PSYCHOMETRIKA—VOL. 71, NO. 4, 631–657
DECEMBER 2006
DOI: 10.1007/s11336-006-1233-5

SEMIPARAMETRIC THURSTONIAN MODELS FOR RECURRENT CHOICES: A BAYESIAN ANALYSIS

ASIM ANSARI

COLUMBIA BUSINESS SCHOOL, COLUMBIA UNIVERSITY

RAGHURAM IYENGAR

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

We develop semiparametric Bayesian Thurstonian models for analyzing repeated choice decisions involving multinomial, multivariate binary or multivariate ordinal data. Our modeling framework has multiple components that together yield considerable flexibility in modeling preference utilities, cross-sectional heterogeneity and parameter-driven dynamics. Each component of our model is specified semiparametrically using Dirichlet process (DP) priors. The utility (latent variable) component of our model allows the alternative-specific utility errors to semiparametrically deviate from a normal distribution. This generates a robust alternative to popular Thurstonian specifications that are based on underlying normally distributed latent variables. Our second component focuses on flexibly modeling cross-sectional heterogeneity. The semiparametric specification allows the heterogeneity distribution to mimic either a finite mixture distribution or a continuous distribution such as the normal, whichever is supported by the data. Thus, special features such as multimodality can be readily incorporated without the need to overtly search for the best heterogeneity specification across a series of models. Finally, we allow for parameter-driven dynamics using a semiparametric state-space approach. This specification adds to the literature on robust Kalman filters. The resulting framework is very general and integrates divergent strands of the literatures on flexible choice models, Bayesian nonparametrics and robust time series specifications. Given this generality, we show how several existing Thurstonian models can be obtained as special forms of our model. We describe Markov chain Monte Carlo methods for the inference of model parameters, report results from two simulation studies and apply the model to consumer choice data from a frequently purchased product category. The results from our simulations and application highlight the benefits of using our semiparametric approach.

Key words: semiparametric models, Dirichlet process priors, choice models.

1. Introduction

Decision makers face recurrent choice events in many different contexts. For example, in psychological research on decision-making, subjects in a choice experiment rank order or choose one from many different choice options on multiple trials of the experiment. Similarly, consumers make repeated choice decisions when deciding which brand to buy in frequently purchased product categories. When analyzing such data, researchers are interested in understanding the underlying determinants of preference evolution and change over time, and in capturing sources of individual differences across decision makers. Preferences can evolve over time because of learning, feedback, decision-inertia or because of changes in importance weights for the attributes of the choice alternatives. Preferences can vary cross-sectionally across decision makers because of differences in importances for attributes, differences in characteristics of decision makers, or because of other unobserved sources of heterogeneity that impact the decision process. In this paper, we develop multivariate Thurstonian models that simultaneously incorporate both temporal and cross-sectional sources of variation in preferences in a semiparametric fashion.

Requests for reprints should be sent to Asim Ansari, Columbia Business School, Columbia University, 517 Uris Hall, 3022 Broadway, New York, NY 10027, USA. E-mail: maa48@columbia.edu.

© 2007 The Psychometric Society

Thurstonian models (Thurstone, 1927, 1931) are widely used in the psychometric literature to model multinomial choices, multivariate binary and ordinal choice data, paired comparisons and rank-ordered data (Albert & Chib, 1993, 1997; Böckenholt, 1990, 2001; Bradlow & Zaslavsky, 1999; Chan & Bentler, 1998; Maydeu-Olivares, 1999, 2001; Takane, 1987; Yao & Böckenholt, 1999). In a Thurstonian specification, multivariate normal latent variables are used to model the observed qualitative choices in terms of an underlying discriminal process. This latent variable representation, where the latent variable can also be interpreted as the "utility" of the decision object, facilitates a clear understanding of the impact of various decision factors on the choice process. In this paper, we build Thurstonian models composed of three components, (1) a base component that captures the relationship between alternative-specific attributes and the choices, (2) a temporal component that specifies the dynamic evolution of parameters using a state-space approach and (3) a heterogeneity component which specifies how utility parameters differ across decision makers. We show how all three components can be constructed semiparametrically using Dirichlet process (DP) priors (Ferguson, 1973, 1974; Blackwell & MacQueen, 1973).

Our use of Bayesian semiparametrics to construct each of the three components yields a very flexible model that nests several common Thurstonian specifications. Specifically, we use Dirichlet Process priors to capture the uncertainties regarding the functional forms for the different distributions that arise in our model components. The use of Dirichlet process priors allows us to avoid the restrictions arising from typical parametric distributional assumptions and yields flexible random distributions for the utility errors, individual-level coefficients and the unobserved components that characterize preference dynamics. In each instance, various aspects of these distributions are automatically determined from the data, thus facilitating robust inferences about the determinants of preference and choice.

In our first component, which specifies the utilities for the choice options, we use a Dirichlet mixture to semiparametrically deviate from the typical Thurstonian assumption that the stochastic components of the utilities are distributed multivariate normal. The Dirichlet mixture yields an infinite mixture of normals for representing the variation in the utility errors. For the second component, in modeling heterogeneity across decision makers, we again use the Dirichlet process to yield flexible random effect distributions capable of accommodating multimodality, skewness, and other deviations from multivariate normality that might be appropriate for specific datasets. The use of DP priors for the heterogeneity component facilitates a seamless adjustment of the population distribution—the resulting heterogeneity distribution can mimic either a finite mixture distribution (see Wedel & Kamakura, 1998) or a continuous distribution such as the multivariate normal, (see Longford, 1993; Fox & Glas, 2001) depending upon the information in the data. Finally, in modeling the temporal component, we use a semiparametric state-space approach to model parameter-driven preference dynamics. The use of the DP to model the transition dynamics generalizes previous approaches for robustifying state-space models in the statistical and engineering literatures.

Bayesian semiparametric models have recently become computationally feasible with the development of Markov chain Monte Carlo (MCMC) methods for sampling from the posterior distribution for the Dirichlet process parameters (Escobar, 1994; Escobar & West, 1998; MacEachern, 1994; Ishwaran & James, 2001). In this paper we show how MCMC methods can be implemented for our model and illustrate these methods via a simulation study as well as an application to real data. The simulation focuses on the use of the Dirichlet process for modeling heterogeneity in multivariate binary choice models. Specifically, we compare the performance of a semiparametric model incorporating heterogeneity based on the Dirichlet process to that of a parametric alternative that assumes normally distributed heterogeneity effects. We show that when individual-specific parameters are generated from a discrete population distribution, the semiparametric model is more accurate in recovering the true parameters as

compared to a parametric model with normal heterogeneity. Furthermore, the semiparametric model conveys the multimodality of the population distribution appropriately. In contrast, when individual-specific parameters arise from a normal population distribution, we find that the semiparametric model performs as well as the parametric model with normal heterogeneity in recovering the true individual-level parameters. In the application, we illustrate our methods on data involving recurrent consumer choice decisions for a frequently purchased product category and show how the model results can illuminate the dynamics of preference evolution over time.

The rest of the paper is as follows. Section 2 describes our modeling framework and develops the general semiparametric Thurstonian specification. Section 3 details two simulation studies that investigate different heterogeneity specifications. Section 4 describes the application of our methods on multinomial data involving repeated consumer purchases and Section 5 concludes with a description of our modeling contributions and suggests extensions for future research.

2. Modeling Framework

The Thurstonian framework provides the foundation for modeling many different types of qualitative responses such as paired comparisons, rank orders, ordinal ratings, multivariate binary choices and multinomial choice decisions. For each data type, the Thurstonian approach implies that an underlying variable representation can be used to model the observed qualitative variables in terms of explanatory variables suggested by psychological theory. The nature of the mapping from the latent variables (utilities) to the observed variables depends upon the type of the qualitative dependent variable. However, in each instance, a vector of latent utilities can be used as a starting point for further model development. In this paper, we describe Thurstonian models for multinomial, multivariate and ordinal choice decisions.

2.1. Longitudinal Data

We focus on longitudinal data situations. Consider data from $i = 1$ to I individuals spanning T time periods. The time periods could correspond to trials in an experiment or waves in a panel survey, or may represent calendar time (e.g., weeks or months). We assume that each individual contributes $j = 1$ to n_i observations. For modeling flexibility, we allow these n_i observations to be spread arbitrarily over the T time periods and use an index $t(ij)$, which takes values in the set $\mathcal{T} = \{1, \ldots, T\}$, to represent the time period pertaining to the jth observation for the ith individual. Such indexing establishes a link between observations and time periods and allows us to model both balanced and unbalanced datasets in a flexible fashion.

In a balanced dataset, (e.g., from a choice experiment) each subject is exposed to the same number of trials. Here, the observations are aligned across subjects and the dataset has one observation for each time period (trial). In such situations, there is no distinction between time periods and observations and $t(ij) = j$ and $n_i = T$, for all i. In contrast, many observational data sets are unbalanced because individuals have varying number of observations, and moreover, these observations are not aligned across individuals. In such situations, calendar time provides a natural metric for modeling longitudinal dynamics. As an example, consider consumer choices occurring in real markets. Choices are only observed when consumers buy in a product category. As purchase incidence occurs stochastically over time, observations are not aligned across consumers. If the time periods correspond with weeks or months, then we could have some consumers making multiple purchases in a given time period, whereas, others may not make any purchase at all during the period. In such contexts, $t(ij)$ is distinct from j and indicates the time period specific to the jth observation for the ith individual.

Given the longitudinal data structure, we use three model components to capture different sources of variation in responses. In the first component, we specify the structure of the underlying choice utilities for a given observation. In the second component, we model the heterogeneity across individuals and in the third component, we focus on the parameter-driven dynamics using a state-space approach.

2.2. Choice Utilities

We begin by describing the nature of the mapping between the observed choices and the underlying utilities for three types of choice decisions.

2.2.1. Multinomial Choice. In many discrete choice settings, decision makers choose one alternative from a choice set of mutually exclusive and collectively exhaustive choice alternatives. For instance, on a shopping trip, consumers typically choose one brand from a set of brands within a product category. Similarly, individuals choose a single mode of transportation to travel from one point to another. Consider a choice set consisting of $(K+1)$ such choice alternatives. If $\mathbf{y}_{ij} = \{y_{ij1}, y_{ij2}, \ldots, y_{ijK+1}\}$ is the vector of $(K+1)$ binary indicators that identify the chosen alternative on an observation j and $\mathbf{v}_{ij} = \{v_{ij1}, v_{ij2}, \ldots, v_{ijK+1}\}$ is the latent vector of random utilities underlying the observed choice, then the link between the observed responses and the latent variables can be represented as follows:

$$y_{ijk} = \begin{cases} 1, & \text{if} \quad \max(\mathbf{v}_{ij}) = v_{ijk} \\ 0, & \text{otherwise.} \end{cases}; \qquad k = 1 \text{ to } K+1. \tag{1}$$

The above implies that decision makers choose the alternative that yields the highest utility. Given the nature of the decision process, the identity of the chosen alternative remains invariant if the same constant is added to all discrete choice utilities in $\mathbf{v}_{ij}$ as only utility comparisons across alternatives are relevant. This location invariance of the utilities can be accommodated by considering only differences in utilities from a base alternative (e.g., the last alternative in the choice set). The differencing yields a set of K transformed utilities $\mathbf{u}_{ij} = \{u_{ij1} = v_{ij1} - v_{ijK+1}, u_{ij2} = v_{ij2} - v_{ijK+1}, \ldots, u_{ijK} = v_{ijK} - v_{ijK+1}\}$ which can be used as a basis for further modeling.

2.2.2. Multivariate Binary Choice. In contrast to the multinomial choice situation described above, in certain decision contexts when the decision maker can choose more than one alternative from the choice set, the observed discrete responses represent a multivariate binary choice situation. For instance, a consumer can buy more than one product on a shopping trip to the store and thus the product category purchase decision can be represented in terms of multiple binary choice decisions. If $\mathbf{y}_{ij} = \{y_{ij1}, y_{ij2}, \ldots, y_{ijK}\}$ is a vector of K binary indicators that identifies the chosen alternatives on an observation and $\mathbf{u}_{ij} = \{u_{ij1}, u_{ij2}, \ldots, u_{ijK}\}$ is the latent vector of utilities underlying these choices, then the mapping between the observed responses and the latent variables is given by

$$y_{ijk} = \begin{cases} 1, & \text{if } u_{ijk} > 0 \\ 0, & \text{if } u_{ijk} \leq 0 \end{cases}; \quad k = 1 \text{ to } K. \tag{2}$$

In such multivariate choice situations, the choice options are not mutually exclusive, and as each utility is compared to a threshold (normalized to zero), differencing of utilities is no longer required for model identification.

2.2.3. Multivariate Ordinal Choice. Extending the multivariate choice situation, we can also model decisions wherein decision makers face multiple ordinal choices. Let $\boldsymbol{y}_{ij}$ be $K \times 1$ vector of ordinal responses for the jth observation of individual i. We assume that the ordinal responses are on a Likert scale ranging from 1 to M. The observed responses can be modeled in terms of a $K \times 1$ vector $\mathbf{u}_{ij}$ of underlying continuous variables. The link between the kth ordinal variable y_{ijk} and the corresponding latent variable u_{ijk} can be written as:

$$y_{ijk} = \begin{cases} 0, & \text{if } u_{ijk} < 0 \\ 1, & \text{if } 0 \le u_{ijk} < \theta_{k1} \\ 2 & \text{if } \theta_{k1} \le u_{ijk} < \theta_{k2} \\ \vdots & \\ M, & \text{if } \theta_{kM-1} \le u_{ijk}, \end{cases} \tag{3}$$

where, the vectors $\boldsymbol{\theta}_k = \{\theta_{k1}, \ldots, \theta_{kM-1}\}$, for $k = 1 \ldots K$ contain the item-specific thresholds. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ contain the thresholds across all the K items. For parsimony we model a situation where the thresholds are assumed to be the same across all individuals. However, we assume that each ordinal variable has a different set of the thresholds associated with it. To ensure identification, we set the first threshold θ_{k0} to 0 for all k. These thresholds are assumed to satisfy the order constraints $0 \le \theta_{k1} \le \theta_{k2} \le \ldots \le \theta_{kM-1}$ for each k. The constraints ensure that the cumulative probabilities for the ordinal responses are non-decreasing.

Irrespective of the type of choice decision, the latent utilities $\mathbf{u}_{ij}$ serve as a starting point for specifying the impact of explanatory variables. Each latent variable u_{ijk} can be expressed in terms of independent variables that include the attributes of the decision options and the characteristics of the decision maker. Let $\mathbf{x}_{ijk}$ be a $p_k \times 1$ vector that contains all independent variables that influence the kth response on the jth observation for individual i. In addition, let $\mathbf{w}_{ijk}$ be a $l_k \times 1$ vector of only those explanatory variables in $\mathbf{x}_{ijk}$ whose impact is assumed to vary across individuals. Notice that $\mathbf{w}_{ijk}$ can contain all the variables in $\mathbf{x}_{ijk}$, in which case, all coefficients are modeled as varying across individuals. In addition, the first elements in $\mathbf{x}_{ijk}$ and $\mathbf{w}_{ijk}$ are set to 1 to incorporate response-specific intercepts. The kth utility equation can be written as

$$u_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\mu}_{t(ij)k} + \mathbf{w}'_{ijk}\boldsymbol{\lambda}_{ik} + e_{ijk}, \quad k = 1 \ldots K, \tag{4}$$

where $\boldsymbol{\mu}_{t(ij)k}$ is a $p_k \times 1$ vector of time-varying population-level response coefficients, the $l_k \times 1$ vector $\boldsymbol{\lambda}_{ik}$ contains the individual-specific coefficients and e_{ijk} represents the unobserved influences that impact the kth utility.

Consolidating across the K utilities, the system of equations that characterizes the relationship between the latent variables and the covariates can be written as:

$$\mathbf{u}_{ij} = \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i + \mathbf{e}_{ij}, \tag{5}$$

where, $\mathbf{X}_{ij}$ is a $K \times L_1$ matrix and $\mathbf{W}_{ij}$ is a $K \times L_2$ matrix. These are composed appropriately from the vectors $\mathbf{x}_{ijk}$ and $\mathbf{w}_{ijk}$ respectively. The exact specification for these matrices, and therefore, the magnitudes of L_1 and L_2 depend upon how the explanatory variables and their coefficients are shared across the K utility equations. The vector $\boldsymbol{\mu}_{t(ij)}$ contains all the time-varying population-level response coefficients across all equations and the vector $\boldsymbol{\lambda}_i$ contains all the individual-specific random coefficients. The vector of errors $\mathbf{e}_{ij}$ is assumed to vary independently across the observations in the dataset.

Thurstonian models assume that the errors $\mathbf{e}_{ij}$ are distributed multivariate normal $N(0, \boldsymbol{\Sigma})$. In this paper, we deviate from the normality assumption and instead, use a Dirichlet mixture distribution to characterize the uncertainty about the distribution of the utility errors. We assume that the unknown continuous distribution for the errors F_1 belongs to the family of scale mixtures

of the multivariate normal distribution. Let $\Phi_{\gamma_{ij}}$ represent the CDF of the normal, $N(0, \gamma_{ij}^{-1}\boldsymbol{\Sigma})$, where γ_{ij} is a positive scale variable, and let

$$F_1(e) = \int_{\gamma_{ij}} \Phi_{\gamma_{ij}}(e) \mathrm{d}G_\gamma(\gamma_{ij}). \tag{6}$$

The scale variable γ_{ij} is assumed to come from a unknown distribution G_γ defined over $\mathcal{R}^+$. The class of normal scale mixture distributions allows considerable flexibility in functional forms and tail behaviors and offers a robust alternative to the typical choice of a multivariate normal. We model the uncertainty about the unknown mixing distribution G_γ using a Dirichlet process prior to yield a Dirichlet mixture distribution over F_1 (Basu & Mukhopadhyay, 2000; Chib & Hamilton, 2002).

The Dirichlet process provides a mechanism for placing a probability distribution on the space of distributions and can be described as follows: Let H be a random distribution function. Given a positive real α and a distribution H_0 defined on a space χ, we say that H comes from the Dirichlet process, i.e., $H \sim DP(H_0, \alpha)$, if, for all Borel sets $A_1, A_2, \ldots, A_M$ that partition χ and for all values of M, the vector of random probabilities:

$$\boldsymbol{p} = (p_1, \ldots, p_M) = (p(A_1), \ldots, p(A_M)) = \left(\int_{A_1} dH(x), \ldots, \int_{A_M} dH(x) \right), \tag{7}$$

where $\sum_{i=1}^{M} p_i = 1$, has the Dirichlet distribution

$$\boldsymbol{p} \sim \mathrm{Dir}(\alpha H_{01}, \alpha H_{02}, \ldots, \alpha H_{0M}), \tag{8}$$

and $H_{0l} = \int_{A_l} dH_0(x)$. The Dirichlet distribution has a density given by the expression,

$$\frac{\Gamma(\alpha)}{\prod_{l=1}^{M} \Gamma(\alpha H_{0l})} \prod_{l=1}^{M} p_l^{\alpha H_{0l} - 1}. \tag{9}$$

From the moments of the Dirichlet distribution, we can see that if $X \sim H$, then the expected probability of X falling into a region A_l is $\int_{A_l} dH_0(x)$, i.e., $\mathrm{E}[H(A_l)] = H_0(A_l)$. Similarly $\mathrm{Var}[H(A_l)] = H_0(A_l)(1 - H_0(A_l))/(\alpha + 1)$. When the precision parameter, α of the Dirichlet process is close to zero, the variance $\mathrm{Var}[H(A_l)]$ is very high. As α gets larger, this variance becomes smaller and the distributions sampled from H become closer to the base distribution H_0. Realizations of a Dirichlet process are discrete probability distributions, i.e., with probability one, H is discrete, but the number of points to which it gives positive probability is unbounded.

In modeling the utility errors, we assume that $\gamma_{ij} \sim G_\gamma$, and as G_γ is random (i.e., unknown) we use a Dirichlet process prior $G_\gamma \sim DP(G_{\gamma 0}, \alpha_\gamma)$, to characterize the uncertainty regarding G_γ. Here, $G_{\gamma 0}$ is the baseline distribution that defines the "location" of the Dirichlet process prior, and the precision parameter $\alpha_\gamma > 0$ determines the concentration of the prior for G_γ about the baseline distribution $G_{\gamma 0}$. We use a gamma distribution $\mathrm{Ga}(\nu_\gamma/2, \nu_\gamma/2)$ as the baseline distribution. A Dirichlet process distribution for G_γ generates a semiparametric Dirichlet mixture distribution for F_1 via the convolution in Equation 6. Given the baseline gamma distribution for the Dirichlet process, the prior expectation of F_1 is a multivariate-t distribution. This is the case as the multivariate-t can be obtained as a scale mixture of a multivariate normal and the chi-squared distribution.

The precision parameter α_γ determines how close the semiparametric distribution G_γ is to the baseline gamma distribution $G_{\gamma 0}$. When α_γ is large ($\alpha_\gamma \to \infty$), a randomly sampled mixing distribution, G_γ, approximates the baseline gamma distribution. Then, the resulting unconditional distribution of the utility errors is similar to a multivariate-t distribution with degrees of freedom ν_γ. Further, if ν_γ is large, the distribution of errors mimics a normal and thus encompasses the typical Thurstonian specification. On the other hand, when α_γ is small

($\alpha_\gamma \to 0$), the sampled mixing distribution has its mass concentrated on a few points and then, the unconditional distribution for the utility errors behaves as a finite mixture of normal distributions, each with the same mean. As the precision parameter α_γ, and the normal covariance matrix $\mathbf{\Sigma}$ are inferred from the data, the Dirichlet process automatically allows us to flexibly model the distribution for the utility errors. For identification, different restrictions are required on $\mathbf{\Sigma}$ for the three types of choice models. For multinomial choice, one of the variances in $\mathbf{\Sigma}$ is constrained to one and for the multivariate binary and ordinal choice situations, $\mathbf{\Sigma}$ is constrained to be a correlation matrix. Having specified the utility component, we now describe the dynamic component of our model.

2.3. Parameter-Driven Dynamics

We use a state-space approach to model the parameter-driven dynamics (Böckenholt, 2002). Specifically, we assume that the vector of population response coefficients $\boldsymbol{\mu}_{t(ij)}$ stochastically evolves over time according to a transition equation. Recall that the index $t(ij)$ represents the time period pertaining to the jth observation for individual i and takes values in the set $\mathcal{T} = \{1, \ldots, T\}$. The transition equation is written as

$$\boldsymbol{\mu}_{t(ij)} = \mathbf{B}_t \boldsymbol{\mu}_{t(ij)-1} + \boldsymbol{\xi}_{t(ij)}, \tag{10}$$

where $\boldsymbol{\mu}_{t(ij)}$ represents the population coefficients for time period $t(ij)$, $\mathbf{B}_t$ is a known transition matrix and $\boldsymbol{\xi}_{t(ij)}$ is a vector of stochastic variables that govern the uncertainty in the state transition process. The transition equation specifies a Markov chain that allows the population coefficients in a given time period to depend upon those from the previous time period. We assume that the transition process starts with a vector $\boldsymbol{\mu}_0$ that has a normal prior with mean $\mathbf{m}_0$ and a covariance matrix $\mathbf{C}_0$.

As with the utility errors, we assume that the transition errors $\boldsymbol{\xi}_{t(ij)}$ come from an unknown distribution F_2 which is a semiparametric scale mixture of the multivariate normal distribution. Let $\kappa_{t(ij)}$ be a positive scale variable distributed according to an unknown distribution G_κ and let F_2 be a convolution of the normal distribution $N(0, \kappa_{t(ij)}^{-1}\mathbf{\Omega})$ and G_κ. We model the uncertainty regarding G_κ and thus consequently, about F_2, using a Dirichlet process prior, i.e., $G_\kappa \sim DP(G_{\kappa 0}, \alpha_\kappa)$. To complete the specification, we assume that the baseline distribution $G_{\kappa 0}$ of the Dirichlet process is given by a gamma distribution, $\text{Ga}(\nu_\kappa/2, \nu_\kappa/2)$. This yields a general stochastic formulation for the transition equation which adds to the research on robust state-space models that has advocated flexible distributions for modeling the transition errors. For example, Masreliez and Martin (1977) and Meinhold and Singpurwalla (1989) propose the use of multivariate-t distributions for robustification of Kalman filter models. In our model, when α_κ is large, the distribution for the transition errors generated from the Dirichlet mixture mimics a multivariate-t. Other researchers (Carter & Kohn, 1994) have used a finite mixture of normals for modeling the transition errors. When α_κ is small, our model mimics this situation. As α_κ is inferred from the data, a range of possibilities can be flexibly handled within our framework. Finally, our treatment of dynamics complements other research in marketing on the use of Kalman filters (Van Heerde, Mela, & Manchanda, 2004; Leichty, Fong, & DeSarbo, 2005). We next specify the final component of our modeling approach.

2.4. Heterogeneity

Given the longitudinal nature of data, it is crucial to account for individual differences. We assume that the random coefficients $\boldsymbol{\lambda}_i$ vary across the individuals according to a population distribution based on the Dirichlet process. Two semiparametric approaches are possible to model the uncertainty about the population distribution. The first approach involves a Dirichlet mixture based on a scale mixture of normals, as for the utilities and time-varying components. In this

approach, the $\boldsymbol{\lambda}_i$ are assumed to come from $N(0, \tau_i^{-1}\boldsymbol{\Lambda})$, where $\tau_i \sim G_\tau$ and $G_\tau \sim DP(G_{\tau 0}, \alpha_\tau)$. As before, one can use a gamma distribution $\text{Ga}(\nu_\tau/2, \nu_\tau/2)$ as the baseline distribution $G_{\tau 0}$. This Dirichlet mixture yields a continuous unconditional population distribution that provides greater flexibility than the normal distribution that is typically used in multilevel models.

Alternatively, a Mixture of Dirichlet Process (MDP) approach can be used to specify a discrete heterogeneity distribution. The heterogeneity distribution is assumed to come from a mixture of Dirichlet processes. In this specification, the random coefficients $\boldsymbol{\lambda}_i$ come from an unknown discrete population distribution G_λ and the uncertainty about G_λ is specified using a Dirichlet process $DP(G_{\lambda 0}, \alpha_\lambda)$. We assume that the baseline distribution $G_{\lambda 0}$ is a multivariate normal $N(0, \boldsymbol{\Lambda})$. As $\boldsymbol{\Lambda}$ is uncertain, the distribution for G_λ is a called a mixture of Dirichlet processes. Distributions generated in this manner can accommodate deviations from normality such as multimodality and can down-weight the effect of outliers. Depending upon the inferred magnitude of the precision parameter α_λ, the population distribution can mimic a normal (the baseline distribution), or a finite mixture with few mass points.

The choice between the two alternative heterogeneity specifications is governed by prior expectations regarding the nature of heterogeneity. If one expects that the population distribution is better handled using a continuous specification, or is fatter tailed or over- dispersed with respect to the multivariate normal distribution, then the first approach involving scale mixtures of normals is more suitable. However, if one expects that the population distribution could be under-dispersed relative to the multivariate normal, then using the mixture of Dirichlet processes approach is desirable. The latter would be the case when a latent class or finite mixture specification is expected. In such instances, the MDP approach is suitable as it automatically provides information about the number of mass-points, or "segments" or "clusters" that characterize heterogeneity.

We summarize below the complete model. In the following, the utility and transition components are specified using Dirichlet mixtures and the heterogeneity component is specified using the MDP approach described above. The complete specification is as follows:

$$
\begin{aligned}
&\mathbf{u}_{ij} = \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i + \mathbf{e}_{ij},\\
&\mathbf{e}_{ij} \sim N(0, \gamma_{ij}^{-1}\boldsymbol{\Sigma}),\\
&\gamma_{ij} \sim G_\gamma,\\
&G_\gamma \sim DP(\text{Ga}(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma}{2}), \alpha_\gamma),\\
&\boldsymbol{\mu}_{t(ij)} = \mathbf{B}_t\boldsymbol{\mu}_{t(ij)-1} + \boldsymbol{\xi}_{t(ij)},\\
&\boldsymbol{\xi}_{t(ij)} \sim N(0, \kappa_{t(ij)}^{-1}\boldsymbol{\Omega}),\\
&\kappa_{t(ij)} \sim G_\kappa\\
&G_\kappa \sim DP(\text{Ga}(\frac{\nu_\kappa}{2}, \frac{\nu_\kappa}{2}), \alpha_\kappa)\\
&\boldsymbol{\lambda}_i \sim G_\lambda,\\
&G_\lambda \sim DP(N(0, \boldsymbol{\Lambda}), \alpha_\lambda).
\end{aligned}
$$

The model as specified above is very general and nests many different Thurstonian specifications as special cases. Special cases include parametric multinomial (McCulloch & Rossi, 1994) and multivariate probit models (Chib & Greenberg, 1998) with and without normally distributed time-varying parameters and normally distributed heterogeneity. As discussed earlier, the model is also capable of accommodating multimodal error distributions and robust specifications (such as those based on the multivariate-t distribution) that are not susceptible to the influence of outliers.

Inference regarding the different parameters can be based on the joint posterior distribution of the unknowns. As the posterior distribution is not analytically available, we use simulation based Bayesian methods to obtain random draws from this posterior. The complexity of the posterior distribution precludes the use of direct methods for simulating from the posterior. We therefore use Markov Chain Monte Carlo (MCMC) methods to simulate dependent draws from the posterior. The MCMC scheme includes a combination of data augmentation, Gibbs sampling, Metropolis-Hastings steps and slice sampling. The priors and the full conditional distributions are specified in the Appendix.

3. Simulation

We describe two small-scale simulation studies that assess the suitability of modeling heterogeneity using the Dirichlet Process. Specifically, we focus on how well the individual-specific coefficients are recovered by different model formulations. To isolate the role of heterogeneity, we use a multivariate probit model (MVP) that is a restricted version of our full model described earlier. In this model, we allow only cross-sectional variation across individuals and do not accommodate any parameter-driven dynamics. Thus, we use the following model specification for individual i and observation j:

$$\mathbf{u}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}_i + \mathbf{e}_{ij}, \quad \mathbf{e}_{ij} \sim N(0, \boldsymbol{\Sigma}). \tag{11}$$

We use two mechanisms for generating the individual-specific coefficients ($\boldsymbol{\beta}_i$'s). In the first simulation study, these coefficients are generated from a discrete distribution. In the second, these come from a continuous distribution.

We use 25 replications in each simulation study. The replicated datasets contain bivariate binary responses for 200 individuals, each having 10 observations generated according to the MVP model in Equation 11. We assume that each of the underlying utilities are influenced by one explanatory variable. For any given replication, we use the same individual-level design matrix $\mathbf{X}$ of dimension (20×4) for all individuals. The first two columns of $\mathbf{X}$ contain the intercepts for the two utility equations, and the next two columns contain i.i.d. uniform random deviates for the two explanatory variables. The design matrices differ across the replicated datasets. Finally, we set $\boldsymbol{\Sigma} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and hence, the vector of latent utilities for individual i follows $\mathbf{u}_i \sim N(\mathbf{X}\boldsymbol{\beta}_i, \mathbf{I})$.

In each simulation study, we estimate two models. The first model (M1) is a Mixture of Dirichlet Process (MDP) model where the $\boldsymbol{\beta}_i$'s follow an unknown heterogeneity distribution G, where, $G \sim DP(N(\boldsymbol{\mu}, \boldsymbol{\Lambda}), \alpha)$. The second model (M2) uses a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ for capturing heterogeneity. For each replicated dataset we estimate the two models using MCMC methods. The full conditional distributions for Model M2 can be obtained with relative ease as special cases of the conditionals for Model M1. For each model we set diffuse but proper priors. For Model M1, we set a multivariate normal prior $N(\mathbf{0}, 1000\mathbf{I})$ for the population mean $\boldsymbol{\mu}$ of the baseline normal distribution and a Wishart prior $W(5, \mathbf{I}/5)$ for the precision matrix $\boldsymbol{\Lambda}^{-1}$. We used a gamma prior Ga(0.01, 100) for the precision parameter α of the Dirichlet process. For Model M2, the priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}^{-1}$ are the same as for M1. For each model and each dataset, we ran the MCMC simulation for 2000 iterations and the results reported in the paper are based on the last 1000 iterations with the first 1000 iterations used as burn-in draws. In all cases, convergence was clearly achieved well before 1000 iterations.

3.1. First Simulation: Bernoulli Population Distribution

In this study, we construct the datasets by drawing the individual-level $\boldsymbol{\beta}_i$'s using independent symmetric Bernoulli deviates for each of the four coefficients. The symmetric Bernoulli

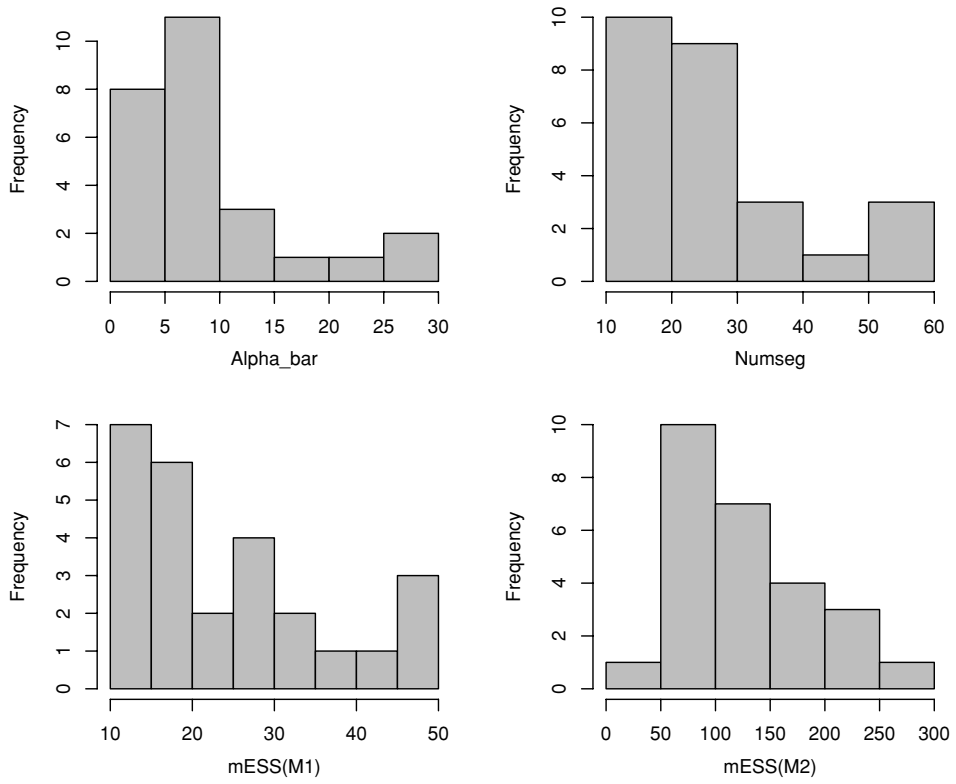

FIGURE 1.
(Study 1) Top Panel - Histograms for $\bar{\alpha}$ (Alpha_bar) and mean number of clusters (Numseg) for Model M1 (MDP model). Bottom Panel - Histograms for mESS of Model M1 (mESS(M1)) and Model M2 (mESS(M2)).

distribution (see Escobar, 1995) is a two atom distribution with mass of 0.5 at a and $-a$. We set the mass points for the first intercept at 2.5 and -2.5 and for the second intercept at 2 and -2. The mass points for the coefficients of the first explanatory variable are set at 1.5 and -1.5 whereas, the mass points for the coefficient of the remaining variable are set at 1.75 and -1.75.

For each dataset, the MCMC iterations for Model M1 (the MDP model) can be used to compute a mean number of "clusters" or "segments" (Numseg) and the posterior mean $\bar{\alpha}$ for the precision parameter α. Recall that the Dirichlet process generates discrete heterogeneity distributions which induces a clustering among the individual-level coefficients. In contrast, the normal heterogeneity assumption in Model M2 implies no such clustering.

The top panel of Figure 1 shows histograms for Numseg and $\bar{\alpha}$ across the 25 replications. The mean for Numseg across these replications is 27.04 which clearly shows that the MDP model captures the discrete nature of the population distribution. The small values of $\bar{\alpha}$ also validate that the individual-specific parameters come from a discrete distribution.

The comparative accuracy of the two models in recovering the individual-level parameters can be assessed using mean error sum of squares (mESS). The mESS for each dataset is calculated using the formula

$$\text{mESS} = \frac{1}{200}\sum_{i=1}^{200}\sum_{l=1}^{4}(\beta_{il} - \hat{\beta}_{il})^2, \tag{12}$$

where, β_{il} is the true parameter and $\hat{\beta}_{il}$ is the estimated posterior mean for the parameter.

The bottom panel for Figure 1 contains histograms for mESS obtained from the 25 replications. The left histogram in the bottom panel is for Model M1 and the right is for Model M2. The plots indicate that the mESS statistics for Model M1 are substantially lower than those obtained for Model M2. In fact, we find that in each of the replications, the mESS for Model M1 is lower than the mESS for Model M2. This clearly demonstrates that when the true heterogeneity distribution is discrete, Model M1 (the MDP model) is more accurate than Model M2 (the normal heterogeneity model) in recovering the true individual-level parameters.

We also compare the marginal predictive distributions for the individual-specific coefficients across the two models. These represent the predictive distributions for a future individual from the population. Consistent with prior expectations, we find that Model M2 generates unimodal predictive distributions for all coefficients. The predictive distributions for Model M1, however, are clearly bimodal with the modes centered around the mass-points used for the symmetric Bernoulli.

Figure 2 shows the marginal predictive distributions for two such individual-specific coefficients (an intercept and slope) for an arbitrary chosen replication. The left column in Figure 2 contains the distributions for M1 (the MDP model). The right column contains the corresponding predictive distributions for M2, the normal heterogeneity model. The figure unambiguously shows that the MDP model is capable of uncovering multimodal population distributions.

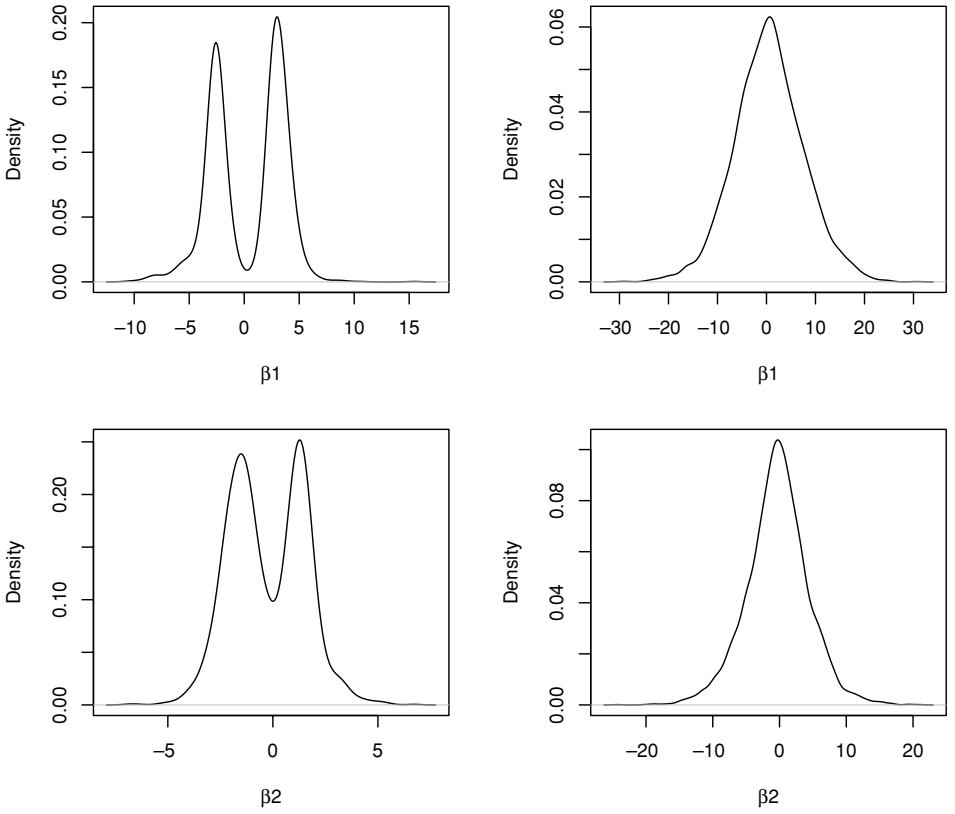

FIGURE 2.
(Study 1) Marginal predictive distributions for the individual-specific coefficients from the MDP model (left column) and the Normal model (right column).

3.2. Second Simulation: Normal Population Distribution

In this study, the individual-level $\boldsymbol{\beta}_i$'s are generated from independent standard normal deviates for each of the four coefficients. We set the population means for the first and second intercepts at 1.0 and 2.0 respectively. The population mean for the coefficients for the two explanatory variables are set at -0.6.

Our interest here is in understanding how well the discrete population distribution G in the MDP model mimics the true normal distribution and to what extent the estimates for the individual-level parameters for this model deviate from the truth because of this discreteness. Even though G is discrete, the number of points to which it gives positive probability is unbounded, and therefore we expect the MDP model to semiparametrically approach the truth.

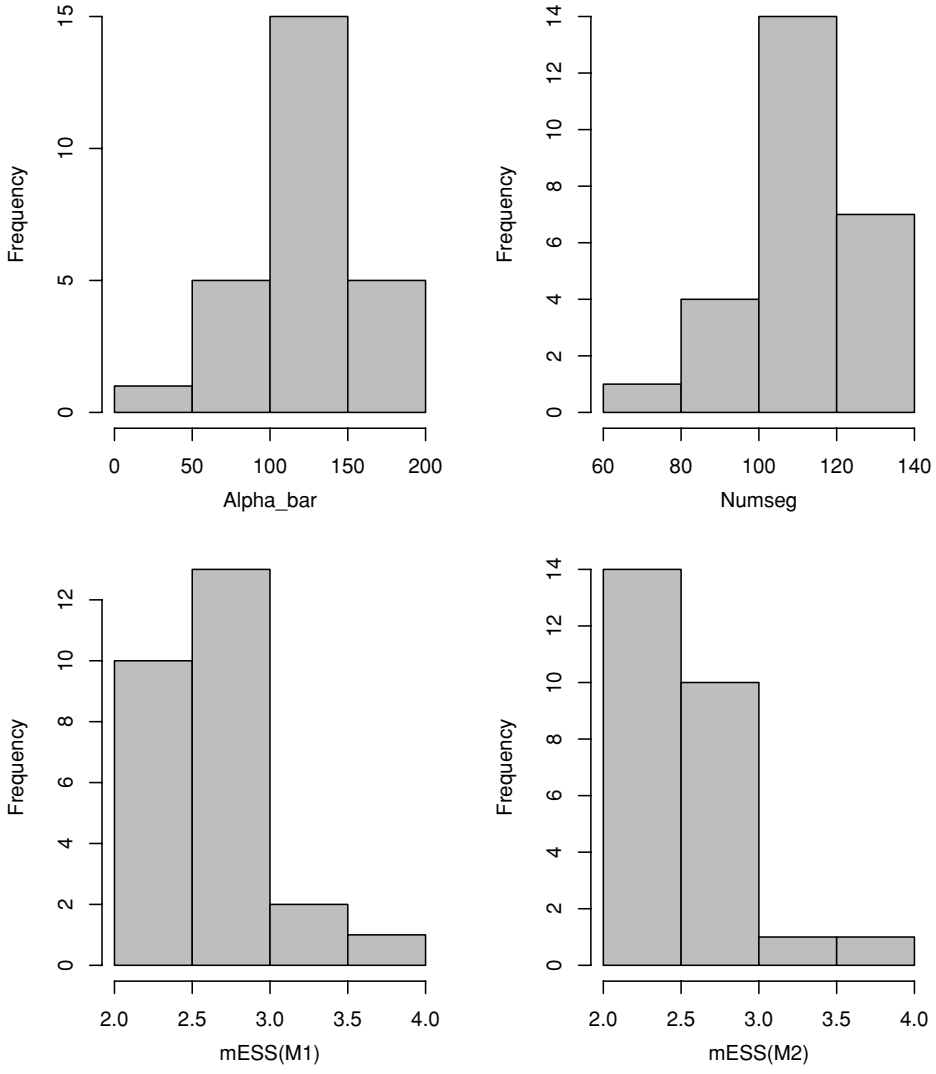

FIGURE 3.
(Study 2) Top Panel - Histograms for $\bar{\alpha}$ (Alpha_bar) and mean number of clusters (Numseg) for Model M1 (MDP model). Bottom Panel - Histograms for mESS of Model M1 (mESS(M1)) and Model M2 (mESS(M2)).

The top panel of Figure 3 shows the histograms for Numseg and $\bar{\alpha}$ across the 25 replications. The mean of Numseg across the replications is 110.3 which clearly shows that the MDP model adjusts to capture the continuous nature of the parameter distribution. The large values of $\bar{\alpha}$ also validate that the base distribution, a normal, is suitable as the population distribution of the parameters.

The bottom panel of Figure 3 shows the histograms of the mean error sum of squares (mESS) across the replications. The left histogram shows the mESS for Model M1 and the right histogram pertains to Model M2. These plots indicate that the mESS statistics for Model M1 are comparable in magnitude to those obtained for Model M2. This clearly demonstrates that the MDP model is as good as a model with normal population distribution in recovering the individual-level parameters even when the true population distribution is a normal.

Figure 4 shows the marginal predictive distributions for two individual-specific coefficients for an arbitrarily chosen replication. The left column in Figure 4 contains the distributions for the Model M1 (MDP model) and the right column contains the corresponding predictive distributions for Model M2. Consistent with prior expectations, Model M2 generates unimodal predictive distributions for all four coefficients. What is noteworthy is that the predictive distributions for Model M1 also indicate unimodality. This shows that the MDP model is flexible in mimicking the true nature of heterogeneity. When the parameters come from a continuous population distribution such as the normal, the Dirichlet process yields discrete distributions that mimic the normal. This is possible because (a) the precision parameter α for the Dirichlet process is inferred from the data and (b) the baseline distribution is a normal.

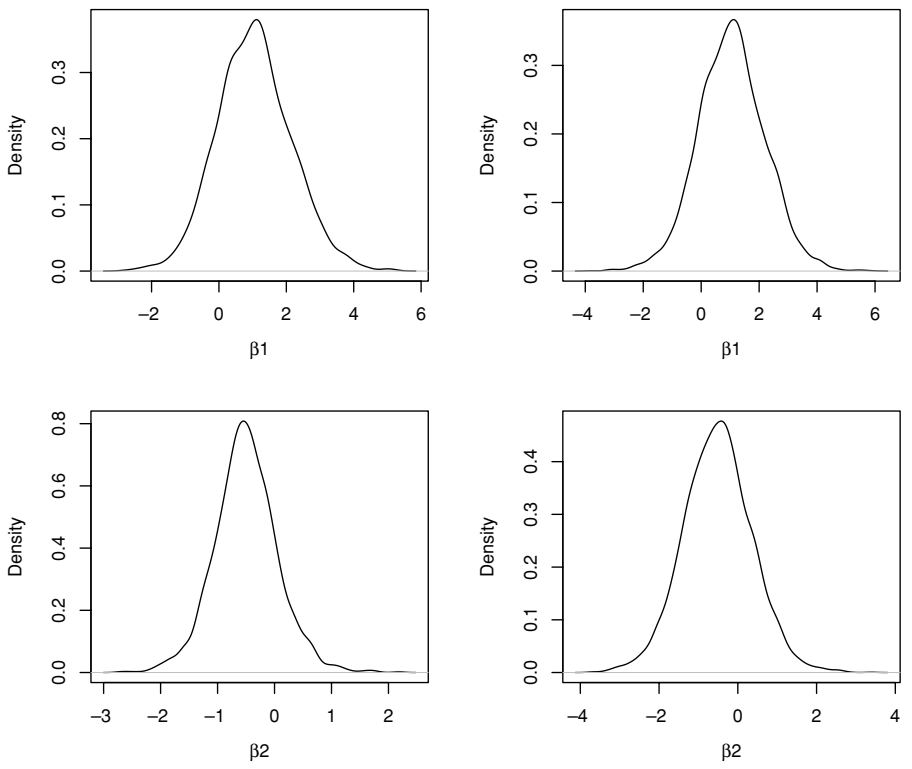

FIGURE 4.
(Study 2) Marginal predictive distributions for the individual-specific coefficients from the MDP model (left column) and the Normal model (right column).

3.3. General Conclusions

In this section, we draw a few general conclusions based on above described simulation studies. The first simulation shows that when the individual-level parameters come from a discrete population distribution, the MDP model adjusts the precision parameter α to indicate the discrete character of the true distribution and highlights multimodality. In contrast, the model with a normal population distribution assumes that each individual is a distinct cluster of its own and the predictive distributions for the individual-level coefficients provide no hint of multimodality. Further, a comparison of the mESS shows that the MDP model recovers the true parameters better than the model with normal population distribution.

The second simulation shows that the Model M1 (MDP model) performs as well as Model M2 (the normal heterogeneity model) when the parameters indeed come from a normal distribution. Thus, there is no loss of accuracy in using an MDP model when the true parameter distribution is a normal. This is further validated from Figure 3 which shows that the mESS from the MDP model are comparable to the mESS from a normal model.

A comparison of Figures 1 and 3 reveals further insights about the nature of the MDP model. This comparison shows that when the underlying distribution is discrete, the MDP model adjusts to give a low value for $\bar{\alpha}$, whereas, when the underlying distribution is continuous, the MDP model yields a high value for $\bar{\alpha}$. Thus, the model is flexible enough to reflect the changes in the data generation process. The histogram for the mean number of clusters (Numseg) also reveals a similar pattern. Thus, overall, the MDP model can appropriately capture parameter heterogeneity as it is clearly better than a model with normal heterogeneity when the true population distribution is discrete and not worse off when the true distribution is a normal.

4. Application

In this section, we use our semiparametric approximation to model multinomial brand choice of households for the canned tuna product category.

4.1. Data and Variables

We use purchase data collected from supermarket scanners. The dataset comes from a panel of households maintained by A.C. Neilsen in a mid-western market in the U.S. and contains choices for four brands of canned tuna over a duration of 29 months beginning from 1985. We have data on the purchase histories of 300 households. These households made a total of 4213 purchases during the data period. The mean interpurchase time is about six weeks, the mean number of observations per household is 14 with a standard deviation of 12 and the average number of observations per month is 145 with a standard deviation of 48. In addition to the information on household choices of the four brands, the data has details of prices, supermarket displays and in-store feature advertisements for the four brands within the stores in the market.

In modeling the choices, we use the following explanatory variables to specify the utility equations for the brands. The price variable (PRICE) represents the price per ounce of a brand. The display variable (DISPLAY) is a binary indicator of whether a brand was on special display in the store on the shopping occasion. Similarly, feature (FEATURE) is a dummy variable that captures whether a brand was featured in the store circular by the retailer. Figure 5 shows how the average market shares, prices, displays and features for the four brands vary over the time-span of the data. The figure shows that, except for short-term fluctuations, the market shares for the four brands were relatively stable over the long run. Also, the feature and display activity remained relatively stable over the time period of the data. In contrast, there is a decrease in brand prices, especially in the last 15 months of the data.

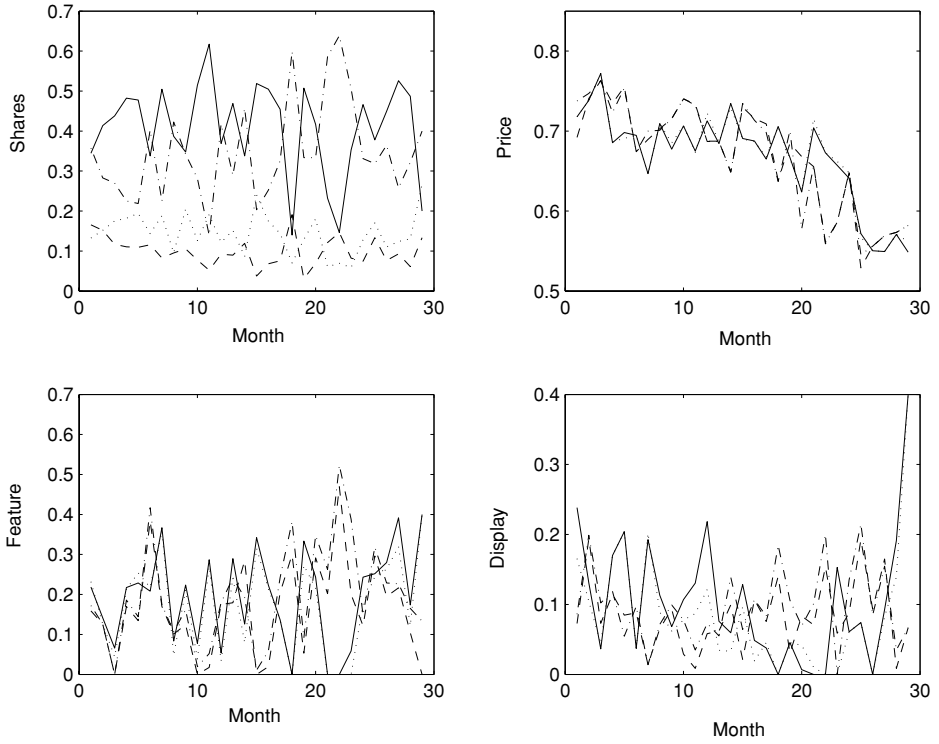

FIGURE 5.
Time variation in Brand Shares, Prices and Promotions for the four brands over the span of the data. Brand 1 (solid line), Brand 2 (dotted line), Brand 3 (small dash-dot line) and Brand 4 (long dash-dot line).

4.2. Model Specifications

To understand the effects of including cross-sectional heterogeneity and state-space dynamics, we estimate three models on the data. The first model, Model 1, is a multinomial probit model that incorporates both cross-sectional heterogeneity and time-varying dynamics. The heterogeneity specification is built as in Section 2.4 to accommodate differences in preference primitives across households by allowing the household-specific parameters, $\boldsymbol{\lambda}_i$, to come from a random distribution, G_λ. The uncertainty about G_λ is modeled by using a DP prior $DP(N(0, \boldsymbol{\Lambda}), \alpha_\lambda)$. The time-varying component is specified as in Section 2.3 to capture parameter-driven monthly dynamics. For simplicity, we assume a random walk transition equation, thus the matrix $\boldsymbol{B}_t$ is set to identity. We use scale mixtures of normals to specify the errors in the transition equation. The mixing distribution G_κ for the scale parameter, κ_t, is modeled semiparametrically using a DP prior $DP(\text{Ga}(\nu_\kappa/2, \nu_\kappa/2), \alpha_\kappa)$. We allow the precision parameters, α_κ and α_λ, for the Dirichlet processes to be inferred from the data.

The second and third models are restrictions of the above model. The second model (Model 2) captures only cross-sectional heterogeneity and does not contain any time-varying parameters. Thus, it is a mixture of Dirichlet process multinomial probit model. The third model (Model 3) incorporates only state-space dynamics but does not accommodate any unobserved sources of heterogeneity across households. Thus, Model 3 is a semiparametric state-space multinomial probit model.

As we have four brands, we obtain three differenced utilities (differenced with respect to a base alternative-Brand 4). In each of the three models we specify these utilities in terms of

six regressors: three differenced marketing variables - price, display and feature and three utility specific intercepts to capture the intrinsic preferences of each brand relative to the base alternative. As the price, feature and display sensitivities are intrinsic to the households, these coefficients are invariant across the three utility equations. This yields a design matrix $\mathbf{X}_{ij}$ of dimension 3×6 for observation j of household i. As all coefficients are assumed to vary across households, the matrix $\mathbf{W}_{ij}$ is the same as $\mathbf{X}_{ij}$.

4.3. Results

In this section, we present the results for the three models. The results are based on MCMC draws from the joint posterior distribution of all unknowns. For each model, we ran sampling chains for 15,000 iterations and convergence was assessed by monitoring the time-series of the draws. The results in this paper are based on 5,000 draws retained after discarding the initial 10,000 burn-in iterations. We begin by reporting model comparison results.

4.3.1. Model Comparison. We base our model comparisons on the Bayes factors (Kass & Raftery, 1995), which account for model fit and simultaneously penalize model complexity. The Bayes factors for two models can be expressed as the ratio of their marginal densities. We computed the Log-marginal likelihood (LML) for each model using the MCMC draws using the GHK method for calculating the CDF of a multivariate normal distribution (Geweke, 1989; Hajivassiliou, 1990; Keane, 1994).

The LML for Model 1 is 6490.17, for Model 2 is 6668.27 and for Model 3 is 8172.71. A comparison of the LML's across the three models provides information about the importance of modeling unobserved cross-sectional heterogeneity and parameter-driven dynamics in our data. Model 1, which includes both dynamics and heterogeneity has the highest LML. We find that Model 3, which does not include household-level unobserved heterogeneity performs the worst. These results show that including unobserved household heterogeneity leads to a greater improvement in the LML when compared to the improvement resulting from the inclusion of state-space dynamics. It is important to note that these results are specific to our dataset. Clearly, it is possible that in other contexts, the time varying component may be more important.

4.3.2. Model Adequacy. The adequacy of a Bayesian model can be assessed using posterior predictive model checking (Gelman, Meng, & Stern, 1996; Hoijtink & Molenaar, 1997; Yao & Bockenholt, 1999). In this approach, the actual data set, D^{obs}, is compared with many replicated data sets $D_1^{\text{rep}}, \ldots, D_l^{\text{rep}}$ that are generated from the model. The datasets can be compared using either test statistics $T(D)$ that are functions of data alone or using more general discrepancy variables $T(D, \gamma)$ which are quantities involving both the data and parameters. These test quantities are chosen to measure departures of the observed data from the assumed model. If the replicated data sets differ systematically from the actual data on some test quantities, then we can ascertain that the model does not adequately capture the data generation process on those aspects that are captured by the test quantities.

A posterior predictive p-value can be approximated from a sample of l MCMC draws using

$$p(D) = \frac{1}{l} \sum_{i=1}^{l} I(T(D_i^{\text{rep}}, \gamma_i) \geq T(D^{\text{obs}}, \gamma_i)), \tag{13}$$

where $I()$ is an indicator function. The expression in (13) calculates the p-value as the proportion of the l replications in which the simulated discrepancy variable exceeds the realized value. An extreme p-value (i.e, close to either zero or one) indicates that the model is inadequate in recovering that aspect of the data that is measured by the test quantity T.

TABLE 1.
Posterior predictive p-values.

Month	Brand1	Brand2	Brand3	Brand4
1	0.673	0.493	0.493	0.443
2	0.490	0.497	0.490	0.567
3	0.357	0.397	0.597	0.563
4	0.553	0.493	0.457	0.477
5	0.533	0.477	0.483	0.577
6	0.563	0.413	0.350	0.693
7	0.473	0.323	0.573	0.680
8	0.480	0.607	0.467	0.643
9	0.607	0.223	0.603	0.583
10	0.400	0.797	0.340	0.530
11	0.640	0.427	0.577	0.253
12	0.670	0.597	0.267	0.633
13	0.503	0.437	0.667	0.613
14	0.363	0.723	0.443	0.537
15	0.640	0.383	0.440	0.573
16	0.437	0.417	0.437	0.770
17	0.500	0.483	0.513	0.490
18	0.613	0.513	0.493	0.437
19	0.443	0.490	0.450	0.707
20	0.440	0.277	0.660	0.553
21	0.550	0.723	0.347	0.557
22	0.420	0.617	0.443	0.433
23	0.447	0.600	0.490	0.467
24	0.430	0.583	0.387	0.613
25	0.460	0.343	0.607	0.583
26	0.493	0.693	0.340	0.567
27	0.563	0.583	0.513	0.120
28	0.517	0.470	0.530	0.670
29	0.773	0.453	0.080	0.237
Aggregate	Brand1	Brand2	Brand3	Brand4
	0.587	0.400	0.410	0.623

For our application, we use *monthly* as well as *aggregate* shares for the four brands as test statistics. As Model 1 had the highest log-marginal likelihood, we report the posterior predictive checks for only this model. Table 1 contains the p-values corresponding to our test statistics.

It is clear from the table that the model is adequate in representing both monthly and aggregate shares as none of the p-values are extreme.

4.4. Parameter Estimates

As Model 1 has the greatest support from the data, we only present the parameter estimates for this model. As is usual for Bayesian inference, we summarize the posterior distributions of the parameters by reporting their posterior means and standard deviations. Figure 6 shows the time evolution of the population-level coefficients in $\boldsymbol{\mu}_t$. The figure shows that preferences vary considerably over the short and long term. The short-term differences are mostly due to sampling variation, whereas the longer-term variation might reflect changes in the relative preference of brands. For instance, the intercept for Brand 3 suggests an increasing trend. This suggests that the intrinsic preference for Brand 3 is increasing relative to that of Brand 4, which is the base brand.

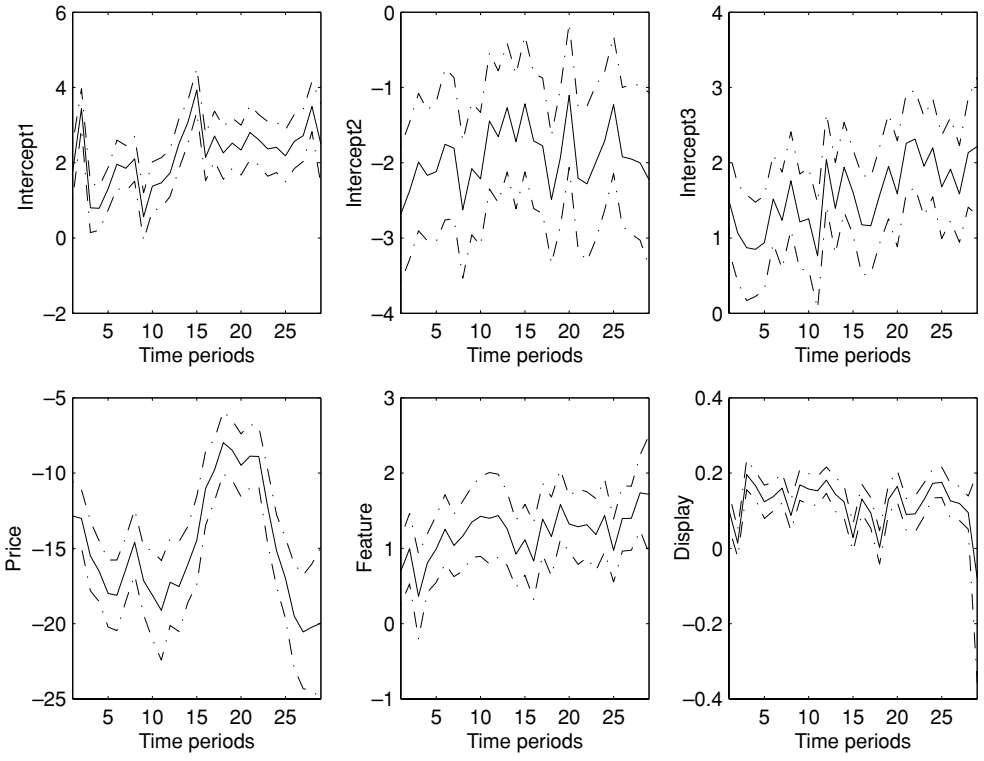

FIGURE 6.
Evolution of brand intercepts and marketing variable sensitivities over time (29 months) in the Bayesian state-space, heterogeneous multinomial model. The solid lines denote the posterior means and the dashed lines denote one standard deviation posterior bands.

Feature sensitivity shows a mildly increasing trend over the long-term. This increase cannot be attributed to increased promotional activity as Figure 5 shows that both feature and display activity was fairly stable across the time period of the data. Figure 6 also shows that the display sensitivity appears to be relatively stable for most of the data. The substantial time variation in the coefficients, when considered together with the change in the LML upon removing parameter-driven dynamics from Model 1, suggests that using a state-space approach is important in our application. In addition, the sensitivities for the marketing variables have the expected signs - the price sensitivity is negative whereas the display and feature sensitivities are positive throughout the 29 months of data.

Table 2 characterizes the magnitude of state-space dynamics and unobserved heterogeneity. The table reports the posterior means and the corresponding posterior standard deviations for the elements of $\mathbf{\Omega}$ and $\mathbf{\Lambda}$. The top panel of Table 2 pertains to the elements in $\mathbf{\Omega}$. Recall that $\mathbf{\Omega}$ represents the scale matrix of the time-varying random shocks that influence the transition equation. Most of the diagonal elements of $\mathbf{\Omega}$ are large relative to their corresponding posterior standard deviations, suggesting that the unobserved effects capture important sources of variation in the population means over time. However, most of the off-diagonal elements do not appear to be significantly different from zero. This suggests that the population-level coefficients for the marketing variables such as price, feature and display in $\boldsymbol{\mu}_t$ evolve more or less independently of each other over time.

The bottom panel shows the posterior means and standard deviations for the elements in $\mathbf{\Lambda}$, the covariance matrix for the baseline normal distribution for the DP prior used in our

TABLE 2.
Time-varying and cross-sectional heterogeneity.

	Intercept 1	Intercept 2	Intercept 3	Price	Feature	Display
Ω						
Intercept 1	0.75 (0.27)	0.34 (0.23)	0.04 (0.16)	−0.56 (0.65)	0.13 (0.18)	−0.09 (0.25)
Intercept 2		0.78 (0.31)	−0.06 (0.16)	−0.62 (0.71)	0.09 (0.18)	−0.07 (0.26)
Intercept 3			0.56 (0.19)	0.06 (0.47)	0.03 (0.14)	−0.01 (0.19)
Price				5.61 (4.32)	−0.18 (0.56)	0.28 (0.83)
Feature					0.60 (0.22)	−0.01 (0.22)
Display						1.07 (0.48)
Λ						
Intercept 1	7.45 (3.32)	0.57 (1.12)	4.64 (2.65)	2.47 (3.22)	0.12 (0.74)	−1.01 (0.98)
Intercept 2		2.49 (1.19)	−1.42 (1.10)	3.09 (2.25)	−0.12 (0.42)	0.59 (0.61)
Intercept 3			5.90 (2.85)	−1.83 (3.13)	0.35 (0.59)	−1.28 (0.91)
Price				13.56 (9.59)	−0.74 (1.13)	0.04 (1.36)
Feature					0.90 (0.36)	0.03 (0.32)
Display						1.53 (0.62)

Note. Posterior Mean and Standard Deviations (in parenthesis) of Ω and Λ.

heterogeneity specification. The diagonal elements in $\mathbf{\Lambda}$ are larger than the corresponding elements in $\mathbf{\Omega}$. This indicates substantial heterogeneity, and is also consistent with the relative changes in the LML when heterogeneity or dynamics are included. Finally, focusing on the covariation in the utility errors, we see that the correlation coefficients, r_{ij}, between utility equations i and j $(i, j = 1, 2, 3,\ i \neq j)$ have the posterior means: $r_{12} = -0.12,\ r_{13} = 0.45,\ r_{23} = 0.24$. For identification, the standard deviation associated with the first utility equation was set to 1. Thus, the remaining free standard deviations within $\mathbf{\Sigma}$ have posterior means $\sqrt{\sigma_{22}} = 1.61$ and $\sqrt{\sigma_{33}} = 1.42$. Given the identification restrictions on both the location and scale of the utilities, an unambiguous interpretation of the correlations and standard deviations in terms of the true covariance matrix of errors is not possible.

We now discuss the posterior estimates of the various statistical quantities associated with the Dirichlet Process Priors. We assumed that the scale mixing parameter associated with state-space transition errors, $\kappa_{t(ij)}$ is distributed according to G_κ, which in turn has a Dirichlet Process Prior with a baseline distribution $\text{Ga}(\nu_\kappa/2, \nu_\kappa/2)$ and a precision parameter α_κ. The precision parameter determines how the unknown distribution differs from the baseline gamma distribution and its value is inferred from the data. Our estimation indicates that α_κ has a mean of 2.24 with a standard deviation of 0.75. In addition, the number of "clusters" has a mean of around 12 with a standard deviation of 1.71. Considering that we have 29 time periods in our data, there seems to be some clustering of the scale parameters.

For the heterogeneity component, we assumed that $\boldsymbol{\lambda}_i$ is distributed with some unknown distribution G_λ, which has a Dirichlet process prior with a base distribution, $N(0, \mathbf{\Lambda})$ and a precision parameter α_λ. We find that the estimated mean of this precision parameter is 7.96 with a standard deviation of 2.45. Further, the estimated number of "clusters" has a mean of 36 and a standard deviation of 7.12. As our sample has 300 consumers, this small estimated number of clusters suggests that the heterogeneity across households can be adequately modeled by a discrete distribution with relatively few mass points.

Further insight on the nature of heterogeneity is obtained by examining the predictive distributions of the household-specific coefficients for a random, yet unobserved household. The predictive distribution for such a future household is easily obtained by using the Polya-urn representation of the Dirichlet process priors (Blackwell & MacQueen, 1973). Figure 7 shows

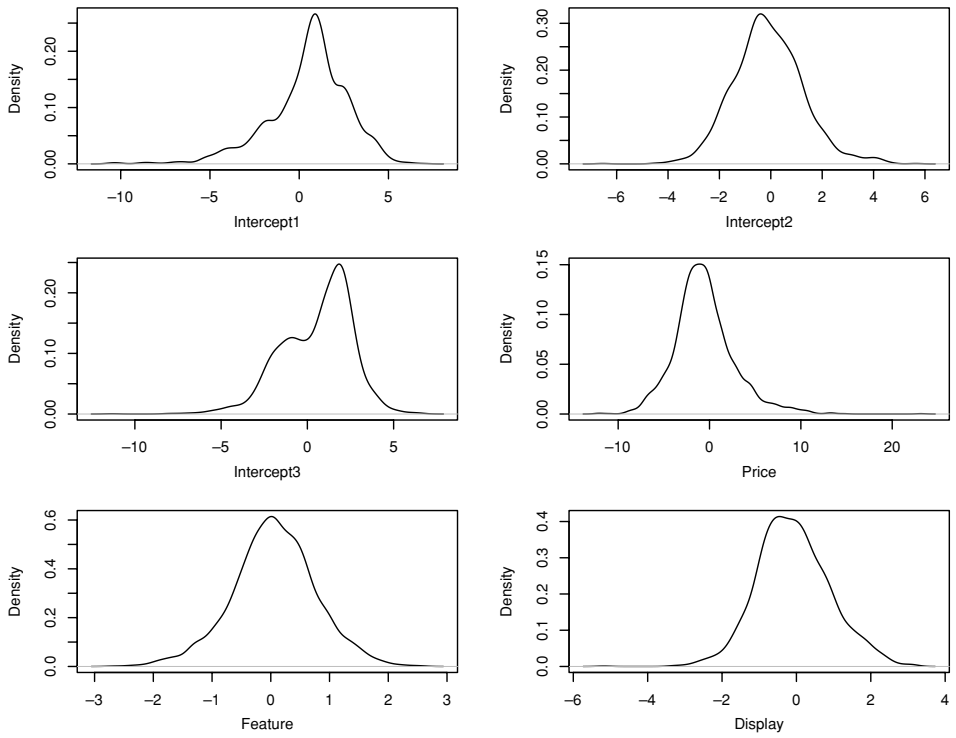

FIGURE 7.
Predictive distributions for brand intercepts and marketing variable sensitivities.

the marginal predictive distributions for the six household-specific parameters in $\boldsymbol{\lambda}_i$. These reveal that a majority of these predictive distributions appear to be unimodal, however, the distribution of the intercept for the third utility shows pronounced bimodality. This suggests the possibility of customer segments with different intrinsic preferences for Brand 3 relative to the base brand (Brand 4).

5. Conclusion

We developed a flexible semiparametric Thurstonian model based on Dirichlet process priors for modeling longitudinal datasets. The model accommodates both cross-sectional heterogeneity and parameter-driven dynamics in a robust fashion and nests several commonly used models as special cases. We showed how MCMC methods can be used to circumvent the need for complex numerical integration necessary for the evaluation of the model posterior.

In two simulation studies, we investigated the benefits of semiparametrically specifying the heterogeneity component of the model. The simulations provide evidence that the mixture of Dirichlet process model is more accurate in parameter recovery compared to a model with a normal population distribution when the true distribution of heterogeneity in the population is discrete and does as well as a model with a normal population distribution when the true population distribution of the parameters is normal. The simulations also showed that the semiparametric model reveals multimodality in the heterogeneity, whereas, a normal population distribution is incapable of doing the same. We also applied our model on household-level purchasing data in the tuna product category. Our model comparison results show that incorporating both parameter-driven dynamics and household-level variation in preferences is important in our application.

While we focused on Thurstonian models involving multivariate binary, multivariate ordinal as well as multinomial data, our methods can be adapted to the analysis of rank-ordered data (Yao & Böckenholt, 1999).

Appendix: Priors and Full conditional distributions

Prior Distributions

The Bayesian approach requires prior distributions over the parameters ($\boldsymbol{\mu}_0$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}$, α_κ, α_λ, α_γ). We specify independent and diffuse (but proper) priors.

The prior for the precision matrix, $\boldsymbol{\Omega}^{-1}$, is Wishart, $W[\rho_\omega, (\rho_\omega \mathbf{R}_\omega)^{-1}]$, where $\rho_\omega \geq L_1$ is the "degree of freedom" parameter for the Wishart and $\mathbf{R}_\omega$ is a $L_1 \times L_1$ scale matrix. For our parameterization, $\mathbf{R}_\omega$ is the expected prior variance of the time-varying parameters, $\boldsymbol{\mu}_t$'s. We set $\rho_\omega = L_1 + 1$ and $\mathbf{R}_\omega = \mathbf{I}$.

Similarly, the prior for the precision matrix $\boldsymbol{\Lambda}^{-1}$ is $W[\rho_\lambda, (\rho_\lambda \mathbf{R}_\lambda)^{-1}]$, where $\rho_\lambda \geq L_2$ and $\mathbf{R}_\lambda$ is $L_2 \times L_2$. We set $\rho_\lambda = L_2 + 1$, and $\mathbf{R}_\lambda = \mathbf{I}$. The prior for initial population mean $\boldsymbol{\mu}_0$ is multivariate normal $N(\boldsymbol{m}_0, \mathbf{C}_0)$. We set $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{C}_0 = 100\mathbf{I}$ to obtain a diffuse prior. For the precision parameters α_γ, α_κ and α_λ of the three Dirichlet Processes, following Escobar and West (1998), we assume independent gamma priors $\text{Ga}(a_\gamma, b_\gamma)$, $\text{Ga}(a_\kappa, b_\kappa)$ and $\text{Ga}(a_\lambda, b_\lambda)$ respectively.

The covariance matrix $\boldsymbol{\Sigma}$ has different restrictions depending upon the type of data. For multinomial data, one of its variance elements is fixed at one. For multivariate binary and ordinal data, $\boldsymbol{\Sigma}$ is a correlation matrix. We focus on the multinomial case. Following Barnard, McCulloch, & Meng (2000) the covariance matrix Σ can be decomposed into a correlation matrix, Γ and a vector $\boldsymbol{s}$, of standard deviations, i.e., $\Sigma = \text{Diag}(\boldsymbol{s}) \times \Gamma \times \text{Diag}(\boldsymbol{s})$, where $\boldsymbol{s} = (\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{JJ}})'$. The correlation matrix $\boldsymbol{\Gamma}$ has $J_f = K(K-1)/2$ non-redundant correlations. Let $vec(\Gamma)$ be a vector of these J_f correlations. We assume a multivariate normal prior over $vec(\Gamma)$, $p(vec(\Gamma)) \propto \exp[-\frac{1}{2}(vec(\Gamma) - vec(\Gamma_0))' H_0 (vec(\Gamma) - vec(\Gamma_0))]$ where $vec(\Gamma) \in A$, a subset of the hypercube$[-1, 1]^{J_f}$ that leads to a proper correlation matrix, $vec(\Gamma_0)$ is a vector of prior means for the correlations, which we set to zero, and H_0 can be set to an identity matrix. Let $\boldsymbol{\tau}$ contain the logarithms of the elements in $\boldsymbol{s}$. We assume that the log-standard deviations in $\boldsymbol{\tau}$ come from independent standard normal priors. For multivariate binary and ordinal data, $\boldsymbol{\Sigma}$ is a correlation and thus its prior can be specified as for $\boldsymbol{\Gamma}$ above.

For the ordinal decisions, priors are needed for the thresholds. The prior distributions for all other unknowns are as for the multivariate binary choice model. As the thresholds are order constrained, we use a transformation of the thresholds proposed by Fahrmeir and Tutz (1994) to obtain unconstrained thresholds and then use a multivariate normal distribution over these unconstrained parameters. Specifically, the threshold vector $\boldsymbol{\theta}_k$ for each question $k \leq K$ can be transformed to a vector $\boldsymbol{\zeta}_k = \{\zeta_{k1}, \ldots, \zeta_{kM-1}\}$ such that $\zeta_{k1} = \log(\theta_{k1})$, and $\zeta_{kj} = \log(\theta_{kj} - \theta_{kj-1})$, $2 \leq j \leq M-1$. The inverse mapping, $\theta_{kj} = \sum_{l=1}^{j} \exp(\zeta_{kj})$, $1 \leq j \leq M-1$, can be used to obtain the original thresholds for the individuals. We assume independent multivariate normal priors $\boldsymbol{\zeta}_k \sim N(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ for the thresholds associated with each ordinal response. For simplicity, the values for $\boldsymbol{\pi}_k$ and $\boldsymbol{\Psi}_k$ can be chosen to be the same across all response variables. Moreover, the different thresholds for a variable k can be considered a priori independent, i.e., $\boldsymbol{\Psi}_k$ can be considered a diagonal matrix.

Full Conditional Distributions

The full conditional distributions for the MCMC steps are described below.

(1) The full conditional distribution for the individual-level parameters $\boldsymbol{\lambda}_i$ is based on the Polya-urn representation of the Dirichlet Process (Blackwell and MacQueen 1973). Integrating over the random population distribution, the prior for $\boldsymbol{\lambda}_i$, is a mixture of distributions of the form:

$$\boldsymbol{\lambda}_i|\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_{i-1} \sim \frac{1}{i-1+\alpha_\lambda}\sum_{m=1}^{i-1}\delta_{\boldsymbol{\lambda}_m} + \frac{\alpha_\lambda}{i-1+\alpha_\lambda}G_{\lambda 0}, \tag{1}$$

where $G_{\lambda 0}$ is the baseline distribution $N(0, \boldsymbol{\Lambda})$. Therefore, the conditional posterior for $\boldsymbol{\lambda}_i$ can be written as:

$$p(\boldsymbol{\lambda}_i \mid \{\boldsymbol{\lambda}_l, l \neq i\}, \{\mathbf{u}_{ij}\}, \{\boldsymbol{\mu}_t\}, \boldsymbol{\Lambda}, \{\gamma_{ij}\}, \boldsymbol{\Sigma}) = q_0 G_b(\boldsymbol{\lambda}_i|.) + \sum_{k\neq i} q_k \delta_{\boldsymbol{\lambda}_k}, \tag{2}$$

where,

- $G_b(\boldsymbol{\lambda}_i|.)$ is the multivariate normal $N(\hat{\boldsymbol{\lambda}}_i, \mathbf{V}_i)$ baseline posterior distribution. The precision matrix $\mathbf{V}_{\lambda_i}^{-1} = \boldsymbol{\Lambda}^{-1} + \sum_{j=1}^{n_i}\mathbf{W}_{ij}'(\gamma_{ij}\boldsymbol{\Sigma}^{-1})\mathbf{W}_{ij}$ and $\hat{\boldsymbol{\lambda}}_i = \mathbf{V}_{\lambda_i}\,[\sum_{j=1}^{n_i}\mathbf{W}_{ij}'(\gamma_{ij}\boldsymbol{\Sigma}^{-1})\tilde{\mathbf{u}}_{ij}]$, where the adjusted utilities $\tilde{\mathbf{u}}_{ij} = \mathbf{u}_{ij} - \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)}$.
- The weight, $q_0 \propto \alpha_\lambda f_i$ where $f_i = \int f_N(\mathbf{u}_i \mid \boldsymbol{\mu}_t, \boldsymbol{\lambda}_i, \{\gamma_{ij}\}, \boldsymbol{\Sigma})\; f_N(\boldsymbol{\lambda}_i)d\boldsymbol{\lambda}_i$ is the marginal density of the utilities for individual i under the baseline prior $N(\mathbf{0}, \boldsymbol{\Lambda})$ and is obtained by integrating out the random effects. Let $\mathbf{W}_i$ be obtained by stacking the observation specific matrices $\mathbf{W}_{ij}$ one below another. Similarly, define $\mathbf{X}_i$ by stacking the matrices $\mathbf{X}_{ij}$ and let $\mathbf{D}_{i\gamma}$ be a diagonal matrix containing the γ_{ij}^{-1}'s associated with the n_i observations of individual i. Finally, let $\tilde{\mathbf{u}}_i$ be obtained by stacking the adjusted utilities. The marginal density is the density at $\tilde{\mathbf{u}}_i$ of the multivariate normal distribution $N(\mathbf{0}, \mathbf{W}_i\boldsymbol{\Lambda}\mathbf{W}_i' + \mathbf{D}_{i\gamma}\otimes\boldsymbol{\Sigma})$. Thus,

$$q_0 = \alpha_\lambda\left[\frac{1}{(2\pi)^{Kn_i/2}|\boldsymbol{\Upsilon}_i|^{1/2}}\exp\left(-\frac{1}{2}\tilde{\mathbf{u}}_i\boldsymbol{\Upsilon}_i^{-1}\tilde{\mathbf{u}}_i'\right)\right],$$

  where $\boldsymbol{\Upsilon}_i = \mathbf{W}_i\boldsymbol{\Lambda}\mathbf{W}_i' + \mathbf{D}_{i\gamma}\otimes\boldsymbol{\Sigma}$. Because of the special structure for $\boldsymbol{\Upsilon}_i$, computational shortcuts based on Woodbury's identiry (Harville, 1997) are readily available for computing its inverse and determinant.
- The weight q_k is proportional to $f_N(\mathbf{u}_i \mid \boldsymbol{\lambda}_k) = \prod_{j=1}^{n_i} f_N(\mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_k, \gamma_{ij}^{-1}\boldsymbol{\Sigma})$, which is the normal density of the utilities for individual i evaluated using individual k's parameters.

Thus, with probability q_0, we sample $\boldsymbol{\lambda}_i$ from the full conditional under the baseline distribution and with probability q_k, we select it from the mass point distribution $\delta_{\boldsymbol{\lambda}_k}$, i.e., we set $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_k$. This results in a mixture with one component being a normal distribution and all other component being point masses.

(2) As the population distribution is discrete, different individuals can share the same parameters. Thus individuals can be naturally grouped into clusters. In any iteration, there is some number I^*, $(0 < I^* \leq I)$ of unique values for the user-specific $\boldsymbol{\lambda}_i$s. Denote the unique parameters as $\boldsymbol{\lambda}_m^*$, $m = 1\ldots I^*$. These can be interpreted as cluster-specific coefficients. In addition, let S_m represent the set of individuals with common random effects $\boldsymbol{\lambda}_m^*$ and let $|S_m|$ be the number of such individuals. Bush and MacEachern (1996) and Ishwaran and James (2001) suggest remixing of these cluster-specific parameters to aid the convergence for the MCMC algorithm. After determining the cluster structure of the individual-specific parameters, the cluster-level parameters $\boldsymbol{\lambda}_m^*$ are recomputed from the conditional density $p(\boldsymbol{\lambda}_m^*|.) = N(\hat{\boldsymbol{\lambda}}_m^*, \hat{\mathbf{V}}_{\lambda_m})$, where $\hat{\mathbf{V}}_{\lambda_m}^{-1} = \boldsymbol{\Lambda}^{-1} + \sum_{l=1}^{|S_m|}\sum_{j=1}^{n_l}\mathbf{W}_{lj}'(\gamma_{lj}\boldsymbol{\Sigma}^{-1})\mathbf{W}_{lj}$ and $\hat{\boldsymbol{\lambda}}_m^* = \hat{\mathbf{V}}_{\lambda_m}[\sum_{l=1}^{|S_m|}\sum_{j=1}^{n_l}\mathbf{W}_{lj}'(\gamma_{lj}\boldsymbol{\Sigma}^{-1})\tilde{\mathbf{u}}_{lj}]$. As the unique values of the random effects can come only from the baseline distribution, and because the baseline is conjugate with the

normal utilities, the full conditional distribution is normal. Selecting a new value for the cluster parameter $\boldsymbol{\lambda}_i^*$ changes the $\boldsymbol{\lambda}_i$s for all the users within that cluster.

(3) The posterior distribution of $\boldsymbol{\Lambda}^{-1}$ is Wishart and can be written as $p(\boldsymbol{\Lambda}^{-1} \mid \boldsymbol{\mu}, \{\boldsymbol{\lambda}_i^*\}) = W(\rho_\lambda + I^*, [\rho_\lambda \boldsymbol{R}_\lambda + \sum_{m=1}^{I^*} \boldsymbol{\lambda}_m^* \boldsymbol{\lambda}_m^{*'}]^{-1})$.

(4) The precision parameter α_λ can be sampled using data-augmentation (West, 1992). At the nth iteration of the MCMC sampler, we first sample a latent variable η_1 from the beta distribution $(\eta_1 | \alpha_\lambda^{(n-1)}, I) \sim Be(\alpha_\lambda^{(n-1)} + 1, I)$, which has mean $(\alpha_\lambda^{(n-1)} + 1)/(\alpha_\lambda^{(n-1)} + 1 + I)$. Then $\alpha_\lambda^{(n)}$ is sampled from the mixture of gamma distributions, i.e.,

$$\begin{aligned}\left(\alpha_\lambda^{(n)} | \eta_1, I^{*(n-1)}\right) \sim \pi_{\eta_1} Ga\left(a_\lambda + I^{*(n-1)}, b_\lambda - \log(\eta_1)\right) \\ + (1 - \pi_{\eta_1}) Ga\left(a_\lambda + I^{*(n-1)} - 1, b_\lambda - log(\eta_1)\right),\end{aligned} \tag{3}$$

where the weights π_{η_1} are defined in odds forms by $\frac{\pi_{\eta_1}}{1-\pi_{\eta_1}} = \frac{a_\lambda + I^{*(n-1)} - 1}{I(b_\lambda - log(\eta_1))}$.

(5) The full conditional for Ω^{-1} is Wishart and is given by $W(\rho_\omega + T, [\sum_{t=1}^{T} \kappa_t (\boldsymbol{\mu}_t - \mathbf{B}_t \boldsymbol{\mu}_{t-1})(\boldsymbol{\mu}_t - \mathbf{B}_t \boldsymbol{\mu}_{t-1})' + \rho_\omega \mathbf{R}_\omega]^{-1})$.

(6) For multinomial choice models, $\boldsymbol{\Sigma}$ is decomposed into a vector of standard deviations and a correlation matrix and the first standard deviation σ_{11} is set to 1 (i.e., the first term in $\boldsymbol{\omega}$, which is the vector of log standard deviations is fixed to 0) for identification. The full conditional distribution for each free element in $\boldsymbol{\omega}$ is unknown and we use Metropolis-Hastings steps to simulate the free elements in $\boldsymbol{\omega}$. A univariate normal proposal density is used to generate candidates for this procedure. If $\omega_l^{(m-1)}$ is the current value of lth component of $\boldsymbol{\omega}$, then a candidate value is generated using a random walk chain $\omega_l^c = \omega_l^{(m-1)} + N(0, \tau)$, where τ is a tuning constant that controls the acceptance rate. The proposed candidate ω_l^c is accepted as the new value $\omega_l^{(m)}$ with probability,

$$\min\left[\frac{L(\omega_l^c|.)p(\omega_l^c)}{L(\omega_l^{(m-1)}|.)p(\omega_l^{(m-1)})}, 1\right], \tag{4}$$

where, $L(\omega_l^c|.)$ is conditioned on other parameters and is the likelihood of observing the latent utilities, evaluated at the candidate value ω_l^c. The quantity $p(\omega_l^c)$ represents the prior normal density evaluated at ω_l^c. If the candidate is rejected, then $\omega_l^{(m)} = \omega_l^{(m-1)}$.

(7) Many different approaches can be used to sample the correlation matrix $\boldsymbol{\Gamma}$. Here, we use slice sampling to sample each non-redundant correlation in $\boldsymbol{\Gamma}$, separately. The details of slice sampling can be found in Neal (2003).

(8) The scale mixing parameters γ_{ij} come from G_γ which comes from a Dirichlet process prior with a gamma distribution as the baseline distribution. For any γ_{ij}, let $\boldsymbol{\gamma}_{-ij}$ denote a vector containing the scale parameters for all other observations. Let l index the elements in this vector. A posterior draw for γ_{ij} can either come from the vector $\boldsymbol{\gamma}_{-ij}$ or from the baseline posterior distribution. Thus, the full conditional for γ_{ij} can be written as:

$$p(\gamma_{ij}|\mathbf{u}_{ij}, \boldsymbol{\lambda}_i, \boldsymbol{\gamma}_{-ij}, \boldsymbol{\mu}_{t(ij)}, \boldsymbol{\Sigma}) \sim q_0 G_b(\gamma_{ij}|.) + \sum_{l \neq (ij)} q_l \delta_{\gamma_l}, \tag{5}$$

where, $G_b(\gamma_{ij}|.)$ is the gamma distribution, Ga(a, b) where $a = (\nu_\gamma + K)/2$, and $b = (\nu_\gamma + (\mathbf{u}_{ij} - \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} - \mathbf{W}_{ij}\boldsymbol{\lambda}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{u}_{ij} - \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} - \mathbf{W}_{ij}\boldsymbol{\lambda}_i))/2$. The weights in (18) are given by

$$\begin{aligned} q_0 &\propto \alpha_\gamma \int f(\mathbf{u}_{ij}|\boldsymbol{\mu}_{t(ij)}, \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}) dG_{\gamma 0}, \\ &\propto \alpha_\gamma f_T(\mathbf{u}_{ij}|\mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i, \boldsymbol{\Sigma}, \nu_\gamma), \end{aligned}$$

where $f_T()$ is the pdf of a t-distribution and $q_l \propto f_N(\mathbf{u}_{ij}|\mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i, \gamma_l^{-1}\boldsymbol{\Sigma})$, where $f_N()$ is the pdf of a normal distribution.

(9) As in the full conditional (2), different observations which share the same scale parameter can be grouped into clusters. In any iteration, there is some number N^*_γ, $(0 < N^*_\gamma \leq N)$ of unique values for the observation-specific γ_{ij}s. Denote the unique scale parameters as γ^*_m, $m = 1 \ldots N^*_\gamma$. In addition, let S_m represent the set of observations with common scale parameters γ^*_m and let $|S_m|$ be the number of such observations. The cluster-level parameters γ^*_m are recomputed from the conditional density $p(\gamma^*_m|.) = \text{Ga}(df^*_{\gamma_m}, \hat{V}^*_{\gamma_m})$, where $\hat{V}^*_{\gamma_m} = (\nu_\gamma + \sum_{(ij)\in S_m}(\mathbf{u}_{ij} - \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} - \mathbf{W}_{ij}\boldsymbol{\lambda}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{u}_{ij} - \mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} - \mathbf{W}_{ij}\boldsymbol{\lambda}_i))/2$ and $df^*_{\gamma_m} = (\nu_\gamma + K|S_m|)/2$. Because of the conjugacy of the baseline gamma distribution with the normal utilities conditional on γ_{ij}, the full conditional distribution is gamma. Selecting a new value for the cluster parameter γ^*_{ij} changes the γ_{ij}s for all the observations within that cluster.

(10) As in (8) above, each scale mixing parameter κ_t associated with time period t, $t = 1$ to T, can be drawn from its full conditional which is a mixture of a Gamma distribution and mass-points concentrated on the scale-mixing parameters associated with the other time periods. This full conditional can be derived in a manner analogous to that for γ_{ij} in step (8).

(11) Analogous to the previous remixing of the scale parameters associated with each observation, the scale parameters associated with dynamics κ_t, $t = 1$ to T, can also remixed as in step (9) with appropriate modifications.

(12) The precision parameter α_κ can be sampled as in step (4) by replacing I, I^*, a_λ and b_λ with T, T^*, a_κ and b_κ respectively.

(13) Similarly, the precision parameter α_γ can be sampled as in step (4) with appropriate changes.

(14) The states $\{\boldsymbol{\mu}_t\}$, $t = 1$ to T, are sampled using the forward-filtering backward-sampling algorithm, in two steps. In a forward step, the moments of the updated distribution of each state is computed using a Kalman filter approach. In the backward step, each parameter is sampled from its conditional distribution conditioned on the preceding draw.

Redefine the model equations in terms of the time periods $t = 1$ to T.

$$\begin{aligned} \widetilde{\boldsymbol{u}}_t &= \mathbf{X}_t\boldsymbol{\mu}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \Sigma_t), \quad \Sigma_t = \Sigma \otimes \mathbf{H}_{n_t}, \\ \boldsymbol{\mu}_t &= \mathbf{B}_t\boldsymbol{\mu}_{t-1} + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t \sim N(\mathbf{0}, \kappa_t^{-1}\boldsymbol{\Omega}), \end{aligned}$$

where, n_t is the number of observations in period t, $\mathbf{H}_{n_t}$ is a diagonal matrix that contains the scale parameters γ_{ij}^{-1} for the observations pertaining to time period t. In addition, $\widetilde{\mathbf{u}}_t$ is the $n_t \times K$ vector of adjusted utilities stacked observation by observation and obtained after subtracting the systematic component $\mathbf{W}_{ij}\boldsymbol{\lambda}_i$ from the utilities for each observation and $\mathbf{X}_t$ is obtained by vertically stacking the $\mathbf{X}_{ij}$ matrices for all observations belonging to time-period t.

In the forward step, the moments of each state are computed recursively. Let the posterior at time $t-1$ be $p(\boldsymbol{\mu}_{t-1}|D_{t-1}) \sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$, where, D_{t-1} is the information set at time-period $t-1$ and includes the utilities up to time period $t-1$ and the other unknowns. The prior for $\boldsymbol{\mu}_t$ can be written as $p(\boldsymbol{\mu}_t|D_{t-1}) \sim N(\mathbf{B}_t\mathbf{m}_{t-1}, \boldsymbol{\Delta}_t)$, where, $\boldsymbol{\Delta}_t = \mathbf{B}_t'\mathbf{C}_{t-1}\mathbf{B}_t + \kappa_t^{-1}\boldsymbol{\Omega}$. After the data for period t is observed, conditional on the utilities for the observations in period t, the posterior at time t can be written as $p(\boldsymbol{\mu}_t|D_{t-1}, \widetilde{\mathbf{u}}_t) \sim N(\mathbf{m}_t, \mathbf{C}_t)$ by combining the prior and the likelihood $\widetilde{\mathbf{u}}_t \sim N(\mathbf{X}_t\boldsymbol{\mu}_t, \Sigma_t)$. Hence, the posterior parameters are given by $\mathbf{C}_t^{-1} = \boldsymbol{\Delta}_t^{-1} + \mathbf{X}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t$ and $\mathbf{m}_t = \mathbf{C}_t[\boldsymbol{\Delta}_t^{-1}\mathbf{B}_t\mathbf{m}_{t-1} + \mathbf{X}_t'\boldsymbol{\Sigma}_t^{-1}\widetilde{\mathbf{u}}_t]$. These posterior moments can be computed and stored in the forward step of the algorithm.

Let $\widetilde{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T\}$ and let the parameters that are not in $\widetilde{\boldsymbol{\mu}}$ be written as $\widetilde{\boldsymbol{\alpha}}$. Then we can write $p(\widetilde{\boldsymbol{\mu}}|D_T, \widetilde{\boldsymbol{\alpha}}) = p(\boldsymbol{\mu}_T|D_T, \widetilde{\boldsymbol{\alpha}})p(\boldsymbol{\mu}_{T-1}|\boldsymbol{\mu}_T, D_{T-1}, \widetilde{\boldsymbol{\alpha}}) \ldots p(\boldsymbol{\mu}_0|\boldsymbol{\mu}_1, D_0, \widetilde{\boldsymbol{\alpha}})$ which depends upon the identity $p(\boldsymbol{\mu}_{T-k}|\boldsymbol{\mu}_{T-k+1}, D_T, \widetilde{\boldsymbol{\alpha}}) = p(\boldsymbol{\mu}_{T-k}|\boldsymbol{\mu}_{T-k+1}, D_{T-k}, \widetilde{\boldsymbol{\alpha}})$. Therefore, $\widetilde{\boldsymbol{\mu}}$ can be sampled using the following steps:

1. Sample $\boldsymbol{\mu}_T$ from $N(\mathbf{m}_T, \mathbf{C}_T)$
2. For $t = T-1, \ldots, 0$, sample $\boldsymbol{\mu}_t$ from $p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t+1}, D_t, \widetilde{\boldsymbol{\alpha}})$,

where, $(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t+1}, D_t, \widetilde{\boldsymbol{\alpha}}) \sim N(\mathbf{q}_t, \mathbf{Q}_t)$. The precision, $\mathbf{Q}_t^{-1} = \mathbf{C}_t^{-1} + \mathbf{B}'_{t+1}(\kappa_{t+1}\boldsymbol{\Omega}^{-1})\mathbf{B}_{t+1}$, and the mean $\mathbf{q}_t = \mathbf{Q}_t[\mathbf{C}_t^{-1}\mathbf{m}_t + \mathbf{B}'_{t+1}(\kappa_{t+1}\boldsymbol{\Omega}^{-1})\boldsymbol{\mu}_{t+1}]$. For the initial state we have $p(\boldsymbol{\mu}_0|\boldsymbol{\mu}_1, D_0, \widetilde{\boldsymbol{\alpha}}) = N(\mathbf{q}_0, \mathbf{Q}_0)$, where, $\mathbf{Q}_0^{-1} = \mathbf{C}_0^{-1} + \mathbf{B}'_1(\kappa_1\boldsymbol{\Omega}^{-1})\mathbf{B}_1$, $\mathbf{q}_0 = \mathbf{Q}_0[\mathbf{C}_0^{-1}\mathbf{m}_0 + \mathbf{B}'_1(\kappa_1\boldsymbol{\Omega}^{-1})\boldsymbol{\mu}_1]$, and $\mathbf{C}_0$ and $\mathbf{m}_0$ are set to fixed values.

(15) For the ordinal model, we need the full conditionals for the thresholds. The transformed thresholds for the ordinal variables can be sampled one at a time as we assume that these are independent a priori. For each threshold ζ_{kl}, a Metropolis step is used, as its full conditional distribution cannot be written in closed form. Given the independence assumption, the covariance matrix $\boldsymbol{\Psi}_k$ for the prior $p(\boldsymbol{\zeta}_k) = N(\boldsymbol{\pi}_k, \boldsymbol{\Psi}_k)$ for response k is a diagonal matrix. The likelihood associated with $\boldsymbol{\zeta}_k$ can be written as

$$L(\{y_{ijk}\} \mid \boldsymbol{\mu}, \{\boldsymbol{\lambda}_i\}, \boldsymbol{\theta}_k(\boldsymbol{\zeta}_k)) = \prod_{i=1}^{I}\prod_{j=1}^{n_i}\prod_{l=0}^{M}\left[\boldsymbol{\Phi}\left[\frac{\theta_{kl} - \varphi_k}{\sqrt{\sigma_{k|-k}}}\right] - \boldsymbol{\Phi}\left[\frac{\theta_{kl-1} - \varphi_k}{\sqrt{\sigma_{k|-k}}}\right]\right]^{\delta_{ijkl}}, \quad (6)$$

where, φ_k represents the conditional mean of kth utility obtained from the multivariate normal distribution $\mathbf{u}_{ij} \sim N(\mathbf{X}_{ij}\boldsymbol{\mu} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i, \gamma_{ij}^{-1}\boldsymbol{\Sigma})$ and $\sigma_{k|-k}$ represents the conditional variance. The quantity $\delta_{ijkl} = 1$ if $y_{ijk} = l$ otherwise zero, and $\boldsymbol{\theta}_k$ is expressed in terms of the unrestricted $\boldsymbol{\zeta}_k$. Even though we sample each ζ_{kl} independently, the same likelihood and the same prior can be used for each threshold associated with response k. The full conditional is proportional to the product of the above likelihood and the normal prior. If $\zeta_{kl}^{(m)}$ is the current value of the lth threshold for question k, then in the $(m+1)^{th}$ step, a candidate threshold $\zeta_{kl}^{(c)}$ is generated specifying a random walk chain $\zeta_{kl}^{(c)} = \zeta_{kl}^{(m)} + h$ where h is an random normal increment with $E[h] = 0$ and $V[h] = \tau$. The value of the tuning constant τ is set so as to ensure rapid mixing. Let $\boldsymbol{\zeta}_k^{(c)}$ be the vector of thresholds for variable k such that its lth component is $\zeta_{kl}^{(c)}$ and all other components are unchanged from the previous iteration. The candidate $\boldsymbol{\zeta}_k^{(c)}$ is accepted or rejected based on the acceptance probability

$$\min\left[\frac{L(y \mid \boldsymbol{\zeta}_k^{(c)})p(\boldsymbol{\zeta}_k^{(c)})}{L(y \mid \boldsymbol{\zeta}_k^{(m)})p(\boldsymbol{\zeta}_k^{(m)})}, 1\right]. \quad (7)$$

If the candidate is accepted then $\zeta_{kl}^{(m+1)} = \zeta_{kl}^{(c)}$, otherwise $\zeta_{kl}^{(m+1)} = \zeta_{kl}^{(m)}$. The process is repeated for each ζ_{kl}, $l = 1$ to $M-1$ and for $k = 1$ to K.

(16) The full conditional distributions associated with the identified set of utilities $\mathbf{u}_{ij}$ differ across the three choice decisions.

In a multinomial choice situation, this is a K-variate normal distribution $N(\mathbf{X}_{ij}\boldsymbol{\mu}_{t(ij)} + \mathbf{W}_{ij}\boldsymbol{\lambda}_i, \gamma_{ij}^{-1}\boldsymbol{\Sigma})$ truncated over a multidimensional cone (see McCulloch and Rossi (1994) for the case of the traditional multinomial probit model). If the chosen brand is brand m, i.e., $y_{ijm} = 1$, then $u_{ijm} > \max(\mathbf{u}_{ij,-m}, 0)$; If $y_{ijm} = 0$, then $u_{ijm} < \max(\mathbf{u}_{ij,-m}, 0)$, where $\mathbf{u}_{ij,-m}$ is a $(K-1)$ - dimensional vector of all the components of $\mathbf{u}_{ij}$ excluding u_{ijm}.

For the multivariate ordinal and binary choice decisions, the underlying utilities are sampled in an analogous manner. However, as a comparison of utilities is not required, the conditional normal distribution for a utility u_{ijk} is truncated from above and below using appropriate

thresholds. For example, in an ordinal choice setting, if $y_{ijk} = l$, then the conditional distribution for u_{ijk} is truncated to lie in the interval $[\theta_{kl-1}, \theta_{kl})$. Similarly, in the binary choice situation, the utility is truncated from above at zero if the response y_{ijk} is zero, and is truncated from below at zero if $y_{ijk} = 1$.

Sampling these identified set of utilities $\mathbf{u}_{ij}$ can be done observation by observation in a data augmentation step as in Albert and Chib (1993) and McCulloch and Rossi (1994) by sampling each component of $\mathbf{u}_{ij}$ in a mini-Gibbs sampling step. The truncated multivariate normal draw for $\boldsymbol{u}_{ij}$ can be pieced together from the truncated conditional univariate normal draws of each component of $\mathbf{u}_{ij}$.

Acknowledgment

The authors thank Kamel Jedidi for comments and suggestions on an earlier draft of this paper.

References

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88,* 669–679.

Albert, J., & Chib, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. *Technical report, Department of mathematics and statistics, Bowling Green state university*.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10,* 1281–1311.

Basu, S., & Mukhopadhyay, S. (2000). Binary response regression with normal scale mixture links. In D. Dey, S. Ghosh, & B., Mallick (Eds.), *Generalized linear models: a Bayesian perspective*. NY: Marcel Dekker Inc., pp. 231–241.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distribution via polya urn schemes. *The Annals of Statistics*, *1,* 353–355.

Böckenholt, U. (1990). Multivariate Thurstonian models. *Psychometrika*, *55,* 391–403.

Böckenholt, U. (2001). Mixed effects analysis of rank ordered data. *Psychometrika*, *66,* 45–62.

Böckenholt, U. (2002). A Thurstonian analysis of preference change. *Journal of Mathematical Psychology*, *46,* 300–314.

Bradlow, E. T., & Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with no answerŠ responses. *Journal of the American Statistical Association*, *94,* 43–52.

Bush, C. A., & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika*, *83,* 275–285.

Carter, C. K., & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, *81,* 541–553.

Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika*, *63,* 360–369.

Chib, S., & Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, *110,* 67–89.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, *89,* 268–277.

Escobar, M. D. (1995). Nonparametrics Bayesian methods in hierarchical models. *Journal of Statistical Inference and Planning*, *43,* 97–106.

Escobar, M. D., & West, M. (1998). Computing Bayesian nonparametric hierarchical models. Discussion Paper, 92-A20, Duke University, ISDS.

Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modeling based on generalized linear models*. New York: Springer-Verlag.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1,* 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2,* 615–629.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66,* 269–286.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion and rejoinder). *Statistica Sinica*, *6,* 733–807.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57,* 1317–1340.

Hajivassiliou, V. (1990) "Smooth simulation estimation of panel data LDV models. Working paper, Deptartment of Economics, Yale University.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer Verlag.

Hoijtink, H. J. A., & Molenaar, W. (1997). A multidimensional item response model: Constrained latent class analysis using the gibbs sampler and posterior predictive checks. *Psychometrika*, *62,* 171–189.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of American Statistical Association*, *96,* 161–173.
Kass, R., & Raftery, A. (1995). Bayes Factors. *Journal of American Statistical Association*, *90,* 773–795.
Keane, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, *62,* 95–116.
Leichty, J. C., Fong, D. K. H., & DeSarbo, W. (2005). Dynamic Models with Individual Level Heterogeneity: Applied to Evolution During Conjoint Studies. Forthcoming, *Marketing Science*.
Longford, N. (1993). *Random coefficient models*. Oxford: Clarendon Press.
MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics Part B-Simulation and Computation*, *23,* 727–741.
Masriliez, C. J., & Martin, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Transactions on Automatic Control*, *22,* 361–371.
Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, *64,* 325–340.
Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66,* 209–228.
McCulloch, R., & Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64,* 207–240.
Meinhold, R. J., & Singpurwalla, N. D. (1989). Robustification of Kalman filter models. *Journal of the American Statistical Association*, *84,* 479–486.
Neal, R. M. (2003). Slice sampling (with discussion). *Annals of Statistics*, *31,* 705–767.
Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data. *Communication and Cognition*, *20,* 45–62.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *79,* 281–299.
Thurstone, L. L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology*, *14,* 187–201.
Van Heerde, H. J., Mela, C. F., & Manchanda, P. (2004). The Dynamic Effect of Innovation on Market Structure. *Journal of Marketing Research*, *41,* 166–183.
Wedel, M., & Kamakura, W. A. (1998) *Marketing segmentation: conceptual and methodological foundations*. New York: Kluwer Academic Publishers.
West, M. (1992) Hyperparameter estimation in Dirichlet process mixture models. Discussion Paper, 92-03, Duke University, ISDS.
Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*, 79–92.

Manuscript received 2 AUG 2005
Final version received 11 OCT 2005
Published Online Date: 5 NOV 2006