing components, and we found the mode for L to be 42, rather a large number. As discussed in Section 2.3, a large number of mixing components is not necessarily indicative of a large number of distinct clusters of households, and the number of "genuine" clusters can be quite a bit less than L. A casual investigation of the

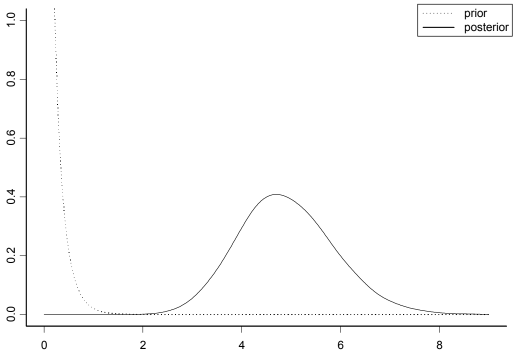


Figure 15: Density of α

Expected No. of Mixing Component	α
1	$2.0e-3$
2	0.2
3	0.4
4	0.5
5	0.7
6	0.9
7	1.1
8	1.3
9	1.5
10	1.7
15	2.8
20	4.1
25	5.5
30	6.9
35	8.5
40	10.2
45	11.9
50	13.8
100	38.1
200	125.1
300	326.1
491	$1.21e5$

Table 6: *A priori* expected number of mixing components given α (H=429)

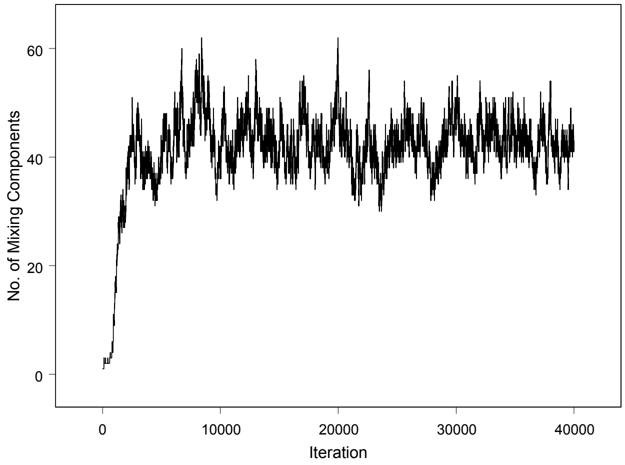


Figure 16: Trace plot of number of mixing components

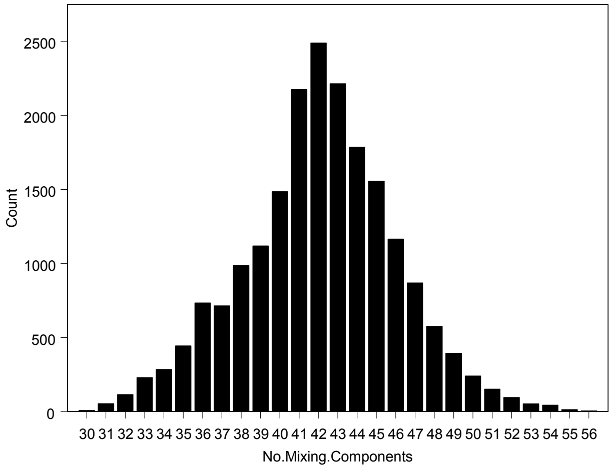


Figure 17: Histogram: number of mixing components

	Mean	StdDev	[5 percentile, 95 percentile]
Brand A	-0.701	0.568	[-1.656, 0.224]
Brand B	1.079	0.626	[0.060, 2.124]
Brand C	0.693	0.699	[-0.401, 1.887]
Feature	0.312	0.307	[-0.196, 0.815]
Display	0.595	0.309	[0.095, 1.102]
Price	-2.175	0.603	[-3.208, -1.239]

Table 7: Posterior mean and standard deviation for μ_0

	Brand A	Brand B	Brand C	Feature	Display	Price
Brand A	8.2886	0.0347	0.1451	−0.0327	−0.0257	0.0316
	(2.8195)					
Brand B	0.3051	9.3487	0.1834	−0.0212	−0.0196	−0.5268
	(1.8682)	(2.9199)				
Brand C	1.4324	1.9230	11.7652	0.0508	−0.1054	−0.3700
	(2.3483)	(2.2710)	(3.8531)			
Feature	−0.1492	−0.1025	0.2763	2.5110	0.0046	0.0337
	(0.9343)	(0.9917)	(1.1161)	(0.7379)		
Display	−0.1172	−0.0950	−0.5725	0.0115	2.5053	0.0846
	(0.9977)	(1.0163)	(1.1868)	(0.4884)	(0.7254)	
Price	0.2776	−4.9130	−3.8708	0.1631	0.4083	9.3026
	(1.7105)	(2.3707)	(2.2624)	(1.0031)	(1.0012)	(3.3095)

Note: 1) lower triangular matrix: estimated covariance matrix
2) posterior standard deviations are in parentheses
3) upper triangular matrix: correlation matrix

Table 8: Estimated Σ_0

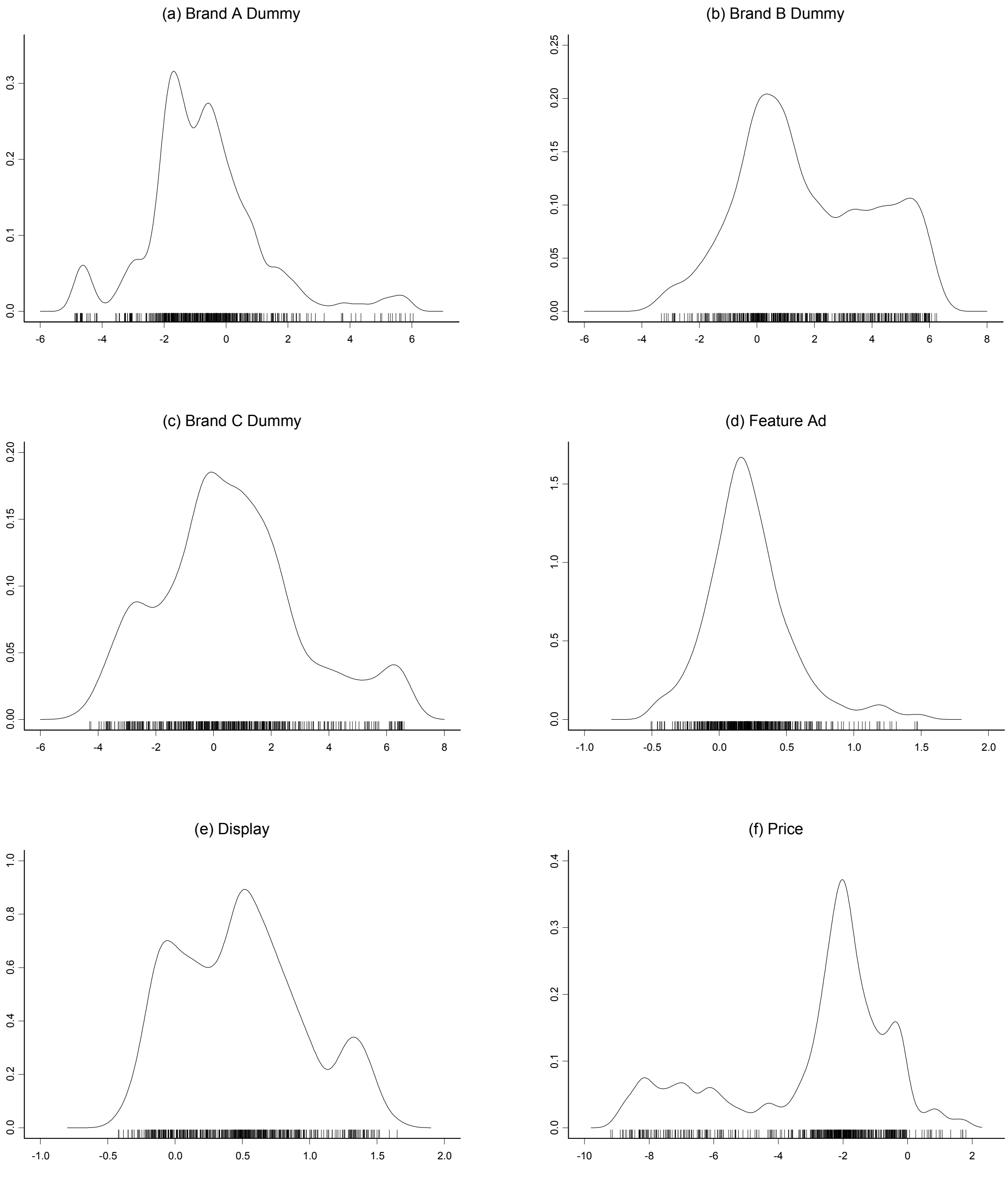


Figure 18: Densities for the components of β_h

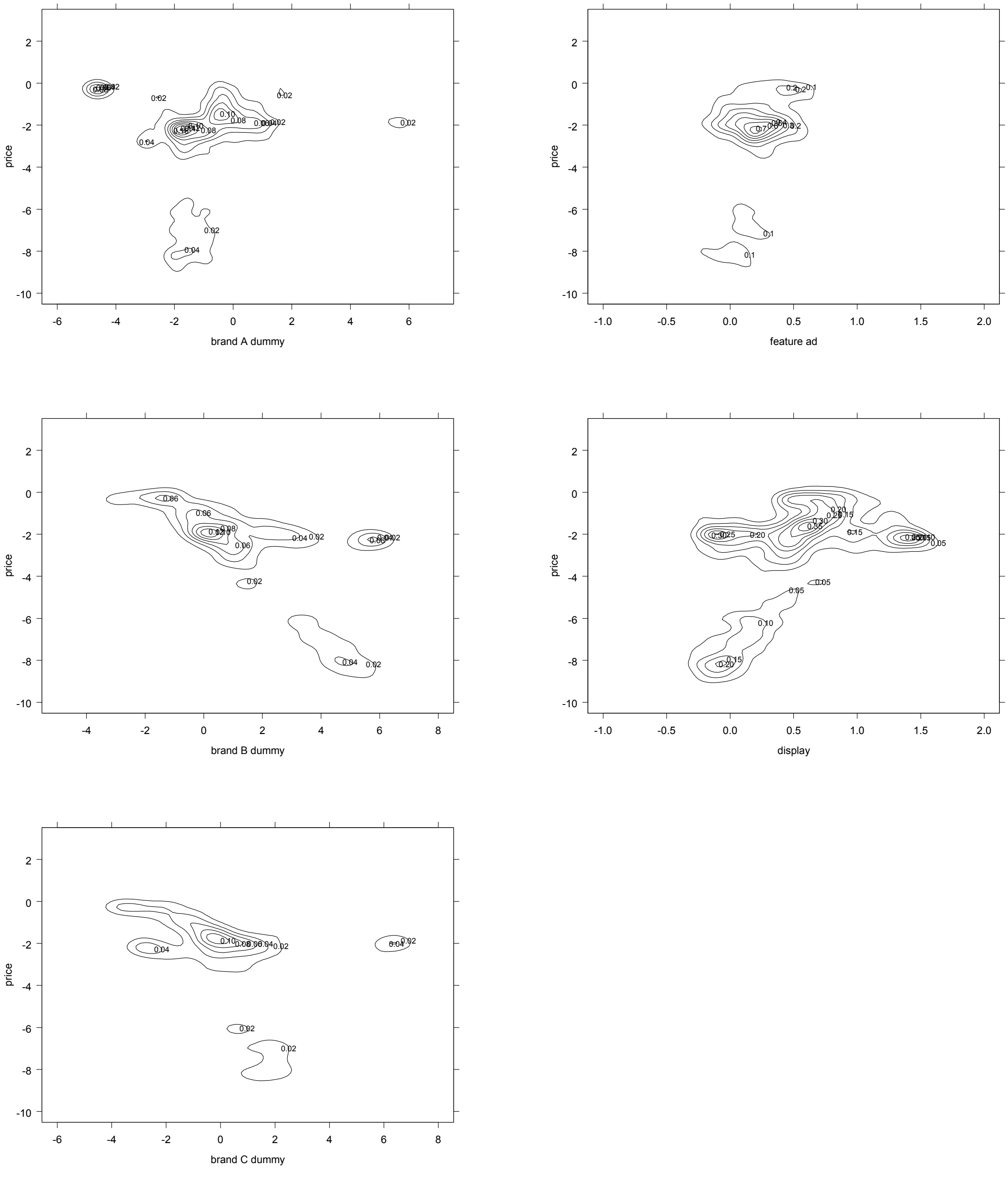


Figure 19: Contour plots based on the posterior density for β_h

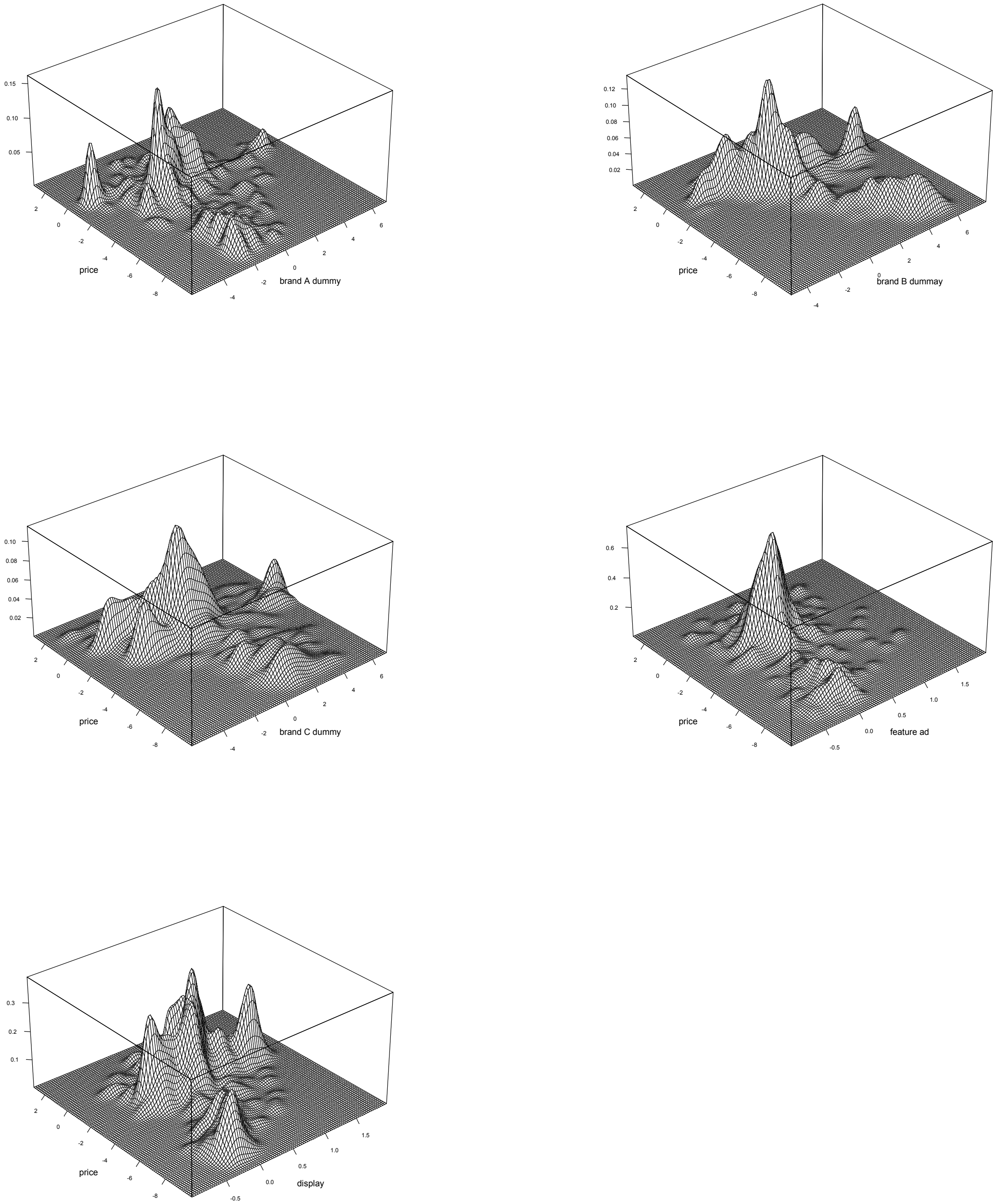


Figure 20: Surface plots based on the posterior density for β_h

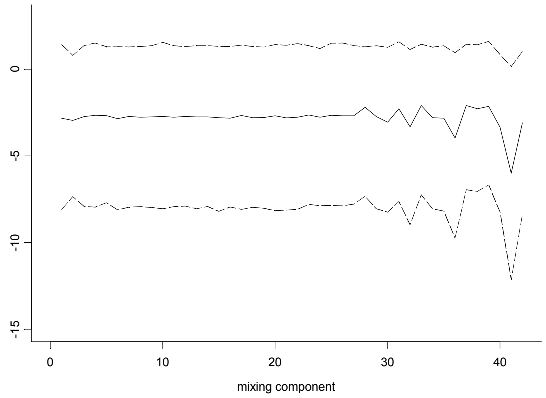


Figure 21: Estimates of the price component of β_i^* $(i = 1, \ldots, 42)$ conditional on $L = 42$

plots discussed in Section 5.2.2 suggests that there may be fewer than 42 clusters and that we may have strongly overlapping mixing components.

We first corrected the label switching problem using the algorithm in Stephens (2000), as discussed in Section 3.2. For the regression coefficients associated with the resulting mixing components, we then obtained the posterior distributions which were based on the values of β_i^* $(i = 1, ..., L)$ for those MCMC iterates for which L was 42.

Let us now concentrate on the element of these vectors that corresponds to net paid price, which we denote the "price coefficient". Fig. 21 presents the posterior means and 90% posterior intervals for the 42 price coefficients after post-processing the MCMC output by Stephens' algorithm; that there do not seem to be pronounced differences among the 42 mixing distributions for the price coefficient, even after controlling for the label switching problem, suggests a pronounced problem with overlapping mixing components.

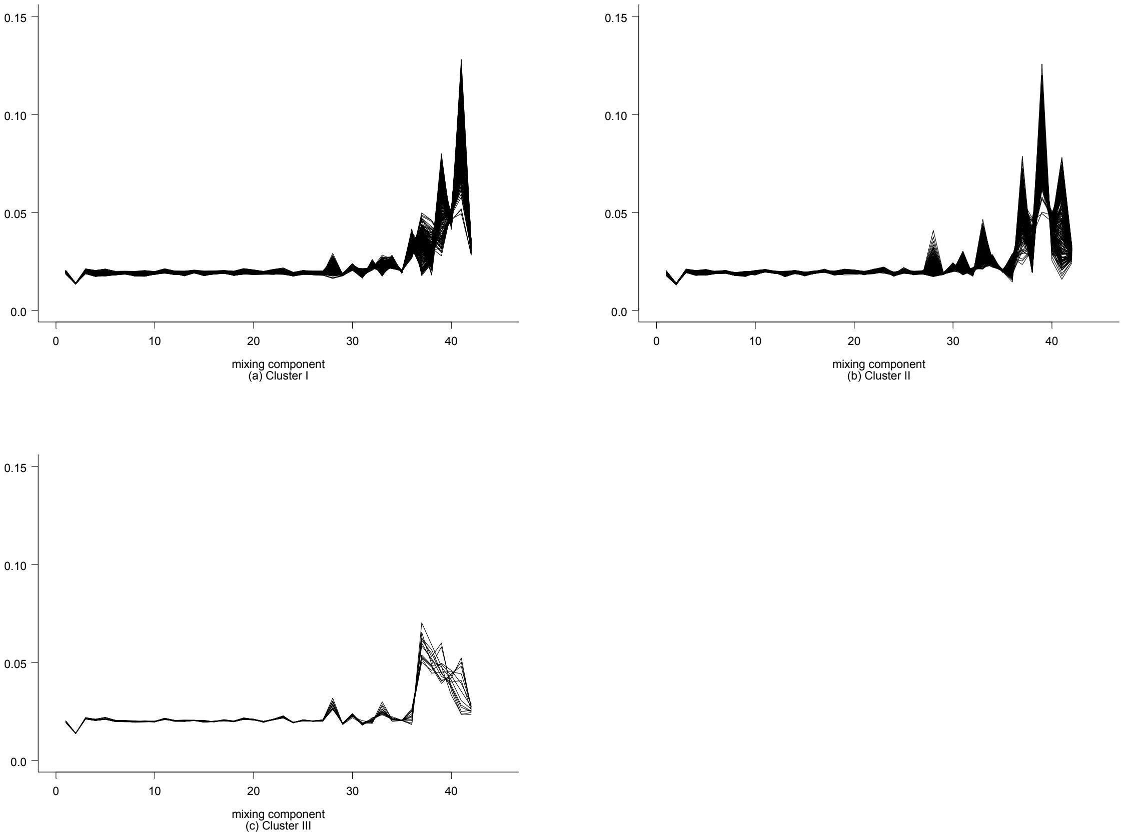
We next take up the pragmatic issue of determining whether, and how, these overlapping mixtures cluster into meaningfully distinct segments.

5.2.4 Clustering and Segmentation Inference

As illustrated in Sections 4.3.3 and 4.3.4, we use the classification probability matrix, R, produced by Stephens' (2000) algorithm. Given an estimate of R, we find $\tilde{c}_h$ so that $\tilde{c}_h = l$ if $r_{hl} = \max(r_{hl}, \ldots, r_{hL})$, $l = 1, \ldots, L$. The number of distinct values in $\tilde{c} = \{\tilde{c}_1, \ldots, \tilde{c}_H\}$, $\tilde{L}$, was 3. The sizes of these three clusters, I, II, and III were 278 (56.5%), 201 (40.9%), and 13 (2.6%), respectively.

Fig. 22 graphs the estimated value of r_{hl} for the households belonging to each cluster. Households belonging to the same cluster tend to show a similar pattern in terms of estimated classification probabilities across mixing components. This suggests that the three cluster scheme based on $\tilde{c}_h$ may succeed in grouping households that are close in terms of β_h.

As suggested in Section 4.3.3, if the new cluster scheme $\tilde{c}_h$ corrected the overlapping mixing problem, there might be significant differences among the three clusters with respect to the regression coefficients. Fig. 23 offers a graphical representation of this with respect to three components of



(a) Cluster I

(b) Cluster II

(c) Cluster III

Figure 22: Plot of R for each cluster

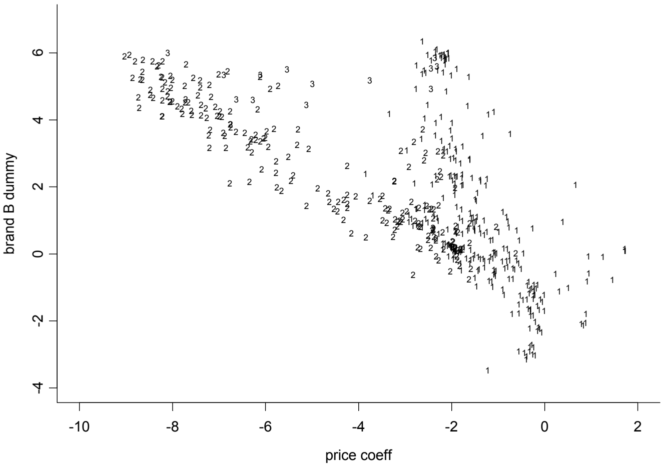


Figure 23: Plot of posterior means of the 429 regression coefficient pairs for price and the Brand B dummy

β_h. Note that the cluster index i $(i = 1, \ldots, 3)$ represents the posterior mean of β_h for each of the $H = 429$ households. Cluster separation is discernible, and households belonging to the same cluster are concentrated largely in the same part of the graph.

Given this new classification scheme of three clusters, $\tilde{c}$, we obtained the posterior distribution of parameters associated with each, $\tilde{\beta}_l^*$ (l =I, II, III) by randomly sampling one value from $\tilde{\beta}_l = \{\beta_h : \tilde{c}_h = l, h = 1, \ldots, H\}$, l =I, II, III at each MCMC iteration. Table 9 gives the posterior means and standard deviations for the regression coefficients across these three clusters. We find that all three clusters highly overlap, as suggested by Fig. 21.

This comparison is, of course, based on the marginal posterior distributions for the regression coefficients in each cluster, and it assumes that these regression coefficients are not highly correlated. This is partially supported by the posterior correlations listed in Fig. 19. To get a visual feel for the correlation between regression coefficients in the set of households, Fig. 23 plots the $H = 429$ pairs of posterior means for the regression coefficients of the Brand C dummy and net paid price. Cluster membership is again indicated by digits $i = 1, 2, 3$.

The correlation here appears to be moderate, and we in fact suspect the correlation may lead to overlapped clusters. To see this point, let us consider Fig. 24, which depicts a hypothetical contour plot of two such clusters. Given the axes x and y, two clusters highly overlap. However, relative to the rotated coordinate system of x' and y', the overlapping clusters essentially disappear, and are seen as quite distinct. In this way, correlation may lead to the type overlapping mixtures evident in Fig. 21, even after correcting for any label switching problems.

	Cluster I	Cluster II	Cluster III
Brand A dummy	-1.0813*(2.3171)**	-0.2493(2.3728)	-1.5215(1.5600)
	[-5.18,2.18]***	[-3.22,4.74]	[-3.96,1.10]
Brand B dummy	0.9196(2.6729)	2.5920(2.6446)	5.2199(2.0360)
	[-2.88,5.92]	[-1.08,6.94]	[1.51,8.53]
Brand C dummy	0.1134(3.2353)	1.3219(2.2836)	-0.1754(2.6020)
	[-4.42,6.32]	[-1.79,5.50]	[-4.95,3.63]
Feature ad	0.2485(0.8146)	0.1745(0.8978)	0.1831(0.8942)
	[-1.06,1.56]	[-1.19,1.77]	[-1.15,1.73]
Display	0.6126(0.9046)	0.2797(0.8663)	0.5413(1.1098)
	[-0.84,2.18]	[-0.98,1.84]	[-0.10,2.46]
Price	-1.3424(1.3872)	-4.9521(3.3352)	-4.8219(3.3325)
	[-3.53,0.83]	[-10.63,-0.93]	[-10.51,-1.05]

Note: *:posterior mean; **: posterior std. dev.;***: [5 percentile, 95 percentile]

Table 9: Posterior means and standard deviations for the regression coefficients in the three clusters

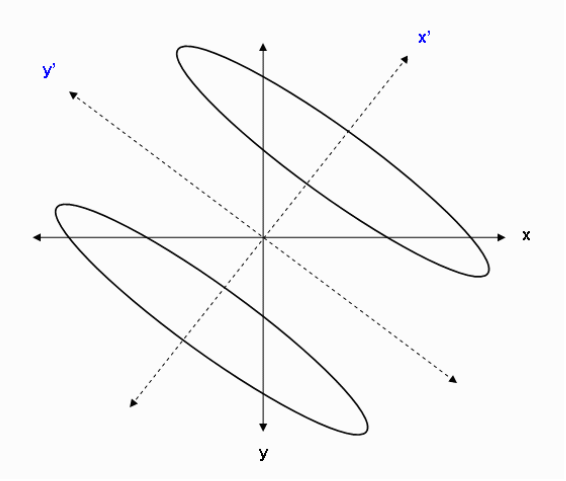


Figure 24: Illustration of the occurence of overlapping clusters by correlation

6 Discussion

In this paper, we proposed a joint methodology that links observed brand choice to marketing and other variables via a logit, probit or other suitable link, and that uses a Dirichlet process prior to capture heterogeneity in regression coefficients across households. We developed an MCMC algorithm for the model with a logit and a probit link, and we applied the logit model to some liquid detergent scanner data.

The resulting distribution of regression coefficients across households is a mixture, capable of capturing two important features: (1) there can be several distinct modes representing meaningfully separated clusters of households, and (2) the distribution of regression coefficients in each cluster is flexible and need not be normal. The overall number of mixing components, however, is typically greater than the number of distinct clusters of households, an artifact of the need to approximate each of the possibly non-normal distributions in the clusters by separate mixtures of several mixing components each.

One of the strengths of the model is that it determines the overall number of mixing components in a straightforward way. In order to determine the meaningfully separate clusters of households, however, we had to add a separate *post hoc* step to combine several mixing components into a cluster. Having identified an appropriate number of mixing components, this step consists of (1) overcoming the labeling problem inherent in mixture models by using the algorithm proposed by Stephens (2000), and (2) using the matrix of estimated classification probabilities to find a small number of household clusters that are intended to meaningfully separate the households into 'substantive' segments.

Our overall modeling approach thus consisted of two consecutive stages, first using the Dirichlet process prior model to estimate the distribution of regression coefficients to capture heterogeneity among households, then parceling households into meaningful clusters. It would be desirable to design an approach that formally combines these two stages, and we hope to report on such a modeling framework in the future.

7 Colophon

Jin Gyo Kim (kimjg@mit.edu), MIT Sloan School of Management, 38 Memorial Drive, E56-323, Cambridge, MA, 02142; Ulrich Menzefricke (menzefricke@rotman.utoronto.ca), Rotman School of Management, University of Toronto, 105 St. George St., Toronto, Ontario, Canada, M5S 3E6; and Fred Feinberg (feinf@umich.edu), University of Michigan Business School, 701 Tappan St., Ann Arbor, Michigan, USA, 48109. Correspondence regarding this manuscript should be addressed to the first author by e-mail to jgkim@mit.edu, or by surface mail to the above address. The authors wish to thank John D. C. Little, Greg Allenby, Michel Wedel and Thomas Otter for their suggestions. All comments welcome.

7.1 Appendix: Estimation of the Probit Model

Whereas the choice probabilities for the logit model in (1) are available in closed form, that is not the case for the probit model in (3), where multi-dimensional integration is required. To avoid explicit multi-dimensional integration, it is convenient to incorporate the unobservable utilities into the MCMC sampling scheme. In the probit model, these utilities are $u_{ht_h} = x_{ht_h}\beta_h + \varepsilon_{ht_h}$, where $\varepsilon_{ht_h} \sim N(0, \Sigma)$.

This utility formulation for the probit model suffers from location and scale identification problems. Several approaches have been proposed to overcome the identification problem (e.g., McCulloch and Rossi 1994). Here, we follow Albert and Chib's (1993) approach, which restricts Σ to be a correlation matrix. Since all diagonal elements in Σ are 1, the number of unknown quantities in Σ becomes $J \times (J-1)/2$. The location identification problem can be avoided by letting one of the brands, say brand J, be the reference brand and setting its latent Gaussian utilities equal to 0, or by setting the intercept of J to be 0 if all brands have their own intercepts.

Letting $u = (u_1, ..., u_H)$, with $u_h = (u_{h1}^{'}, ..., u_{hT_h}^{'})^{'}$, we thus must incorporate the additional variables u and Σ into the MCMC sampler by using the following conditional distributions of the full joint posterior distribution for $p(u, \Sigma, \beta, \alpha, \mu_0, \Sigma_0|y)$:

1. $p(u|\Sigma, \beta, \alpha, \mu_0, \Sigma_0; y)$
2. $p(\Sigma|u, \beta, \alpha, \mu_0, \Sigma_0; y)$
3. $p(\beta_h|u, \Sigma, \beta_{-h}, \alpha, \mu_0, \Sigma_0; y)$ for each $h = 1, ..., H$.
4. $p(\beta_i^*|u, \Sigma, L, \mu_0, \Sigma_0; y)$ for each $i = 1, ..., L$. Here S denotes the cluster structure, that is, $S = (S_1, ..., S_H)$, with $S_h = i$ if $\beta_h = \beta_i^*$ $(h = 1, ..., H)$.
5. $p(\mu_0, \Sigma_0|u, \Sigma, \beta, \alpha; y)$, and
6. $p(\alpha|u, \Sigma, \beta, \mu_0, \Sigma_0; y)$.

Let us discuss these distributions in turn.

7.1.1 Sampling From the Conditional Distribution of u

The conditional distribution for each u_{ht_h} is

$$p(u_{ht_h}|\Sigma, \beta, \alpha, \mu_0, \Sigma_0; y) = p(u_{ht_h}|\Sigma, \beta_h; y) \propto N_J(x_{ht_h}\beta_h, \Sigma) \prod_{i=1, i\neq j}^{J} I_{u_{hit_h} < u_{hjt}},$$

where the chosen brand is $y_{ht} = j$. Each element of u_{ht_h} can be efficiently sampled by the inversion method, which is an efficient method for sampling from a truncated distribution. Let us suppose that we want to sample from a doubly truncated distribution, $F(w|\bullet)I_{a<w<b}$, where $F(\bullet)$ is a target probability density function. Let $\Phi(\bullet)$ denote the cumulative distribution of $F(\bullet)$. Then, sample w so that $w = \Phi^{-1}(o)$, where o is a value uniformly selected from an interval $(\Phi(a), \Phi(b))$.

7.1.2 Sampling From the Conditional Distribution of Σ

Recall that the $(J \times J)$ covariance matrix for the utilities, Σ, is restricted to be a correlation matrix for identification purposes. Let $\Omega = \{\sigma_{ij}\}$, $i < j$, denote all unknown quantities in Σ. Then, we assume the prior distribution for Ω to be

$$p(\Omega) \propto \prod_{i<j} n(\sigma_{ij}|a, s^2) I_A I_B,$$

where $n(\sigma_{ij}|a, s^2)$ denotes a univariate normal distribution with mean a and variance s^2, A denotes the event that Σ is a positive definite matrix, and B denotes the event that all elements of Σ are in [-1,1].

The conditional posterior distribution for Σ thus is

$$p(\Sigma|u, \beta, \alpha, \mu_0, \Sigma_0; y) = p(\Sigma|u, \beta) \propto \left(\prod_{h=1}^{H} \prod_{t_h=1}^{T_h} N_J(u_{ht_h}|x_{ht_h}\beta_h, \Sigma) \right) p(\Omega),$$

where $N_J(\cdot|x_{ht_h}\beta_h, \Sigma)$ denotes the J-dimensional normal density with mean $x_{ht_h}\beta_h$ and correlation matrix Σ.

This conditional posterior distribution can be easily sampled by a slice sampler (Neal 1997). Slice sampling is a form of the auxiliary variable technique for facilitating the design of an improved MCMC sampling algorithm. Define a function

$$f(\{\sigma_{ij}\}|\{\sigma_{ij}\}^-) = \left(\prod_{h=1}^{H} \prod_{t_h=1}^{T_h} N_J(u_{ht_h}|x_{ht_h}\beta_h, \Sigma) \right) n(\sigma_{ij}|a, s^2) I_{-1 \le \sigma_{ij} \le 1} I_A, i \neq j,$$

where $\{\sigma_{ij}\}^-$ denotes a set of all elements in Ω except σ_{ij}. Then, sample σ_{ij} by using the slice sampler.

7.1.3 Sampling From the Conditional Distribution of β_h

As given in (6), we need the following quantities for household $h = 1, ..., H$.

- $q_{0h} \propto \alpha \int \left(\prod_{t_h=1}^{T_h} n_J(u_{ht_h}|x_{ht_h}\beta_h, \Sigma) \right) n(\beta_h|\mu_0, \Sigma_0) d\beta_h = \alpha n_{JT_h}(u_h|X_h\mu_0, X_h\Sigma_0 X_h^t + I_{T_h} \otimes \Sigma)$, where u_h is a JT_h-dimensional vector, $X_h = (x_{h1}^{'}, ..., x_{hT_h}^{'})^{'}$ is an $(JT_h \times k)$ matrix, I_{T_h} is a $(T_h \times T_h)$ identity matrix, and $\otimes$ denotes the Kronecker product,
- $q_{ih}^* \propto \prod_{t_h=1}^{T_h} N_J(u_{ht_h}|x_{ht_h}\beta_i^*, \Sigma)$ is the likelihood for u_h conditional on $\beta_h = \beta_i^*$,
- q_{0h} and q_{ih}^*, $i = 1, ..., L$, are such that $1 = q_{0h} + \sum_{i=1}^{L} n_{ih} q_{ih}^*$,
- $G_b(\beta_h|\mu_0, \Sigma_0; y) = N(\mu^*, \Sigma^*)$, where $\mu^* = \Sigma^* \left(\sum_{t_h=1}^{T_h} \Sigma^{-1} u_{ht_h} + \Sigma_0^{-1}\mu_0 \right)$ and $\Sigma^* = \left(\sum_{t=1}^{T_h} x_{ht_h}^{'} \Sigma^{-1} x_{ht_h} + \Sigma_0^{-1} \right)^{-1}$,

A new value of β_h can now be readily sampled in two steps:

1. Draw a cluster label at random from the integers $\{0, 1, ..., L\}$ with distribution $\{q_{0h}$, $q_{1h}^*, ...,$ $q_{Lh}^*\}$. Denote this label c_p.
2. If $c_p \in \{1, ..., L\}$, let $\beta_h = \beta_{c_p}^*$. If $c_p = 0$, draw the new value for β_h from $N(\mu^*, \Sigma^*)$.

7.1.4 Sampling From the Conditional Distribution of β_i^*

Let $\mathcal{H}_i$ denote the set of households for which $\beta_h = \beta_i^*$ $(i = 1, ..., L)$. Then the conditional posterior distribution for β_i^* is $N(\mu_i^*, \Sigma_i^*)$, where $\mu_i^* = \Sigma_i^*(\Sigma_0^{-1}\mu_0 + \sum_{h \in \mathcal{H}_i} \sum_{t_h=1}^{T_h} x_{ht_h}^{'} \Sigma^{-1} x_{ht_h})$ and $\Sigma_i^* = (\Sigma_0^{-1} + \sum_{h \in \mathcal{H}_i} \sum_{t_h=1}^{T_h} x_{ht_h}^{'} \Sigma^{-1} x_{ht_h})^{-1}$.

7.1.5 Sampling From the Conditional Distribution of μ_0, Σ_0, and α

Sampling from the conditional distributions for μ_0, Σ_0, and α is identical to the case of the logit model.

References

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

[2] Allenby, G. M., N. Arora, and J. L. Ginter (1998), “On the Heterogeneity of Demand,” *Journal of Marketing Research*, 35, 384-389.

[3] Allenby, Greg M. and Peter E. Rossi (1999), “Marketing Models of Consumer Heterogeneity,” *Journal of Econometrics*, 89 (March/April), 57-78.

[4] Andrews, Rick, L., and Imran S. Currim (2001), “A Comparison of Segment Retention Criteria for Finite Mixture Models,” working paper, Department of Business Administration, University of Delaware.

[5] Andrews, Rick L., Andrew Ainslie and Imran S. Currim (2002), “An Empirical Comparison of Logit Choice Models with Discrete vs. Continuous Representations of Heterogeneity,” working paper.

[6] Antoniak, C. E. (1974), “Mixtures of Dirichlet processes with Applications to Bayesian Nonparametric Problems,” *Annals of Statistics*, 2, 1152-1174.

[7] Blackwell, D. and J. B. MacQueen (1973), “Ferguson Distributions via Pólya Urn Schemes,” *Annals of Statistics*, 1, 353-355.

[8] Celeux, G., M. Hurn, and C. P. Robert (2000), “Computational and Inferential Difficulties with Mixture Posterior Distributions,” *Journal of the American Statistical Association,* 95, 451, 957-970.

[9] Chamberlain, G. (1980), “Analysis of Covariance with Qualitative Data,” *Review of Economic Studies*, 47, 225-238.

[10] Chib, S. and E. Greenberg (1998), “Analysis of Multivariate Probit Models,” *Biometrika*, 85, 2, 347-361.

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

[12] Diebolt, J. and C. Robert (1994), ”Estimation of Finite Mixture Distributions Through Bayesian Sampling,” *Journal of Royal Statistical Society*, Ser. B, 56, 363-375.

[13] Escobar, M. D. (1994),“ Estimating Normal Means with a Dirichlet process Prior,” *Journal of the American Statistical Association*, 89, 425, 268-277.

[14] ______ and M. West (1995), “Bayesian Density Estimation and Inference Using Mixtutes,” *Journal of the American Statistical Association*, 90, 430, 577-588.

[15] ______ and M. West (1998), “Computing Bayesian Nonparametric Hierarchical Models,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds: D. Dey, P. Muller, D. Sinha, New York: Springer-Verlag, 1-22.

[16] Ferguson, T. A. (1973), “A Bayesian Analysis of Some Nonparametric Problems,” *Annals of Statistics*, 1, 209-230.

[17] ______ (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in H. Rizvi, J. Rustagi, and D. Siegmund (eds.) *Recent Advances in Statistics*, New York: Academic Press, 287-303.

[18] Frühwirth-Schnatter, S. (2001), "Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models," 194-209.

[19] Gilks, W. R. and P. Wild (1992), "Adaptive Rejection Sampling for Gibbs sampling",. *Applied Statistics,* 41, 337-348.

[20] Gönül, Füsun and Kannan Srinivasan (1993), "Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models: Methodological and Managerial Issues," *Marketing Science*, 12 (3), 213-229.

[21] Hajivassiliou, V., D. McFadden and P. Ruud (1996), "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives: Theoretical and Computational Results," *Journal of Econometrics*, 72(1-2), 85-134.

[22] McCulloch, R. E. and P.E. Rossi (1994), "Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207-240.

[23] McCulloch, R., N. Polson, and P. Rossi (2000), "Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173-193.

[24] Neal, R. M. (1998), "Markov Chain Sampling Methods for Dirichlet process Mixture Models," *Technical Report,* No. 9815, Department of Statistics, University of Toronto.

[25] ______ (1997), "Markov Chain Monte Carlo Methods Based on 'Slicing' the Density Function," *Technical Report,* No. 9722, Department of Statistics, University of Toronto.

[26] ______ (1991), "Bayesian Mixture Modeling by Monte Carlo Simulation," *Technical Report,* No. CRG-TR-91-2, Department of Computer Science, University of Toronto.

[27] Redner, R. A. and H. F. Walker (1984), "Mixture Densities, Maximum Likelihood, and the EM Algorithm," *SIAM Review*, 26, 195-239.

[28] Richardson, S.. and P. J. Green (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society.* Ser. B, 59(4), 731-792.

[29] Roeder, K.(1994), "A Graphical Technique for Determining the Number of Components in a Mixture of Normals," *Journal of the American Statistical Association*, 89, 426, 487-495.

[30] Sheather, S. J. and M. C. Jones (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society,* Ser. B, 53, 683-690.

[31] Stephens, M. (2000), "Dealing with Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 795-809.

[32] Wedel, Michel, Wagner Kamakura, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scott Neslin, and Carsten Stig Poulsen (1999), "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters*, 10 (3), 219-232.

[33] West, M. (1992), "Modelling with Mixtures," (with discussion), in *Bayesian Statistics 4*, eds. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smit, Oxford, U.K.: Oxford University Press, 503-524.