Explaining Preference Heterogeneity with Mixed Membership Modeling

Marc R. Dotson
Marriott School of Management
Brigham Young University
marc_dotson@byu.edu

Joachim Büschken
Catholic University of Eichstätt-Ingolstadt
joachim.bueschken@kuei.de

Greg M. Allenby
Fisher College of Business
Ohio State University
allenby.1@osu.edu

February 14, 2017

Explaining Preference Heterogeneity with Mixed Membership Modeling

Abstract

Choice models produce part-worth estimates that tell us what product attributes individuals prefer. However, to understand the drivers of these preferences we need to model consumer heterogeneity by specifying covariates that explain cross-sectional variation in the part-worths. In this paper we demonstrate a way to generate covariates for the upper level of a hierarchical Bayesian choice model that leads to an improvement in explaining preference heterogeneity. The covariates are uncovered by augmenting the choice model with a grade of membership model. We find improvement in model fit and inference using the covariates generated with the proposed integrated model over competing models. This paper provides an important step in both a proper accounting for extremes in preference heterogeneity and a continued synthesis between marketing models and mixed membership models, which include models for text data.

Keywords

Choice models, mixed membership models, hierarchical Bayes, grade of membership, preference heterogeneity

1 Introduction

The fact that consumers are heterogeneous in their preferences gives rise to marketing as a discipline and an industry. Choice models and associated decision tools that account for this heterogeneity allow firms to better understand what consumers prefer and have become a standard for product development and product line optimization. However, explaining preference heterogeneity remains an elusive problem. In this paper we develop an expanded choice model that improves our ability to explain preference heterogeneity by employing a novel approach to model discrete data, including binary and ratings survey data, that describe the drivers of consumer preference.

Choice modeling is an effective tool for determining what product attributes individuals prefer but it has proven less successful at explaining the heterogeneity in consumer preferences. Explaining preference heterogeneity includes identifying covariates that serve as drivers of preference and enable targeting and promotion activities. The use of hierarchical Bayes in choice modeling allows for both individual-level attribute part-worth utilities and aggregate-level preference heterogeneity parameters. Part-worth estimates tell us what attributes consumers prefer. Parameters describing preference heterogeneity are conditioned on covariates that help explain cross-sectional variation in the part-worths.

Finding covariates that are predictive of part-worths has proven difficult. The primary benefit when using a random effect distribution of heterogeneity has been accounting for unexplained heterogeneity. Using discrete variables describing possible drivers of preference as covariates, such as demographics and psychographics, is standard. However, survey data are typically used as covariates where the number of covariates makes it impractical to include interactions. Additionally, we have growing access to new sources of discrete multivariate data outside of surveys, including text data, that we expect will be a rich source of information for explaining choice yet incorporating it isn't obvious. We propose modeling this discrete multivariate data as part of the choice model in order to uncover covariates that can better explain preference heterogeneity.

In this paper we develop an expanded hierarchical Bayesian choice model where covariates for the upper level are from an integrated grade of membership model (Woodbury et al., 1978; Erosheva et al., 2007). The grade of membership model is related to latent Dirichlet allocation, which serves as a touchstone within topic modeling (Blei et al., 2003). Both are part of a larger class of models known as mixed membership models that provide individual-level, low-dimensional representations of discrete multivariate data by accounting for interactions or co-occurrence (Airoldi et al., 2014). We propose modeling discrete variables describing potential drivers of preference where the co-occurrence or interaction among drivers will help further explain preference heterogeneity. We apply our model within the robotic vacuum and smartphone categories and find we can both explain preference heterogeneity and predict choice better than models that use observed covariates directly or assume a variant latent structure.

This paper contributes to efforts at using mixed membership models to improve marketing models. The application of this class of models to marketing contexts is still in its infancy. Extant research has focused on latent Dirichlet allocation (LDA), using product reviews and online forums to inform market structure (Lee and Bradlow, 2011; Netzer et al., 2012) and to identify preferences for product features (Archak et al., 2011). Most recently, Tirunillai and Tellis (2014) use LDA to conduct brand analysis while Büschken and Allenby (2016) develop a sentence-constrained LDA to better predict review ratings. However, mixed membership models have yet to be employed in the context of choice modeling. We believe this paper provides an important step in this regard.

The remainder of the paper will be organized as follows. We specify our model in Section 2. In Section 3, we walk through a set of simulation experiments. We detail our empirical applications in Section 4. In Section 5, we compare results from our proposed model, with covariates uncovered using the grade of membership model, and alternative models. We discuss implications of and extensions to this research in Section 6.

2 Model Specification

2.1 Hierarchical Bayesian Choice Model

Hierarchical Bayesian choice models allow for the estimation of both individual and aggregate-level preference parameters, even in the presence of few observations per individual (Rossi and Allenby, 2003; Rossi et al., 2005). Decision tools associated with choice modeling make use of individual-level preference parameter estimates to forecast the results of various product policies while aggregate-level parameter estimates are employed to explain the source of individual preferences.

The likelihood in hierarchical Bayesian choice modeling is typically assumed to be a multinomial logit model such that the probability of individual n choosing product alternative j is a function of the attributes x_j that compose the given alternative and the part-worths or individual-level preferences β_n for the attributes:

$$Pr(y_n = j|\beta_n) = \frac{\exp{(x_j'\beta_n)}}{\sum_{p=1}^{P} \exp{(x_p'\beta_n)}} \tag{1}$$

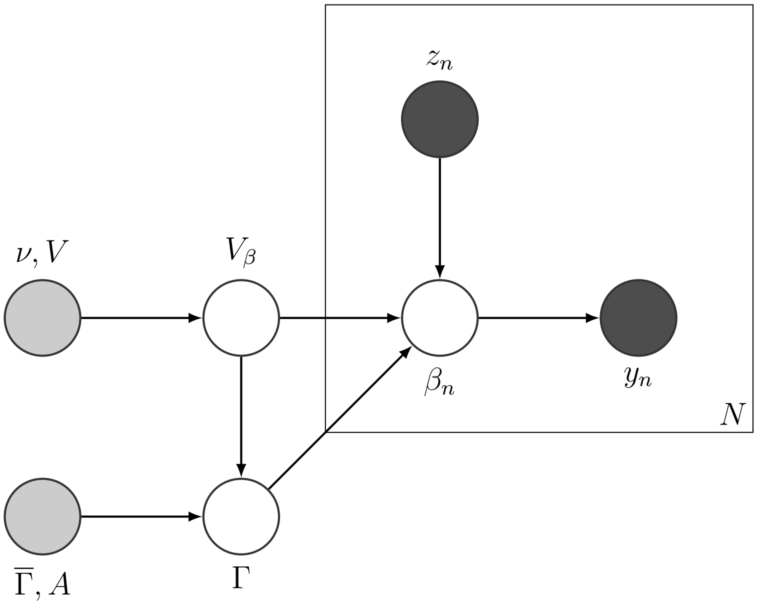
where there are a total of P alternatives to consider. The distribution of heterogeneity, or upper level, models preference heterogeneity in the individual-level β_n's. The distribution of heterogeneity is typically assumed to be multivariate normal and is characterized as:

$$\beta_n = \Gamma' z_n + \xi_n, \quad \xi_n \sim \mathrm{N}(0, V_\beta) \tag{2}$$

where z_n is a vector of covariates for individual n and Γ is a matrix of coefficients that maps variation in z_n to variation in β_n. The mean of the distribution of heterogeneity $\Gamma' z_n$ is where the analyst can specify individual-specific covariates z_n that explain variation in the part-worths. Information is shared through the estimates of Γ and the heterogeneity covariance matrix V_β to estimate individual-level β_n's (Rossi et al., 2005).

The directed acyclic graph (DAG) in Figure 1 provides a visual representation of the hierarchical Bayesian choice model. The DAG utilizes plate notation, where a plate represents replication for the enclosed variables. In the DAG, white nodes represent parameters to be estimated, grey nodes represent fixed hyper-parameters, and black nodes represent observed data.

Figure 1: Hierarchical Bayesian Choice Model



From the use of plate notation in Figure 1, we can see that the hierarchical Bayesian choice model has both aggregate and individual levels. To be clear, at the aggregate level, $\overline{\Gamma}$ is the mean and A is the precision matrix for a conjugate normal prior on Γ and ν and V are the degrees of freedom and scale matrix for a conjugate inverse Wishart prior on V_β. At the individual level, y_n is a vector of observed choices and z_n are the observed covariates for individual n. We can see that the covariates $\{z_n\}_{n=1}^N$ are chosen independent of the model specification. As discussed, the covariates $\{z_n\}_{n=1}^N$ are the key to our ability to explain preference heterogeneity. We will use DAGs, beginning with Figure 1, to help motivate the proposed model.

Following Figure 1, the joint posterior distribution of the standard hierarchical Bayes choice model is:

$$p(\{\beta_n\}_{n=1}^N, \Gamma, V_\beta | \{y_n\}_{n=1}^N, \overline{\Gamma}, A, \nu, V) \propto \left[\prod_{n=1}^N p(y_n|\beta_n) p(\beta_n|\Gamma, V_\beta) \right] p(\Gamma|V_\beta, \overline{\Gamma}, A) p(V_\beta|\nu, V) \tag{3}$$

where $\prod_{n=1}^N p(y_n|\beta_n)$ is the likelihood, $\prod_{n=1}^N p(\beta_n|\Gamma, V_\beta)$ is the distribution of heterogeneity, and $p(\Gamma|V_\beta, \overline{\Gamma}, A)$ and $p(V_\beta|\nu, V)$ are the priors (Rossi et al., 2005). The known design matrix X and covariates $\{z_n\}_{n=1}^N$ are suppressed in Equation (3).

A variety of covariates have been employed to explain preference heterogeneity in the choice modeling literature. For example, Allenby and Ginter (1995) used demographic variables, Lenk et al. (1996) included expertise, and Chandukala et al. (2011) specified consumer needs to explain variation in β_n. However, explaining preference heterogeneity has not met with much success generally (Rossi et al., 1996; Horsky et al., 2006).

One unresolved issue is that discrete covariates are often employed without a practical way to include interactions. The problem is one of dimensionality. The number of interaction terms is J choose M, where J is the number of covariates and M is the number of desired interactions. For example, with $J = 30$ covariates and $M = 2$, there are 435 possible two-way interactions, to say nothing of higher-level interactions where $M > 2$. While Chandukala et al. (2011) employ variable selection to determine which covariates matter, we are interested in a model general enough to account for interactions from traditional survey data as well as accommodate new sources of discrete data.

We propose using a non-standard model that accounts for the interaction or co-occurrence of variables to uncover covariates from discrete multivariate data for use in a choice model's random effect distribution of heterogeneity. Specifically, we propose combining a hierarchical Bayesian choice model with a grade of membership model to uncover covariates that account for interactions in order to explain preference heterogeneity better than using observed covariates directly. We first detail the grade of membership and the class of mixed membership models before specifying our expanded choice model.

2.2 The Grade of Membership Model

The grade of membership (GoM) model was developed to classify disease patterns using discrete patient-level clinical data (Woodbury et al., 1978; Clive et al., 1983). It has since been applied to modeling survey data (Erosheva et al., 2007; Gross and Manrique-Vallier, 2014). In these applications, each respondent answers a battery of survey questions with categorical responses. The research interest is to identify the patterns of co-occurrence in the categorical responses across respondents along with how each respondent relates to the patterns of co-occurrence. The GoM model characterizes these patterns of co-occurrence as profiles of archetypal respondents. Each respondent is a partial member of each of the profiles based on how similar their responses are to each pattern of co-occurrence.

Assume we have a collection of J discrete variables each with n_j categorical responses. The probability of respondent n selecting the lth category for question j is a function of the profiles λ describing the patterns of response co-occurrence across respondents and respondent n's membership vector g_n describing their partial membership in each profile:

$$Pr(w_{n,j} = l | g_n, \lambda) = \sum_{k=1}^{K} g_{n,k} \lambda_{j,k}(l) \tag{4}$$

where there are K profiles and K is specified by the analyst. The membership vector g_n for the nth respondent is constrained so that each element is non-negative and $\sum_{k=1}^{K} g_{n,k} = 1$. The λ is composed of $J \cdot K$ total vectors $\lambda_{j,k}$ each of length n_j that specify how likely each categorical response l is for question j for a hypothetical respondent that is only a member of profile k. Each $\lambda_{j,k}$ is also constrained with non-negative elements so that $\sum_{l=1}^{n_j} \lambda_{j,k}(l) = 1$.

To illustrate, consider responses to a battery of select-all-that-apply questions (i.e., pick any/J) such that $n_j = 2$ for all $J = 30$. Each respondent selects or indicates a subset of the $J = 30$ statements or items that apply to them in answer to the question: "What benefits does cereal provide that are important to you?" Figure 2(a) displays the items

Figure 2: Modeling Pick Any/J Data with a GoM Model

(a) Respondent n's Responses and Membership Vector g_n

What benefits does cereal provide that are important to you?
Item 1: It's a helpful way to get a serving of milk at the same time Item 2: Cereal is a good source of fiber Item 3: My kids will eat cereal for breakfast Item 4: Cereal isn't just for breakfast, it's a good snack anytime Item 11: I want to make sure my family has breakfast in the morning Item 15: Cereal is easy to prepare
g_n "Kids Breakfast" 0.60 "Healthy Snack" 0.20 "Source of Fiber" 0.20

(b) Aggregate-Level Profiles Defined by the Probability of Each Item $\lambda_{j,k}(1)$

$\lambda_{j,k}(1)$	"Kids Breakfast"	"Healthy Snack"	"Source of Fiber"
Item 1	0.67	0.70	0.34
Item 2	0.22	0.85	**0.95**
Item 3	**0.97**	0.13	0.04
Item 4	0.32	**0.92**	0.10
⋮	⋮	⋮	⋮
Item 30	0.04	0.13	0.14

selected (i.e., $w_{n,j} = 1$) for a given respondent n together with their membership vector g_n. Figure 2(b) displays $\lambda_{j,k}(1)$ describing $K = 3$ aggregate-level profiles in terms of the likelihood of selecting each of the $J = 30$ items. Note that since $n_j = 2$ for all $J = 30$, each $\lambda_{j,k}$ is a vector with two elements such that $\lambda_{j,k}(0)$ is the complement of the values listed in Figure 2(b). Thus $\lambda_{j,k}(0) + \lambda_{j,k}(1) = 1$ for each $\lambda_{j,k}$. This simple illustration was selected in order to ensure the profiles displayed in Figure 2(b) were easy to read. The profiles displayed in Figure 2(b) should not be confused with topics from an LDA model. To be clear, the GoM model generalizes beyond the simple case of $n_j = 2$ for all J.

Using Figure 2, we can see how profiles emerge based on what items co-occur. For example, if item 11 "I want to make sure my family has breakfast in the morning" and item 3 "My kids will eat cereal for breakfast" are selected together frequently across respondents, this pattern may be part of a profile describing concern with breakfast for children. In Figure 2(a), the membership vector g_n describes the partial membership respondent n has in each of the $K = 3$ profiles – "Kids Breakfast," "Healthy Snack," and "Source of Fiber" – where the number of profiles $K = 3$ has been specified by the analyst and the weight given to each profile is determined by how similar respondent n's response pattern matches each of the aggregate-level profiles. For this particular respondent, they are primarily a member of the "Kids Breakfast" profile, with a weight of 0.60, while still being a partial member of the remaining two profiles. The membership vector g_n has non-negative elements and is constrained to sum to 1.

The aggregate-level values $\lambda_{j,k}(1)$ in Figure 2(b) describe how likely it is for each item to occur within each profile. The profiles are composed of all $J = 30$ items with the item that is most likely within each profile in bold. Based on common response patterns across respondents, the profiles describe archetypal or extreme respondents, ones that in this case are either concerned wholly with cereal for "Kids Breakfast," a "Healthy Snack," or a "Source of Fiber," where the profile names have been determined by the analyst based on which items differentiate each profile. Thus each membership vector g_n describes where a respondent n is located within a convex hull defined by the profiles. These profiles account for the co-occurrence of the discrete items while reducing the dimensionality from J to K.

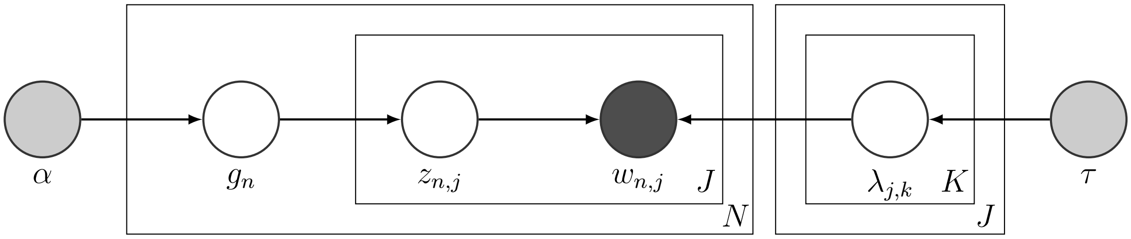
With this illustration in mind, we can apply Equation (4) to show that the probability of respondent n selecting item 1 "It's a helpful way to get a serving of milk at the same time" is a function of g_n, their partial membership in each profile, and $\lambda_{1,k}(1)$, how likely it is for item 1 to be selected in each profile. This results in a probability of $(0.60)(0.67) + (0.20)(0.70) + (0.20)(0.34) = 0.61$. Erosheva et al. (2007) use this relationship to introduce a latent profile assignment for each item.

Assuming that the J responses are conditionally independent given each membership vector g_n, the GoM likelihood is:

$$p(\{w_n\}_{n=1}^N | \{z_n\}_{n=1}^N, \lambda) p(\{z_n\}_{n=1}^N | \{g_n\}_{n=1}^N) = \prod_{n=1}^N \prod_{j=1}^J \prod_{k=1}^K g_{n,k}^{I(z_{n,j}=k)} \lambda_{j,k}(w_{n,j})^{I(z_{n,j}=k)} \qquad (5)$$

where z_n is a J-dimensional vector of latent profile assignments for respondent n, following notation typical to data augmentation (Tanner and Wong, 1987; Rossi and Allenby, 2003). Note that the latent variables z_n in Equation (5) are different from the observed covariates specified in Equation (2). Both $p(\{w_n\}_{n=1}^N|\{z_n\}_{n=1}^N, \lambda)$ and $p(\{z_n\}_{n=1}^N|\{g_n\}_{n=1}^N)$ are multinomial distributions.

Figure 3: The Grade of Membership Model



The DAG in Figure 3 provides a visual representation of the GoM model. The plate notation demonstrates the three model levels: item, respondent, and aggregate. The aggregate-level λ describing profiles is homogeneous while the respondent-level membership vectors g_n are heterogeneous. To be clear, α and τ are both hyper-parameters for conjugate Dirichlet priors on g_n and λ. Following Figure 3, the joint posterior distribution of the grade of membership model is:

$$p(\{z_n\}_{n=1}^N, \{g_n\}_{n=1}^N, \lambda|\{w_n\}_{n=1}^N, \alpha, \tau) \propto \left[\prod_{n=1}^N p(w_n|z_n, \lambda)p(z_n|g_n)p(g_n|\alpha)\right] p(\lambda|\tau) \qquad (6)$$

where $\prod_{n=1}^N p(w_n|z_n, \lambda)p(z_n|g_n)$ is the likelihood and $\prod_{n=1}^N p(g_n|\alpha)$ and $p(\lambda|\tau)$ are priors.

In the marketing literature, it has been argued that identifying extreme responses is important for designing and promoting successful new products (Allenby and Ginter, 1995). For example, extreme response behavior can be used to more efficiently target prospects with a high probability of adopting an innovation. Conceptualizing consumer heterogeneity as a continuous distribution of preferences has been shown to aid in the identification of extreme responses (Allenby et al., 1998; Allenby and Rossi, 1998). The

GoM model represents discrete response behavior as a continuous proximity to a limited number of extreme profiles. Given that marketers often search for a limited number of product offerings for reasons of efficiency or resource limitations, a concept of heterogeneity that expresses differences among consumers in the space of a small number of extreme response profiles is appealing. We utilize the GoM model given this characterization of heterogeneity, which includes the respondent-level membership vectors g_n, in the development of our proposed model.

2.2.1 Relationship with Finite Mixture Models

Having a respondent-level membership vector g_n that consists of non-negative, real-valued latent variables that sum to one is the distinctive feature of mixed membership models, the class of models that includes the GoM model and LDA. Contrast this with the general form of a finite mixture model (Kamakura and Russell, 1989):

$$p(x_n) = \sum_{k=1}^{K} g_k p_k(x_n) \tag{7}$$

where x_n is response data for respondent n. We see that the finite mixture model has a membership vector g_k at the aggregate level while the GoM model in Equation (4) has a membership vector g_n at the individual level. This feature is common to all mixed membership models and illustrates why they are often referred to as individual-level mixture models.

Finite mixture models are a special case of mixed membership models (Erosheva et al., 2007; Galyardt, 2014). However, our use of the GoM within the class of mixed membership models is different than the typical use of finite mixture models in choice modeling. Instead of specifying a mixture of distributions of heterogeneity, we are interested in using the respondent-level membership vector g_n to serve as covariates that can further explain preference heterogeneity.

2.2.2 Relationship with Factor Analysis

Factor analysis is another related model and has long been a standard approach in marketing for dimension reduction (Stewart, 1981). The basic assumption is that a set of variables can be reduced to one or more latent constructs called factors. The data are assumed to arise in the following fashion:

$$x_{n,j} = c_j + \sum_{k=1}^{K} \zeta_{n,k} \lambda_{j,k} + \eta_{n,j}, \quad \eta_{n,j} \sim \mathrm{N}(0,1) \tag{8}$$

where c_j is a constant vector, ζ_n is a respondent-level vector of factor scores, and the collection of $\lambda_{j,k}$ is a matrix of aggregate-level regression coefficients known as factor loadings. The form of factor analysis in Equation (8) is similar to that of the GoM model in Equation (4), with factor scores in place of the membership vector and factors in place of the profiles. Erosheva (2002) even demonstrates that the GoM model is equivalent to a binary factor analysis with an identity link function. However, there are key differences in the two approaches.

Factor analysis and GoM models differ in terms of their underlying assumptions, modeling objectives, and the type of data each method can process (Manton et al., 1994; Marini et al., 1996). First, standard factor analysis, as demonstrated in Equation (8), assumes continuous data. Even using a cut-point model, which assumes the observed data are discrete indicators of latent continuous variables, the latent constructs (i.e., factors) are still considered to be continuous. On the other hand, the GoM model assumes both discrete data and discrete latent constructs (i.e., profiles).

Second, the objective of factor analysis is to uncover latent constructs underlying a set of *variables*. The objective of the GoM model is to both uncover profiles representing extreme characterizations of *respondents* and measure each respondent's proximity to these profiles. In other words, the GoM model has the description of respondents and respondent heterogeneity as the objects of inference. Finally, unlike factor analysis, the

GoM model can handle a combination of multinomial, ordinal, and other discrete multivariate data. For more detail on the comparison between factor analysis and the GoM model, see Appendix B.

2.3 Integrated Hierarchical Bayesian Choice and GoM Model

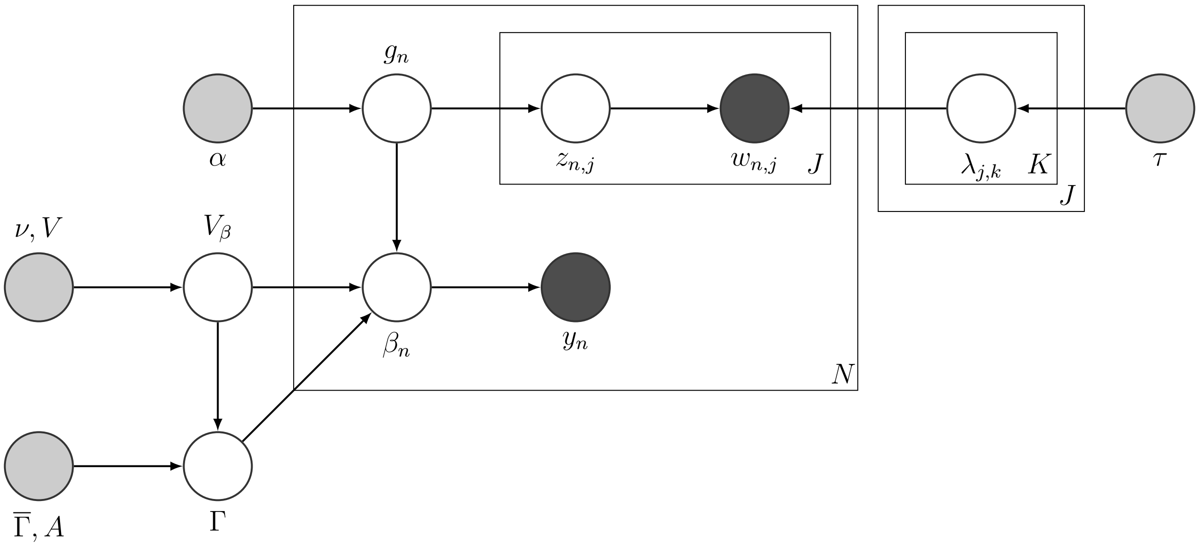
The proposed model integrates a hierarchical Bayesian choice model with a GoM model in order to use discrete multivariate data to uncover covariates that explain preference heterogeneity. A related concept is presented in the form of a supervised latent Dirichlet allocation (sLDA). In the sLDA topic model, each collection of discrete data (i.e., document, in the context of topic modeling) is paired with and used to be predictive of a response, such as using movie reviews to predict movie ratings (Blei and McAuliffe, 2007). We employ the same kind of pairing between a collection of discrete data and response, however our response is part-worth utility parameters and the collection of discrete data is from a battery of survey questions.

The individual-level choice model remains a multinomial logit, as specified in Equation (1), and the distribution of heterogeneity remains multivariate normal as in Equation (2). Since there is a separate g_n for each respondent in the GoM model in Equation (4), we use these membership vectors as covariates to explain heterogeneity in the part-worths β_n (i.e., $\beta_n = \Gamma' g_n + \xi_n$). Thus the likelihood of the integrated model is:

$$\begin{aligned} & p(\{y_n\}_{n=1}^N, \{w_n\}_{n=1}^N | \{\beta_n\}_{n=1}^N, \Gamma, V_\beta, \{z_n\}_{n=1}^N, \{g_n\}_{n=1}^N, \lambda) \\ & \qquad = \prod_{n=1}^N p(y_n|\beta_n) p(\beta_n|g_n, \Gamma, V_\beta) p(w_n|z_n, \lambda) p(z_n|g_n). \end{aligned} \tag{9}$$

Figure 4 illustrates the proposed integrated hierarchical Bayesian choice and GoM model. From the DAG we can see that the proposed model is a three-level model where only the categorical responses w_n and choices y_n for each respondent are observed. The homogeneous profiles λ account for the interaction or co-occurrence among items and

Figure 4: Integrated Hierarchical Bayesian Choice and GoM Model



provide for the dimension reduction we need to use this collection of discrete data as covariates in the model of preference heterogeneity.

Figure 4 combines the DAGs in Figure 1 and Figure 3 to illustrate that the membership vector g_n serves as the link between the choice and GoM components of the model. Thus g_n is informed by both the categorical responses w_n and the chosen alternatives y_n. The proposed model is more complete than a model where g_n is estimated separately from choice and a model where the mean of the latent profile assignments z_n serve as the link since using parameters estimated together in the integrated model allows us to properly account for uncertainty. Because g_n is identified when informed by w_n alone in the GoM model, g_n is also identified in the proposed model when identified by both w_n and y_n.

Following Figure 4, the joint posterior distribution of the proposed model is:

$$p(\{\beta_n\}_{n=1}^N, \Gamma, V_\beta, \{z_n\}_{n=1}^N, \{g_n\}_{n=1}^N, \lambda | \{y_n\}_{n=1}^N, \overline{\Gamma}, A, \nu, V, \{w_n\}_{n=1}^N, \alpha, \tau)$$

$$\propto \left[\prod_{n=1}^N p(y_n|\beta_n) p(\beta_n|g_n, \Gamma, V_\beta) p(w_n|z_n, \lambda) p(z_n|g_n) p(g_n|\alpha) \right] p(\Gamma|V_\beta, \bar{\Gamma}, A) p(V_\beta|\nu, V) p(\lambda|\tau) \tag{10}$$

where $\prod_{n=1}^{N} p(y_n|\beta_n)p(\beta_n|g_n, \Gamma, V_\beta)p(w_n|z_n, \lambda)p(z_n|g_n)$ is the likelihood and $\prod_{n=1}^{N} p(g_n|\alpha)$, $p(\Gamma|V_\beta, \bar{\Gamma}, A)$, $p(V_\beta|\nu, V)$, and $p(\lambda|\tau)$ are the priors. A complete list of the variables in Equation (10) are detailed in Table 1.

Table 1: Variables in a Hierarchical Bayesian Choice Model with a GoM Model

Choice Variables	Description
N	number of respondents
H	number of choice tasks for each respondent n
P	number of alternatives in each choice task
L	number of attribute levels in each choice task
y_n	H-dim vector of choices for respondent n
β_n	L-dim vector of part-worths for respondent n
Γ	$K \times L$ matrix representing the mean of the random effects distribution of heterogeneity
V_β	$L \times L$ covariance matrix of the random effects distribution of heterogeneity
GoM Variables	**Description**
K	number of profiles
J	number of categorical questions
n_j	number of categorical responses for question j
w_n	J-dim vector of respondent n's categorical responses
z_n	J-dim vector of respondent n's profile assignments
g_n	K-dim membership vector for respondent n
λ	collection of probability distributions $\lambda_{j,k}$ over the n_j response options for each question j and profile k

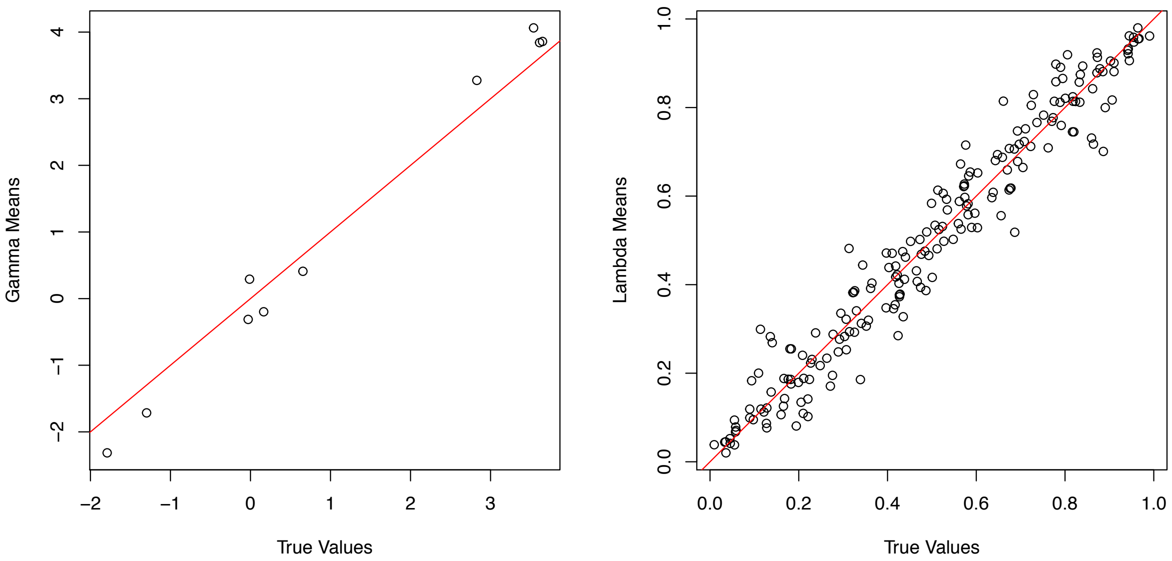
3 Simulation Experiments

We ran a set of simulation experiments to validate the proposed model, demonstrate empirical identification of the proposed and competing models, and discuss the boundary conditions of the proposed model in terms of extreme profile membership.

3.1 Model Validation

We validate our proposed model by generating data where $K = 2$, $N = 200$, $J = 13$, $n_j = 2$ for all J, $H = 50$, $P = 4$, and $L = 5$ and recovering parameter values. Each true parameter value was within or near the bounds of a 95% credible interval. We display the aggregate-level posterior means of Γ and λ in Figure 5. The posterior means line up along the diagonal, indicating parameter recovery. Note that the λ estimates are constrained to be within the $0-1$ bounds.

Figure 5: Model Validation



We employ a Markov chain Monte Carlo estimation procedure with both random-walk Metropolis-Hastings and Gibbs steps. A Gibbs sampler similar to that detailed in Erosheva et al. (2007) is used to estimate the GoM portion of the proposed model. However, since the membership vector g_n is included in the distribution of preference heterogeneity, we use a random-walk Metropolis-Hastings algorithm to estimate g_n's that are predictive of the part-worths. The remaining choice model portions of the proposed model utilize standard estimation methods. Details on generating data and estimation are provided in Appendix A.

3.2 Empirical Identification

We demonstrate empirical identification of the proposed and competing models using simulation experiments. Each of the models differs in terms of the upper-level structure of the hierarchical Bayesian choice model. The three models are a model with observed binary covariates (i.e., the binary covariates model), a model with the membership vector from an integrated GoM model as covariates (i.e., the proposed or membership vector model), and a model with the factor scores from an integrated factor analysis as covariates (i.e., the factor scores model).

The upper level in a hierarchical Bayesian choice model is a multivariate regression, as demonstrated in Equation (2). The difference between the proposed and competing models is in the treatment of the observed covariates. The binary covariates model assumes that the covariates are exogenous predictors of the part-worths or β_n's. Both the proposed membership vector and the competing factor scores models assume that the covariates are endogenous, with measurement indicators of a latent structure serving as the true drivers of the β_n's. The membership vector and the factor scores model differ with respect to assumptions concerning this latent structure, as detailed previously.

We generate choice and categorical response data according to each of these three models. This allows us to estimate each model given each of the datasets. Note that we do not generate data from an intercept model, which is equivalent to specifying a standard multivariate normal distribution for the β_n's. Fitting a latent variable model such as the membership vector model or factor scores model is likely to overfit such data since, in the absence of informative covariates, the prior variance of the β_n's is unbounded. Table 2 provides details on generating data according to each of the three models.

We calculate two fit statistics for each of the models on each of the datasets. The first is the Newton-Raftery approximation of the log marginal density (LMD) (Newton and Raftery, 1994), a standard Bayesian measure for model fit. The second is the average hit probability, a standard measure of predictive fit. A hit probability is the average posterior

Table 2: Parameters for Data Generation

	Binary	Membership	Factor
Parameters	Covariates	Vector	Scores
N number of respondents	400	400	400
N^* number of hold-out respondents	400	400	400
H number of choice tasks for each n	15	15	15
P number of alternatives for each h	4	4	4
L number of attribute levels for each p	12	12	12
J number of categorical questions	30	30	30
n_j number of multinomial outcomes of j	2	2	2
K number of latent profiles or factors	*NA*	3	3

probability of a set of observed choices given a specific model. The hit probability is averaged over a set of hold-out respondents N^*, observations H, and post-burn-in MCMC draws R. The hit probability for a given model M is:

$$\text{HP}(M) = \frac{1}{N^*} \sum_{n^*=1}^{N^*} \left[\frac{1}{H} \sum_{h=1}^{H} \left(\frac{1}{R} \sum_{r=1}^{R} Pr(j|\beta^M_{n^*,r}, X_h)_{n^*} \right) \right] \tag{11}$$

where j is the observed choice from the design matrix X_h for each observed choice task H and $\beta^M_{n^*,r}$ are respondent n^*'s estimated coefficients for each of the R post-burn-in MCMC draws for model M. These $\beta^M_{n^*,r}$ are drawn from the distribution of heterogeneity $N(\Gamma_r^{M\prime} z_n^M, V_{\beta,r}^M)$ for the binary covariates model. However, the hold-out sample covariates for the proposed membership vector model $g^M_{n^*}$ and the competing factor scores model $\zeta^M_{n^*}$ do not have distributions of heterogeneity to draw from.

To address this for the membership vector model, we generate initial profile assignments $z^M_{n^*,j} \sim \text{Multinomial}(g_{n^*})$ for each of the respondents' J questions in the hold-out sample where $g^M_{n^*} \sim \text{Dirichlet}(\alpha^*)$. The profile assignments $z^M_{n^*,j}$ and membership vectors $g^M_{n^*,r}$ are then successively updated for each of the post-burn-in MCMC draws using the hold-out sample response data $w_{n^*,j}$, $\lambda^M_{j,k,r}$, and α^*:

$$z^M_{n^*,j,r} = \arg\max_k \left(\text{Multinomial}(p^M_1, \ldots, p^M_K)\right), \text{ where } p^M_k \propto g^M_{n^*,k,r}\lambda^M_{j,k,r}(w_{n^*,j}),$$

$$g^M_{n^*,r} \sim \text{Dirichlet}(z^M_{n^*,j,r} + \alpha^*).$$

We then draw $\beta^M_{n^*,r} \sim N(\Gamma^{M\prime}_r g^M_{n^*,r}, V^M_{\beta,r})$ for the hold-out sample. The concentration parameter α^* for the Dirichlet prior is set to be uninformative by using the α^* that minimizes the Kullback-Leibler divergence between the distributions of heterogeneity for the in-sample respondents β^M_n and the hold-out sample $\beta^M_{n^*}$ respondents. For the factor scores model, we first generate draws of $\zeta^M_{n^*,r}$ from the sample response data $w_{n^*,j}$ and across-subject factor loadings $\lambda^M_{j,k,r}$ using standard conjugate results from Bayesian factor modeling (Lee, 2007). We then generate $\beta^M_{n^*,r} \sim N(\Gamma^{M\prime}_r \zeta^M_{n^*,r}, \Sigma^M_{\Gamma,r})$ for the hold-out sample.

Table 3: Empirical Identification

	Data		
	Binary Covariates	Membership Vector	Factor Scores
Model	*LMD* | *Hit Prob.*	*LMD* | *Hit Prob.*	*LMD* | *Hit Prob.*
Binary Covariates	**-606** | **0.895**	-586 | 0.728	-2422 | 0.600
Membership Vector	-783 | 0.700	**-519** | **0.800**	-2410 | 0.480
Factor Scores	-610 | 0.644	-535 | 0.656	**-2288** | **0.612**

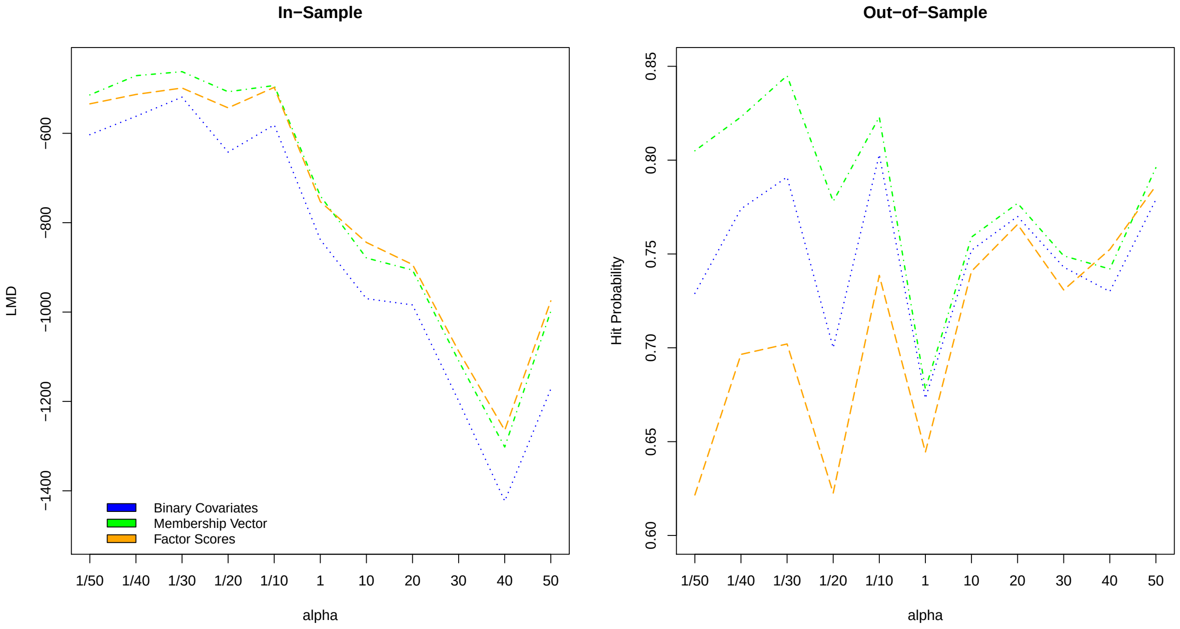
The simulation experiment results are included in Table 3, where the data-generating model in each column is in bold. Note that in order for the in-sample LMD to be comparable across models, it is obtained without considering the likelihood contribution of endogenous covariates in the case of the membership vector model and the factor scores model. The results demonstrate that we can recover the true model from the data. That is, we find for all datasets that the true model is the best-fitting model, in terms of both in-sample and out-of-sample fit. We conclude that choice data are sufficient to identify the true upper-level model.

3.3 Boundary Conditions

We also explore the boundary conditions of the proposed model through simulation experiments. For this analysis, we focus on identifying the proposed membership vector model depending on the distribution of respondent-level membership vectors g_n. Given a value of the concentration parameter α for the Dirichlet prior, different distributions of g_n can be generated. Higher values of α lead to a more uniform distribution across the profiles, implying that each respondent is close to an equally weighted combination of the K extreme profiles. Lower values of α imply weights concentrated more heavily on only one of the K profiles, or more extreme profile membership. To evaluate the role of α, we generate 11 datasets using the proposed membership vector with differences resulting from varying α values between $\frac{1}{50}$ and 50, or from more extreme to less extreme profile membership. The resulting g_n in each dataset are then used to generate choice data.

Figure 6 plots fit for all three models for all 11 datasets. Again, we compare the models based on in-sample fit (LMD) and out-of-sample fit (average hit probabilities) for the choice data only. With respect to out-of-sample predictions, results from Figure 6 suggest that the true model can be identified across a wide range of values for $\alpha = [\frac{1}{50}, 30]$ as evidenced by higher hit probabilities compared to all other models. When in-sample fit is considered, the range in which the proposed membership vector model performs best is smaller ($\alpha \leq \frac{1}{10}$). We conclude that the proposed model does especially well relative to the competing models when respondents have more extreme profile membership (i.e., are located toward the corners of the convex hull defined by the K profiles). In other words, the proposed model performs best when respondents more closely resemble one of the archetypal extreme profiles. Interestingly, the alternative latent variable model (i.e., the factor scores model) performs worst among all the models in this situation.

Figure 6: Boundary Conditions Based on Extreme Profile Membership



4 Empirical Applications

We use data from two surveys of preferences regarding robotic vacuums and smartphones using national samples in the United States and Germany, respectively. While the robotic vacuums data come from an emerging market, the smartphones data represent a well-established market, providing us a broad test for our proposed model. For the robotic vacuums data, a total of 332 respondents were carefully screened to ensure that the product options under consideration were relevant to them. In particular, qualified respondents had to own a robotic vacuum, currently be shopping for their first robotic vacuum, or might consider a robotic vacuum sometime in the next five years. For the smartphones data, a total of 147 respondents were similarly screened to ensure they were in the market for a new smartphone.

Before the conjoint experiment, respondents were asked to detail why the product was relevant to them or anyone in their household. For the robotic vacuums data, respondents selected from a list of 11 statements on cleaning that robotic vacuums might help address and a list of 7 statements that described problems with robotic vacuums. The combined list of 18 statements regarding cleaning and robotic vacuums is provided in Table 4. For the smartphones data, respondents selected from a list of 53 statements on smartphones that described their interests and usage. A subset of the 53 statements regarding smartphones is provided in Table 5. Thus our discrete data consist of two possible categories (i.e., $n_j = 2$) for all $J = 18$ or $J = 53$ where not selecting an item is coded as a 0 and selecting an item is coded as a 1.

Table 4: Statements on Cleaning and Robotic Vacuums

No.	Item
1	I enjoy coming home to a clean house.
2	I don't feel relaxed when I know my home isn't clean.
3	I worry about pet hair and dander in the home.
4	I have trouble keeping the floor beneath my furniture clean.
5	I worry about germs and dirt on my floor and carpet.
6	I get anxious about having guests when my home is dirty.
7	I don't like going to someone's home that is dirty.
8	I don't like touching dirty things.
9	I don't spend much time cleaning.
10	I spend over two hours per week cleaning.
11	I have a cleaning person who cleans for me.
12	Robotic vacuums are too expensive.
13	Robotic vacuums are too complicated to program, set up, and operate.
14	Robotic vacuums often need to be "rescued" because they get stuck.
15	Robotic vacuums need to have their trash containers changed too often.
16	Robotic vacuums don't do a good enough job cleaning the floor and carpet.
17	Robotic vacuums don't spend enough time on really dirty spots on the floor.
18	Robotic vacuums scare household pets.

Standard models using this discrete data as observed covariates in the random effects distribution of heterogeneity don't have a practical way to include interactions, even though interactions should be expected. For example, in the robotic vacuums data, we would expect that respondents who select statement 5 "I worry about germs and dirt on

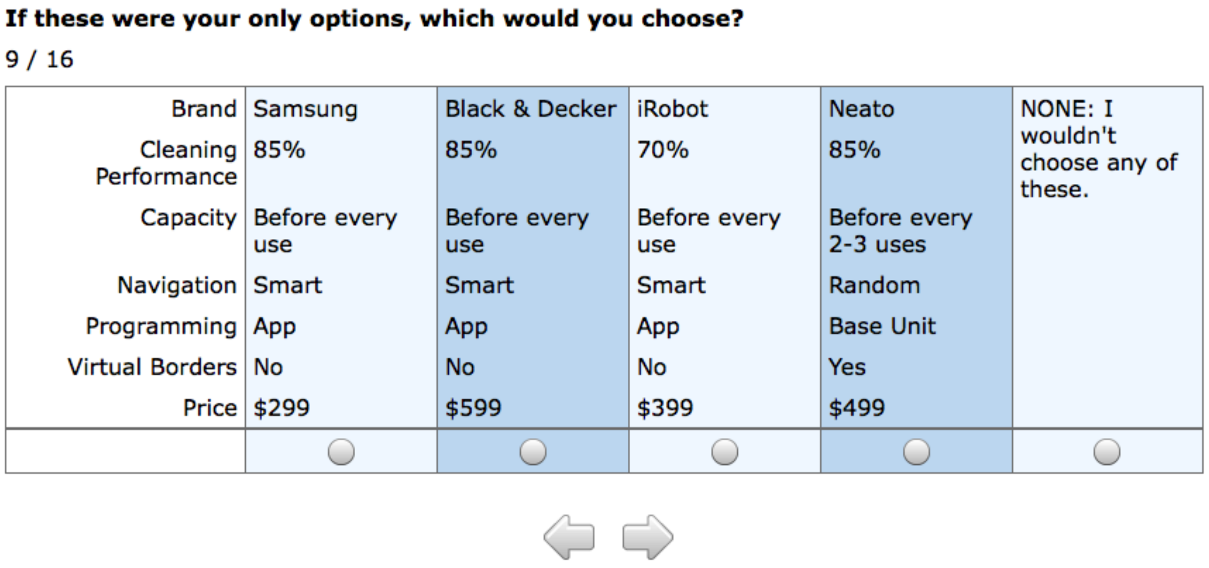
Table 5: Statements on Smartphone Interest and Usagge

No.	Item
1	The security of the OS on my SP is very important to me.
2	The apps on my SP only run with the newest OS.
3	A more recent OS is worth a higher price.
4	A more recent OS shows that a SP is up-to-date.
5	It's always useful to have a more recent OS.
6	I don't care about the OS on my SP.
7	I want to use my SP to make payments.
8	My SP should be handy.
9	I think a smaller SP is more useful.
10	A small display is important to me.
⋮	⋮
43	I need large memory on my SP.
44	I listen to music a lot on my SP.
45	I stream music and video on my SP, I don't store it on my SP.
46	I like to watch HD movies on my SP.
47	I don't feel that display resolution makes a difference.
48	My SP is always switched on, never off.
49	During the night, I always switch my SP off.
50	I switch my SP to vibrate in the night.
51	I switch my SP to vibrate only when necessary.
52	I like to surf the web on my SP.
53	I use QR codes with my SP.

my floor and carpet" also select statement 10 "I spend over two hours per week cleaning" and that this interaction would have an impact on explaining preferences in the random effects distribution of heterogeneity. However, if we were to include two-way interactions, we would add an additional 153 covariates, to say nothing of the dimensionality introduced by higher-level interactions.

After selecting from applicable statements on cleaning and robotic vacuums or on smartphone interest and usage, respondents proceeded through a series of choice tasks where they were asked to select which of a given number of product alternatives they most preferred. For the robotic vacuums data, this set of alternatives included an outside option to not select any of the given alternatives. Each alternative was composed of

Figure 7: Example Robotic Vacuums Choice Task



If these were your only options, which would you choose?

9 / 16

Brand	Samsung	Black & Decker	iRobot	Neato	NONE: I wouldn't choose any of these.
Cleaning Performance	85%	85%	70%	85%	
Capacity	Before every use	Before every use	Before every use	Before every 2-3 uses	
Navigation	Smart	Smart	Smart	Random	
Programming	App	App	App	Base Unit	
Virtual Borders	No	No	No	Yes	
Price	$299	$599	$399	$499	
	○	○	○	○	○

separate attributes. Figure 7 is a screenshot of one of these choice tasks from the robotic vacuums data. The estimable attribute levels, excluding the reference levels in red, are included in Tables 6 and 7.

Table 6: Robotic Vacuums Attribute Levels

Attributes	Levels				
Brand	Outside Option	Neato	iRobot	Samsung	Black & Decker
Performance	70%	85%			
Capacity	Every use	Every 2-3 uses			
Navigation	Random	Smart			
Programming	Base unit	App			
Virtual Borders	No	Yes			
Price	$299	$399	$499	$599	

For the robotic vacuums data, we see from Table 6 that the attributes are defined in terms of brand, price, and different features, including the vacuum's performance (i.e., what percentage of dirt and debris it picks up), capacity (i.e., how often it needs to be emptied), the type of navigation (i.e., does it change directions by just bumping into

Table 7: Smartphones Attribute Levels

Attribute	Levels			
Display Size	4 in.	4.7 in.	5 in.	5.5 in.
Display Resolution	Standard	HD		
Camera (front)	4 MP	6 MP	8 MP	12 MP
Memory	4 GB	16 GB	32 GB	64 GB
Price	$400	$600	$800	

things or is it "smart" and able to scan and determine an optimal path), where it can be programmed, and whether or not virtual borders can be set to keep the robotic vacuum away from certain areas of the home. For the smartphones data, we see from Table 7 that the attributes are defined in terms of functional attributes, including display size (i.e., the effective size of the phone), display resolution, camera quality for the front camera, available memory, and price. A summary of the datasets using model notation is provided in Table 8.

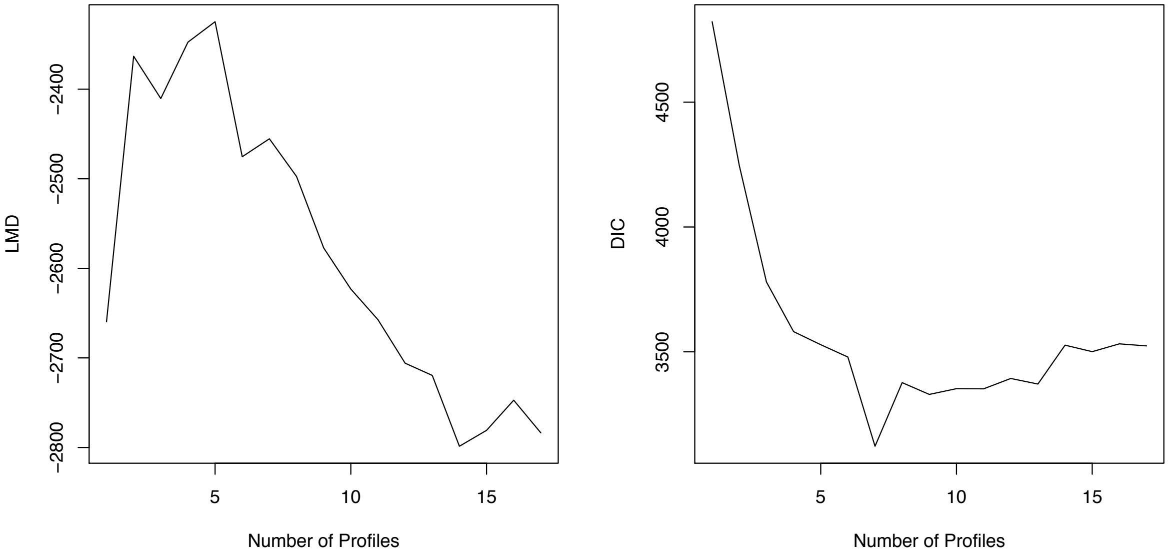
Table 8: Data Summary

Variables	Robotic Vacuums	Smartphones
N total number of respondents	332	147
H number of choice tasks for each respondent n	16	17
P number of alternatives in each choice task	5	3
L number of attribute levels in each choice task	12	12
J number of categorical questions	18	53
n_j number of categorical responses for each question j	2	2

5 Results

We report the results of four models. The intercept model only includes an intercept in the upper level model (i.e., $\beta_n = \gamma + \xi_n$) and serves as a baseline. The binary covariates model includes all 18 or 53 dummy-coded statements from Tables 4 or 5, respectively, as covariates in the upper level model (i.e., $\beta_n = \Gamma' z_n + \xi_n$) and represents the typical way

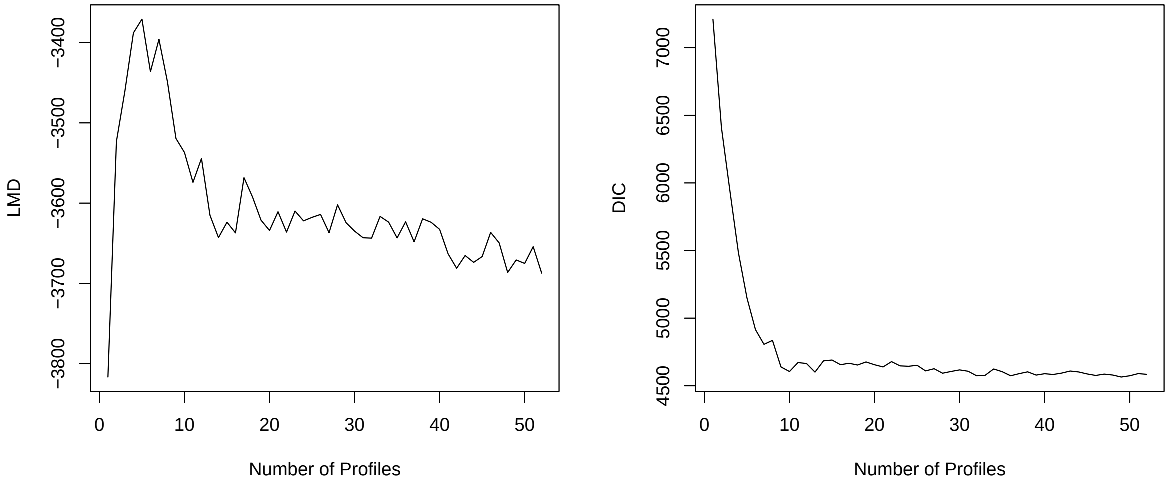
Figure 8: Selecting K for the Robotic Vacuums Dataset



these discrete covariates would be used in practice. The membership vector model is our proposed model, which uses the membership vectors g_n from the integrated choice and GoM model as covariates for $K = 5$ profiles (i.e., $\beta_n = \Gamma' g_n + \xi_n$). Finally, the factor scores model utilizes an alternative assumption about the latent structure and uses the factor scores ζ_n from an integrated choice and factor analysis model as covariates for $K = 9$ factors (i.e., $\beta_n = \Gamma' \zeta_n + \xi_n$).

The number of profiles K is determined by the analyst. We ran an isolated GoM model on the statements in Tables 4 and 5, respectively, and compared two measures of fit, following the review on model selection criteria by Joutard et al. (2007). The first is the Newton-Raftery approximation of the log marginal density (LMD) (Newton and Raftery, 1994). The second is the deviance information criterion (DIC) Spiegelhalter et al. (2002). Values closer to zero indicate improvement in fit for both measures. Figure 8 includes charts for the values of both LMD and DIC for models with $K = 2$ to $K = 18$ for the robotic vacuums data. According to the LMD, $K = 5$ is best. According to the DIC, $K = 7$ is best. With the range of possible models narrowed, we ran the proposed

Figure 9: Selecting K for the Smartphones Dataset



membership vector model for $K = 5$ to $K = 7$. Comparing results to find profiles that are sufficiently differentiated and non-repeating, the model with $K = 5$ was deemed best. The same process was repeated for the smartphones data. Figure 9 includes LMD and DIC for models with $K = 2$ to $K = 53$. According to the LMD, $K = 5$ is best. According to the DIC, $K = 10$ is best. We ran the membership vector model for $K = 5$ to $K = 10$, compared results and profiles, and chose $K = 9$. The number of factors K for both datasets was selected using LMD alone.

In-sample model fit is measured using LMD. Out-of-sample model fit is measured using both LMD and average hit probability from a hold-out sample of 20% of respondents from each dataset, respectively. We follow the same procedure to compute the average hit probability as specified in Section 3.2. We ran each model for 50,000 iterations, saving every 50th draw, and used the final 20,000 iterations for inference. We checked for but found no substantial evidence of label switching.

Table 9 includes all three measures of model fit for both datasets with the best fit for each measure in bold. We can see that across all measures, the proposed membership

Table 9: Model Fit

	In-Sample	Out-of-Sample	
Robotic Vacuum Models	*LMD*	*LMD*	*Hit Prob.*
Intercept ($\beta_n = \gamma + \xi_n$)	-1845	-4862	0.371
Binary Covariates ($\beta_n = \Gamma' z_n + \xi_n$)	-1904	-5450	0.392
Membership Vector ($\beta_n = \Gamma' g_n + \xi_n$)	**-1806**	**-3716**	**0.449**
Factor Scores ($\beta_n = \Gamma' \zeta_n + \xi_n$)	-1854	-4360	0.376
Smartphone Models	*LMD*	*LMD*	*Hit Prob.*
Intercept ($\beta_n = \gamma + \xi_n$)	-869	-1453	0.472
Binary Covariates ($\beta_n = \Gamma' z_n + \xi_n$)	-859	-2092	0.449
Membership Vector ($\beta_n = \Gamma' g_n + \xi_n$)	**-744**	**-1014**	**0.526**
Factor Scores ($\beta_n = \Gamma' \zeta_n + \xi_n$)	-1103	-1195	0.508

vector model using covariates uncovered with an integrated choice and mixed membership model have more explanatory and predictive power. Again, note that we compute the in-sample LMD based on choices only, which allows us to compare models that treat covariates as either endogenous, in the case of the integrated models, or exogenous, in the case of the baseline models. This is different from the out-of-sample LMD, which is based on simulating β_{n^*}'s for hold-out sample respondents from the hierarchical prior, using hold-out sample respondents' categorical response data and cross-sectional parameter estimates. Since the fit of the data is driven entirely the distribution of heterogeneity, model choice and the question how the β_{n^*}'s are being generated becomes critical. The out-of-sample LMD, which considers the full distribution of hold-out probabilities across tasks and respondents, together with predictive fit evident in the average hit probabilities, provide strong evidence that the mixed membership approach is preferred for modeling preference heterogeneity. In short, being able to adequately capture the full distribution of preferences is essential to identifying the correct model.

Another model and measure of fit were also considered. The model included interactions directly. However, in running this alternative model, problems manifested themselves with only two-way interactions. First, the flexibility of the model induced by including so many covariates clearly allowed for overfitting. As we increased the number of iterations in the Markov chain, we continued to see an improvement in in-sample fit with no change in out-of-sample fit and no sign of convergence. Second, the number of interactions would make interpretation infeasible. For these reasons we don't report the results of this model. The alternative measure of model fit was the Bayesian equivalent to the adjusted R^2 (Gelman and Pardoe, 2006). However, this proved inappropriate as the response in the upper-level model is latent.

The proposed model also improves inference regarding the drivers of preference heterogeneity. To illustrate, we look at the model output for the robotic vacuums data. First, let's consider the posterior means of Γ from the binary covariates model. Table 10 displays the complete Γ matrix. The attribute levels are on the left and each column in the matrix is associated with the intercept or one of the statements from Table 4. The posterior means highlighted in red and green are more than two standard deviations below and above zero, respectively. This matrix should inform a marketer concerning the drivers of preference for promotion and targeting strategies. However, making sense of the significant values or considering how these items may interact is cumbersome.

For example, we can use Table 10 to infer that respondents who are concerned about germs and dirt (i.e., statement 5 "I worry about germs and dirt on my floor and carpet") prefer any brand of robotic vacuum relative to the outside good while not being concerned about getting the highest level of performance. We might expect this is because they are cleaning frequently (e.g., statement 10 "I spend over two hours per week cleaning") and having a robotic vacuum is simply one part of a larger cleaning solution. Without a way to properly account for interactions, we aren't able to understand these more detailed explanations of preference heterogeneity.

The proposed model accounts for such interactions by identifying differentiated respondent profiles. Table 11 details the profiles as described by the estimates of $\lambda_{j,k}(1)$. Recall that because $n_j = 2$ for all J, the profiles can be presented in terms of $\lambda_{j,k}(1)$

Table 10: Robotic Vacuums Binary Covariates Model Γ Estimates

Attribute Levels	Int.	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
Neato	1.52	0.03	0.22	-0.05	-0.31	2.19	-1.36	-0.78	-2.18	-0.74	-0.29	3.75	-1.03	1.65	-0.83	-0.66	0.56	-1.25	-0.17
iRobot	2.76	0.26	0.79	-0.31	-0.50	2.41	-1.73	-1.65	-2.53	-0.26	-0.26	3.27	-0.84	1.80	-0.95	-0.24	0.52	-0.69	0.22
Samsung	2.96	-0.46	0.42	-0.19	-0.23	2.27	-1.53	-1.02	-1.82	-0.76	-0.39	3.69	-0.95	1.37	-1.38	-0.83	1.00	-0.77	-0.27
Black & Decker	2.69	-0.42	0.21	-0.53	-0.10	1.99	-0.86	-1.70	-1.85	-0.55	-0.05	4.23	-1.81	2.77	-0.60	-0.58	0.82	-1.18	-0.07
Performance: 85%	0.91	0.43	0.29	0.47	0.68	-1.17	0.10	0.93	0.68	-0.25	0.21	-0.43	0.92	-0.31	0.83	0.34	0.92	-0.15	0.61
Capacity: Every 2-3 uses	0.47	-0.22	0.24	0.09	0.26	-0.02	0.21	0.07	0.03	-0.36	-0.15	-0.07	0.02	0.19	-0.10	0.14	0.43	-0.19	-0.14
Smart Navigation	0.25	0.47	-0.32	0.30	0.14	0.21	0.40	0.18	0.06	0.21	0.11	-0.34	-0.21	-0.14	-0.14	-0.22	-0.74	0.31	-0.38
App Programming	-0.50	0.14	0.48	-0.07	-0.12	0.57	-0.07	-0.10	-0.12	0.63	0.22	-1.15	0.05	0.34	-0.29	0.16	-0.38	-0.04	-0.21
Virtual Borders	1.01	-0.22	-0.29	0.23	0.23	-0.10	0.07	-0.40	0.12	0.02	-0.03	-0.18	-0.08	-1.18	0.30	0.21	-0.63	0.04	0.44
$399	-1.10	-0.22	0.07	-0.36	-0.40	0.43	0.82	-1.11	0.73	-1.43	0.03	1.90	-1.16	-0.71	-0.07	0.65	-0.22	-0.28	-0.41
$499	-3.04	-0.87	0.37	0.13	-1.29	1.22	1.14	-1.97	1.02	-2.54	-0.30	3.22	-2.93	-1.48	0.07	1.57	-0.87	0.33	-0.60
$599	-4.83	-0.96	0.81	0.61	-1.41	0.77	0.76	-1.75	0.90	-2.69	-0.29	4.28	-3.41	-2.56	-0.71	2.04	-1.64	0.16	-0.43

since $\lambda_{j,k}(0)$ is simply its complement. Because respondents were qualified by owning or being interested in a robotic vacuum, it isn't surprising that every profile as described in Table 11 has statement 1 "I enjoy coming home to a clean house" occurring with high probability. Profile 1 is differentiated from the other models by statement 12 "Robotic vacuums are too expensive," statement 10 "I spend over two hours per week cleaning," and statement 17 "Robotic vacuums don't spend enough time on the really dirty spots on the floor" occurring with high probability. We name this profile "Constantly Cleaning."

Profile 2, like profile 1, has statement 12 "Robotic vacuums are too expensive" occurring with high probability, but is further differentiated by statement 9 "I don't spend much time cleaning" and statement 4 "I have trouble keeping the floor beneath my furniture clean." We name this profile "Difficulty with or Little Cleaning." Profile 3 is differentiated by statement 6 "I get anxious about having guests when my home is dirty," statement 7 "I don't like going to someone's home that is dirty," and statement 2 "I don't feel relaxed when I know my home isn't clean" occurring with high probability. We name this profile "Anxious about Cleanliness."

Profile 4 is differentiated overwhelmingly by statement 1 "I enjoy coming home to a clean house" with a weight of 0.66. The next statement co-occurs with a weight of 0.24. We name this profile "Prioritizes a Clean House." Finally, profile 5 is differentiated by statement 5 "I worry about germs and dirt on my floor and carpet" and statement 3 "I worry about pet hair and dander in the home." We name this profile "Specific Cleaning Concerns." The profile names for the robotic vacuums data are summarized in Table 12.

Table 13 displays the matrix of estimated coefficients Γ that maps variability in the membership vectors to variability in the part-worths. Again, the posterior means highlighted in red and green are more than two standard deviations below and above zero, respectively. Recall that these profiles are archetypal or extreme where each individual is a partial member in each profile as defined by the weights of their membership vector g_n. Because of this, we have rescaled the coefficients in this matrix to represent a 10%

Table 11: Robotic Vacuums Membership Vector Model $\lambda_{j,k}(1)$ Estimates

No.	Statements	$\lambda_{j,1}(1)$	$\lambda_{j,2}(1)$	$\lambda_{j,3}(1)$	$\lambda_{j,4}(1)$	$\lambda_{j,5}(1)$
1	I enjoy coming home to a clean house.	0.68	0.90	0.95	0.66	0.96
2	I don't feel relaxed when I know my home isn't clean.	0.27	0.41	0.86	0.24	0.94
3	I worry about pet hair and dander in the home.	0.12	0.44	0.69	0.09	0.89
4	I have trouble keeping the floor beneath my furniture clean.	0.12	0.62	0.71	0.09	0.78
5	I worry about germs and dirt on my floor and carpet.	0.20	0.41	0.77	0.16	0.95
6	I get anxious about having guests when my home is dirty.	0.26	0.58	0.90	0.14	0.95
7	I don't like going to someone's home that is dirty.	0.16	0.55	0.89	0.09	0.92
8	I don't like touching dirty things.	0.13	0.17	0.84	0.07	0.86
9	I don't spend much time cleaning.	0.26	0.53	0.09	0.10	0.11
10	I spend over two hours per week cleaning.	0.35	0.41	0.76	0.19	0.89
11	I have a cleaning person who cleans for me.	0.04	0.05	0.07	0.06	0.18
12	Robotic vacuums are too expensive.	0.59	0.92	0.72	0.24	0.33
13	Robotic vacuums are too complicated to program, set up, and operate.	0.05	0.26	0.14	0.08	0.13
14	Robotic vacuums often need to be "rescued" because they get stuck.	0.21	0.38	0.78	0.11	0.38
15	Robotic vacuums need to have their trash containers changed too often.	0.22	0.12	0.44	0.18	0.37
16	Robotic vacuums don't do a good enough job cleaning the floor and carpet.	0.08	0.24	0.44	0.20	0.18
17	Robotic vacuums don't spend enough time on the really dirty spots on the floor.	0.28	0.16	0.21	0.20	0.10
18	Robotic vacuums scare household pets.	0.13	0.32	0.32	0.11	0.37

Table 12: Robotic Vacuums Profile Names

No.	Robotic Vacuums
1	Constantly Cleaning
2	Difficulty with or Little Cleaning
3	Anxious about Cleanliness
4	Prioritizes a Clean House
5	Specific Cleaning Concerns

Table 13: Robotic Vacuums Membership Vector Model Γ Estimates

Attribute Levels	P1	P2	P3	P4	P5
Neato	-1.63	0.65	-1.87	1.84	0.67
iRobot	-1.12	0.78	-1.94	1.93	0.70
Samsung	-1.19	0.56	-2.03	1.86	0.89
Black & Decker	-1.47	0.78	-2.01	1.86	0.80
Cleaning Performance: 85%	-0.15	0.13	2.08	-0.04	-0.13
Capacity: Every 2-3 uses	-0.13	0	0.19	0.13	0.12
Smart Navigation	0.12	0.02	0.21	-0.06	0.12
App Programming	0.03	-0.07	-0.09	0	0.09
Virtual Borders	0.46	-0.05	0.07	-0.10	-0.02
$399	0.27	-1.82	0.01	0.13	0.19
$499	0.32	-4.08	-0.21	0.20	0.36
$599	0.35	-5.33	-0.70	0.16	0.58

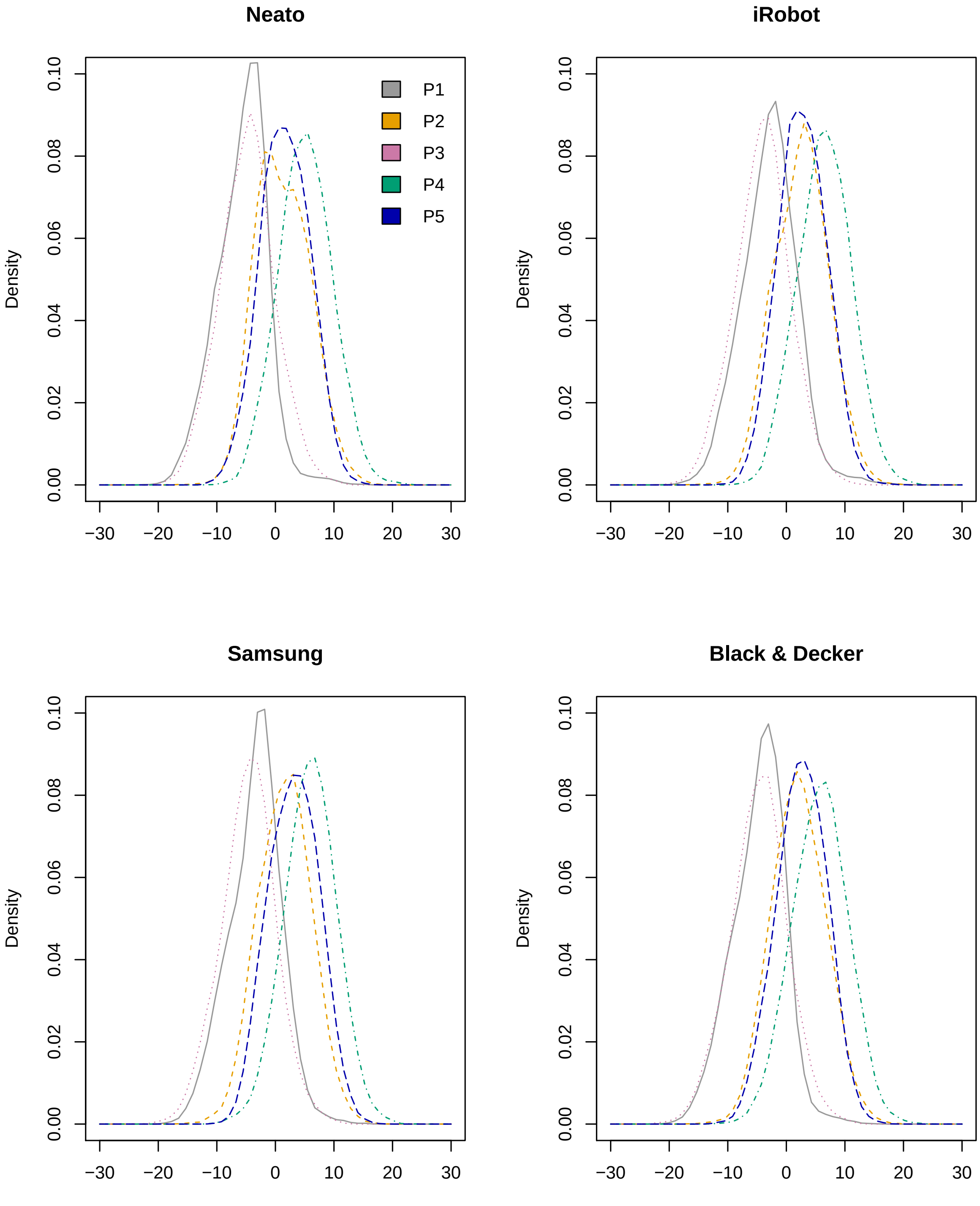
increase in partial membership for each profile. For example, if an individual increase by 10% in profile 2 "Difficulty with or Little Cleaning," then their part-worth for the highest price point would decrease by 5.33.

As with Table 10, the matrix in Table 13 should inform a marketer concerning the drivers of preference for promotion and targeting strategies. However, using the proposed model, we are able to explain preferences in terms of the extreme profiles. For example, profile 5, "Specific Cleaning Concerns" includes statements 5 "I worry about germs and dirt on my floor and carpet" and 10 "I spend over two hours per week cleaning" with high probability. With this profile we can answer what was only suggested from Table 10, that the more an individual is aligned with this profile, the more they prefer a high-capacity robotic vacuum relative to the outside good. In other words, since they are cleaning often, they want a robotic vacuum with high capacity in order to effectively assist.

We can better inform targeting and promotion strategies using the proposed model. We can use the estimate of Γ as a roadmap for targeting by matching what respondents prefer with a more detailed explanation of what is driving those preferences. For example, for consumers who are primarily in profile 2 "Difficulty with or Little Cleaning," we know that price promotions should be especially effective since they have a need for robotic vacuums but are very price sensitive. The dimension-reduction provided by employing an integrated GoM model makes this plausible with the 12×5 Γ matrix in Table 13 compared with a similar task using the 12×19 Γ matrix in Table 10 from the alternative model or an even larger Γ matrix that includes interactions directly.

Accounting for the co-occurrence or interactions among items is akin to segmenting the market. The blocks of significant attribute level coefficients in Table 13 are reminiscent of such segmentation solutions. Unlike mixture models, which are typical in clustering applications, where a respondent is assigned to a single category, mixed membership models like the GoM allow for the more realistic description of each respondent being a partial member of each profile, with weights determined heterogeneously. This partial membership allows for the distribution of preference to move into the extremes. To illustrate, Figure 10 provides the marginal posterior distributions of heterogeneity for the robotic vacuum brands in the study. The densities in each plot correspond to the five profiles from the proposed model where, for convenience in visualizing this difference, each respondent has been assigned to the profile with the largest mean posterior weight $g_{n,k}$. The pattern of the densities demonstrate how the proposed model is able to capture extremes in the distribution of heterogeneity. It is this accounting for extremes in preference heterogeneity that improves the model fit and our ability to conduct inference.

Figure 10: Robotic Vacuums Marginal Posterior Distributions of Heterogeneity



6 Discussion

In this paper we show that modeling interactions or co-occurrence among discrete multivariate data does more to explain consumer preferences than the discrete covariates on their own. This is accomplished by integrating a grade of membership model, part of the class of mixed membership models, with a choice model to estimate membership vectors for use in a hierarchical Bayesian random effects distribution of heterogeneity.

Choice modeling remains an essential fixture of marketing research. However, finding covariates that are explanatory of preference heterogeneity has proven difficult. Our proposed model provides a novel way to account for interactions, and provide dimension reduction, for survey data that explain variation in part-worth utilities. The empirical applications utilize typical survey response data in both an emerging market and an established category to demonstrate the use of the proposed model. However, with growing access to unstructured collections of discrete data, we see this approach as an important step to utilizing such data, including text, to improve choice modeling.

Latent Dirichlet allocation, as another mixed membership model, performs in a similar way to the GoM. Text data results in the same kind of sparse matrix as the multinomial data used in the GoM model, with LDA proceeding with words instead of items or statements and a single document for each individual. The dimension reduction using text is even more dramatic when starting with potentially thousands of unique words in the count matrix. However, the amount of data needed to run LDA with words composing the collection of discrete data is significant due to the large number of words in any given vocabulary. Without enough data, there are a variety of developments in topic modeling that are ripe for application within marketing, including using Dirichlet process priors (Ferguson, 1973; Antoniak, 1974) as a kind of distribution of heterogeneity over topic proportions. We leave the practical problems of using text in the place of traditional survey questions as an extension to this research.

Another extension relates to estimating the concentration parameter α and the optimal size of K. In working to properly account for extremes in the distribution of heterogeneity, we have seen that generating the predictive distribution of heterogeneity is sensitive to the Dirichlet hyperprior. While we have minimized the influence of α in generating that predictive distribution, one might consider how to inform an additional model layer so that α can be estimated instead. Additionally, while there isn't a consensus as to which measure of model fit provides the gold standard for determining the size of K, there are a number of extant methods for navigating across possible model dimensions that could be employed to include K as a parameter in the model (Green, 1995; Green et al., 2015). The technical details of how to incorporate such methods into the proposed model is left for future research.

More generally, we see the use of mixed membership models as a model-based approach to classifying consumers that yields a more realistic description of the individual as being a mixture of various extreme consumer profiles in a way that allows us to more realistically get into the extremes of the distribution of heterogeneity. This paper serves as a step toward fulfilling a broader need to provide more complete descriptions and explanations of consumer preference heterogeneity.

A Appendix: Membership Vector Model

A.1 Generating Data

Following Equation (9), we can generate data for each of N respondents as follows. First, fix the number of profiles K, the number of respondents N, the number of categorical questions J, and the categorical levels for each question n_j (e.g., $n_j = 2$ for all J in the case of pick any/J data), the number of choice tasks H, the number of alternatives in each choice task P, and the number of attribute levels L. Next, set the true values of Γ, V_β, and λ. We set α and τ to be vectors of 1 for the Dirichlet priors, creating a uniform distribution on the respective simplex to mirror our lack of information regarding partial membership and profile composition.

For each respondent n, we proceed as follows:

1. Draw $g_n \sim \text{Dirichlet}(\alpha)$, the membership vector.

2. For each of the $j = 1, \ldots, J$ questions:

   (a) Draw $z_{n,j} \sim \text{Multinomial}(g_n)$, a profile assignment.

   (b) Draw $w_{n,j} \sim \text{Multinomial}(\lambda_{j,k=z_{n,j}}(1), \ldots, \lambda_{j,k=z_{n,j}}(n_j))$, a categorical response drawn from the appropriate entry in $\lambda_{j,k}$ indexed by j and $k = z_{n,j}$.

3. Draw $\beta_n \sim \text{Normal}(\Gamma' g_n, V_\beta)$.

4. For each of the $h = 1, \ldots, H$ choice tasks:

   (a) Generate a design matrix $X_{n,h}$.

   (b) Compute latent utility for each alternative p:

$$U_{n,h,p} = X_{n,h,p}\beta_n + \varepsilon_{n,h,p};\; p = 1, \ldots, P$$

   where $\varepsilon_{n,h,p} \sim \text{EV}(0, 1)$.

(c) Let $y_{n,h} = \arg\max_p \left(\{U_{n,h,p}\}_{p=1}^{P}\right)$.

A.2 Estimation

Following Equation (10) and the DAG in Figure 4, we proceed with estimation as follows for R iterations:

1. For each of the $n = 1, \ldots, N$ respondents:

   (a) For each of the $j = 1, \ldots, J$ questions, draw $z_{n,j}$ using $g_{n,k}$, the partial membership respondent n has in each profile k, and $\lambda_{j,k}(w_{n,j})$, the probability of the chosen response $w_{n,j}$ for each profile k:

   $$z_{n,j} = \arg\max_k \left(\text{Multinomial}(p_1, \ldots, p_K)\right), \text{ where } p_k \propto g_{n,k}\lambda_{j,k}(w_{n,j}).$$

   (b) Draw g_n^{new} using a random-walk Metropolis-Hastings step where g_n^{old} is initialized at $1/K$ for all K elements and $g_n^{new} = \text{Dirichlet}(g_n^{old} \times s_p)$, where s_p is the specified step size. The larger s_p is, the closer g_n^{new} will be to g_n^{old}.

   (c) Accept g_n^{new} with probability

   $$\alpha_{\text{accept}} = \min\left(1, \frac{\left[\prod_{j=1}^{J} p(z_{n,j}|g_n^{new})\right] p(g_n^{new}|\alpha)p(\beta_n^{old}|g_n^{new}, \Gamma, V_\beta)p(g_n^{old}|g_n^{new})}{\left[\prod_{j=1}^{J} p(z_{n,j}|g_n^{old})\right] p(g_n^{old}|\alpha)p(\beta_n^{old}|g_n^{old}, \Gamma, V_\beta)p(g_n^{new}|g_n^{old})}\right)$$

   where the random-walk proposal density $p(g_n^{old}|g_n^{new})$ and $p(g_n^{new}|g_n^{old}) \sim$ Dirichlet.

   (d) Draw β_n^{new} using a random-walk Metropolis-Hastings step where β_n^{old} is initialized at 0 and $\beta_n^{new} = \beta_n^{old} + \epsilon, \epsilon \sim N(0, V_\beta \times s_\beta)$, where s_β is the specified step size. The smaller s_β is, the closer β_n^{new} will be to β_n^{old}.

   (e) Accept β_n^{new} with probability

   $$\alpha_{\text{accept}} = \min\left(1, \frac{p(y_n|X_n, \beta_n^{new})p(\beta_n^{new}|g_n^{old}, \Gamma, V_\beta)}{p(y_n|X_n, \beta_n^{old})p(\beta_n^{old}|g_n^{old}, \Gamma, V_\beta)}\right)$$

where the multivariate normal random-walk proposal density cancels out.

2. Draw Γ and V_β using $B = \Gamma' G + \Xi$ where G is a matrix with each g_n^{old} as a row vector, B is a matrix with each β_n as a row vector, and $\Xi \sim N(0, V_\beta)$.

3. Draw λ using counts of the augmented variable z:

$$\lambda_{j,k} \sim \text{Dirichlet}(p_{k,1}, \ldots, p_{k,n_j}), \text{ where } p_{k,l} \propto 1 + \sum_{n=1}^{N} I(z_{n,j} = l)$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

B Appendix: Factor Analysis and Factor Scores Model

B.1 Comparing Factor Analysis and the GoM

To detail the differences between factor analysis and the grade of membership (GoM) model, it is useful to write down a factor analytic model for binary data in the form of a cut-point model (Lee, 2007). In this model, observed responses from respondent n, w_n, are generated as follows:

$$w_{n,j} = 1 \text{ if } z_{n,j} > 0 \; \forall j \in \{1, \ldots, J\}, \quad z_n \sim N(\Lambda\zeta_n, \Sigma) \tag{12}$$

where w_n is a $J \times 1$ vector of observed binary responses, z_n is a $J \times 1$ vector of latent continuous responses, ζ_n is a $K \times 1$ vector of factor scores, the $J \times K$ matrix Λ indicates factor loadings, Σ indicates a $J \times J$ covariance matrix of the z's, and the threshold 0 is a fixed, arbitrarily chosen cut-point. The model in Equation (12) can easily be extended to ordinal data via additional cut-points (Johnson and Albert, 2006). Apart from the normal errors specification of this model and the subsequent need for identification constraints because of its scale invariance, this is a standard factor model and equivalent to a model in which multiple observed responses are regressed on unobserved factor scores.

The probability of observing a single response $w_{n,j}$, given this specification, can be expressed as an integral over the z space:

$$\begin{aligned} Pr(w_{n,j} = 1|\zeta_n, \Lambda) &= \int_0^\infty p(z_{n,j}|\{z_{n,-j}\}, \zeta_n, \Lambda, \Sigma)dz \\ &= \int_0^\infty N(\zeta'_{n,k}\gamma_{j,k}, \sigma^2_{n,j})dz \\ Pr(w_{n,j} = 0|\zeta_n, \Lambda) &= 1 - \int_0^\infty N(\zeta'_{n,k}\gamma_{j,k}, \sigma^2_{n,j})dz \end{aligned} \tag{13}$$

where $\sigma^2_{n,j}$ is the univariate variance of $z_{n,j}$, conditional on $\{z_{n,-j}\}$. In the case of conditionally independent regression errors, $\sigma^2_{n,j} = \sigma^2_n$. Equation (13) expresses the probability of observing a given response as the integral over the latent z space truncated at 0, given unobserved unit-level factor scores and across-unit factor loadings.

The GoM model expresses the probability of observing response l as an individual-level, multinomial mixture model of K profiles in which each response option for each question has profile-specific multinomial choice probabilities that are mixed over unit-specific weights (Erosheva et al., 2007):

$$Pr(w_{n,j} = l | g_n, \lambda) = \sum_{k=1}^{K} g_{n,k} \lambda_{j,k}(l) \tag{14}$$

in which the g_n indicate respondent-level weights of the profiles and $\lambda_{j,k}(l)$ is the probability of observing response l to question j given exclusive membership to profile k. As with every mixture model, the GoM model imposes the following constraints on the weights: $0 \leq g_{n,k} \leq 1$ and $\sum_{k=1}^{K} g_{n,k} = 1$. In the case of binary responses, the GoM model is simply:

$$\begin{aligned} Pr(w_{n,j} = 1 | g_n, \lambda) &= \sum_{k=1}^{K} g_{n,k} \lambda_{j,k}(1) \\ Pr(w_{n,j} = 0 | g_n, \lambda) &= \sum_{k=1}^{K} g_{n,k} (1 - \lambda_{j,k}(1)) \end{aligned} \tag{15}$$

where the g_n are subject to the same constraints.

Comparing the factor model in Equation (13) to the GoM model in Equation (15) reveals several differences between the two models. First, the GoM model is an individual-level mixture model of K latent profiles. The factor model is essentially a linear multivariate regression model. This leads to a different interpretation of g_n, compared to the latent factor scores ζ_n. The g_n represent convex weights over a multidimensional latent space whereas the factor scores are the set of latent sources of observed responses, each of which is unidimensional and which contribute to the observed responses in a linear fashion. The important difference lies in what the underlying construct is. In the GoM model, a profile is defined as a set of response probabilities across *all* J questions and their response options. In factor analysis, a factor is assumed to exist independently from the measurements.

Second, and because of its mixture model property, the GoM model allows us to capture response heterogeneity in two ways. First, it allows for unit-level latent scores g_n, which captures differences among respondents. Second, it captures heterogeneity in responses through profile-specific response probabilities $\lambda_{j,k}$. Individual response behavior is described in terms of the similarity of individual and profile-specific response probabilities. An individual's response behavior more similar to one of the profiles across *all* responses is expressed by a higher weight of that profile for that individual.

Third, factor analysis makes specific assumptions concerning the distribution of observed responses. More specifically, factor analysis assumes that for the data in Equation (13), the z are distributed multivariate normal. The GoM model, in comparison, makes no assumption about the joint distribution of observed responses.

This suggests that whether or not the GoM model is to be preferred over a factor model is essentially a question of which model is an adequate description of respondent heterogeneity in a particular application. The GoM model describes heterogeneity as similarity between individuals and extreme profiles. The number of extreme profiles in the GoM model is defined a priori and can be used to reduce the dimensionality of the response space. Factor analysis is often used for the same purpose, but it lacks the property of locating individuals in the convex space spanned by extreme response behavior.

B.2 Estimating the Factor Scores Model

For the upper level of a hierarchical choice model we consider a standard factor model of indicators Z:

$$\beta_i = \Gamma \xi_i + \zeta_i, \quad \zeta_i \sim N\left(0, \Sigma_\zeta\right) \tag{16}$$

and

$$Z_i = \Lambda \xi_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \Sigma_\epsilon\right) \tag{17}$$

where:

β_i is a vector of part-worths of respondent i of length M

Γ is a $M \times q$ matrix of (common) regression coefficients

ξ_i is a vector of (latent) factor scores of respondent i of length q

ζ_i is a vector of errors of length M

Z_i is a vector of observed (continous) variables of length J

Λ is a $J \times q$ matrix of (common) factor loadings

ϵ_i is a vector of errors of length J

Typically, in an application of factor models in marketing, $q \ll J$, giving rise to the dimensionality reduction property of the factor model. Because of the distributional assumption made in Equation (17), this model is not applicable to binary data. We therefore extend the model to accommodate binary data using a binary probit specification:

$$\begin{aligned} w_{i,j} &= 1 \quad if \quad Z_{i,j} > 0 \\ w_{i,j} &= 0 \quad if \quad Z_{i,j} \leq 0 \end{aligned} \tag{18}$$

In this model, the Z are latent continuous variables giving rise to binary W via a fixed cutpoint $c = 0$. Conditional on ξ_i, this model factorizes as follows:

$$p\left(\beta_i | \Gamma \xi_i, \Sigma_\zeta\right) p\left(Z_i | \Lambda \xi_i, \Sigma_\epsilon\right) p\left(W_i | Z_i, c\right) \tag{19}$$

where the last expression is an indicator function. The complete choice model including prior distributions is then:

$$p\left(y_i|X_i,\beta_i\right)p\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)p\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)p\left(X_i|Z_i,c\right)p\left(\xi_i|0,I_q\right)p\left(\Sigma_\zeta\right)p\left(\Sigma_\epsilon\right)p\left(\Gamma\right)p\left(\Lambda\right) \quad (20)$$

where:

X_i is a matrix of attributes of the choice options

y_i is the observed set of multinomial outcomes

I_q is the identity matrix of dimension $q \times q$

The prior specification for ξ_i follows standard conventions for the factor model in which mean and the scale of the latent factor scores are a priori fixed for identification of the model. Further following conventions, we specify Σ_ϵ as a diagonal matrix implying uncorrelated errors of the regression in Equation (17). Equation (20) suggest the following Gibbs sampling scheme for the integrated choice and factor analysis model with binary covariates:

1. $p\left(\beta_i|y_i,X_i,\Gamma,\xi_i,\Sigma_\zeta\right)\propto p\left(y_i|X_i,\beta_i\right)p\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)$
2. $p\left(\Sigma_\zeta|\beta,\Gamma,\xi\right)\propto\prod_{i=1}^N p\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)p\left(\Sigma_\zeta\right)$
3. $p\left(\Gamma|\beta,\xi,\Sigma_\zeta\right)\propto\prod_{i=1}^N p\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)p\left(\Gamma\right)$
4. $p\left(\xi_i|\beta_i,\Gamma,\Sigma_\zeta,Z_i,\Lambda,\Sigma_\epsilon\right)\propto p\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)p\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)p\left(\xi_i|0,I_q\right)$
5. $p\left(Z_i|\xi_i,\Lambda,\Sigma_\epsilon,X_i,c\right)\propto p\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)p\left(X_i|Z_i,c\right)$
6. $p\left(\Lambda|Z,\xi,\Sigma_\epsilon\right)\propto\prod_{i=1}^N p\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)p\left(\Lambda\right)$
7. $p\left(\Sigma_\epsilon|Z,\Lambda,\xi\right)\propto\prod_{i=1}^N p\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)p\left(\Sigma_\epsilon\right)$

For sampling steps 1-3, 6 and 7 we use standard results from the literature (Rossi et al. 2006, Johnson and Albert 2006). For the other steps, we proceed as follows:

B.2.1 Sampling ξ

We start by noting that both likelihood contributions to the ξ_i as well as the (fixed) prior are normal distributions:

$$N\left(\beta_i|\Gamma\xi_i,\Sigma_\zeta\right)N\left(Z_i|\Lambda\xi_i,\Sigma_\epsilon\right)N\left(\xi_i|0,I_q\right) \tag{21}$$

the resulting exponent of the product of these terms is:

$$\left(\beta_i-\Gamma\xi_i\right)^T\Sigma_\zeta^{-1}\left(\beta_i-\Gamma\xi_i\right)+\left(Z_i-\Lambda\xi_i\right)^T\Sigma_\epsilon^{-1}\left(Z_i-\Lambda\xi_i\right)+\xi_i^TI_q\xi_i \tag{22}$$

Collecting terms and dropping constants with respect to ξ_i yields:

$$-2\xi_i^T\left(\Gamma^T\Sigma_\zeta^{-1}\beta_i+\Lambda^T\Sigma_\epsilon^{-1}Z_i\right)+\xi_i^T\left(\Gamma^T\Sigma_\zeta^{-1}\Gamma+\Lambda^T\Sigma_\epsilon^{-1}\Lambda+I_q\right)\xi_i \tag{23}$$

which implies that the conditional posterior distribution of ξ_i is a multivariate normal distribution with:

$$N\left(\left(\Gamma^T\Sigma_\zeta^{-1}\Gamma+\Lambda^T\Sigma_\epsilon^{-1}\Lambda+I_q\right)^{-1}\left(\Gamma^T\Sigma_\zeta^{-1}\beta_i+\Lambda^T\Sigma_\epsilon^{-1}Z_i\right),\left(\Gamma^T\Sigma_\zeta^{-1}\Gamma+\Lambda^T\Sigma_\epsilon^{-1}\Lambda+I_q\right)^{-1}\right)$$

B.2.2 Sampling Z

The sampling of Z_i is achieved via a truncated normal distribution:

$$p\left(Z_i|\xi_i,\Lambda,\Sigma_\epsilon,X_i,c\right)\propto\begin{cases}N_0^\infty\left(\Lambda\xi_i,\Sigma_\epsilon\right) & if \quad x_i=1\\ N_{-\infty}^0\left(\Lambda\xi_i,\Sigma_\epsilon\right) & if \quad x_i=0\end{cases}$$

where the subscript and superscript of N indicate the lower and upper cut-point, respectively. Since we assume that Σ_ϵ is a diagonal matrix we can sample the $z_{i,j}$ independently from univariate normal distributions:

$$p\left(z_{i,j}|\xi_i, \Lambda_j, \sigma_{\epsilon_j}, X_{i,j}, c\right) \propto \begin{cases} N_0^{\infty}\left(\Lambda_j^T \xi_i, \sigma_{\epsilon_j}^2\right) & \mathit{if} \quad x_{i,j} = 1 \\ N_{-\infty}^{0}\left(\Lambda_j^T \xi_i, \sigma_{\epsilon_j}^2\right) & \mathit{if} \quad x_{i,j} = 0 \end{cases} \tag{24}$$

We note that the binary probit model in (24) is not scale-identified. One way to solve this problem is estimate Σ_ϵ and to post-process the results (Rossi et al. 2006). Another strategy is to fix $\Sigma_\epsilon = I_J$ which is the strategy we apply.

References

Airoldi, Edoardo M, David Blei, Elena A Erosheva, Stephen E Fienberg. 2014. *Handbook of Mixed Membership Models and Their Applications*. 1st ed. Chapman & Hall/CRC.

Allenby, G M, James L Ginter. 1995. Using Extremes to Design Products and Segment Markets. *Journal of Marketing Research* **32**(4) 392–403.

Allenby, Greg M, Neeraj Arora, James L Ginter. 1998. On the Heterogeneity of Demand. *Journal of Marketing Research* **35**(3) 384–389.

Allenby, Greg M, Peter E Rossi. 1998. Marketing Models of Consumer Heterogeneity. *Journal of Econometrics* **89**(1-2) 57–78.

Antoniak, Charles E. 1974. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics* **2**(6) 1152–1174.

Archak, Nikolay, Anindya Ghose, Panagiotis G Ipeirotis. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science* **57**(8) 1485–1509.

Blei, David M, Jon D McAuliffe. 2007. Supervised topic models. *Neural Information Processing Systems*.

Blei, David M, Andrew Y Ng, Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3** 993–1022.

Büschken, Joachim, Greg M Allenby. 2016. Sentence-Based Text Analysis for Customer Reviews. *Marketing Science* (forthcoming).

Chandukala, Sandeep R, Yancy D Edwards, Greg M Allenby. 2011. Identifying Unmet Demand. *Marketing Science* **30**(1) 61–73.

Clive, Jonathan, Max A Woodbury, Ilene C Siegler. 1983. Fuzzy and Crisp Set-Theoretic-Based Classification of Health and Disease. *Journal of Medical Systems* **7**(4) 317–332.

Erosheva, Elena A. 2002. Grade of Membership and Latent Structure Models With Application to Disability Survey Data. *Ph.D. thesis, Department of Statistics, Carnegie Mellon University* .

Erosheva, Elena A, Stephen E Fienberg, Cyrille Joutard. 2007. Describing Disability Through Individual-Level Mixture Models for Multivariate Binary Data. *The Annals of Applied Statistics* **1**(2) 346–384.

Ferguson, Thomas S. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1**(2) 209–230.

Galyardt, April. 2014. Interpreting Mixed Membership. *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall/CRC, 39–65.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, Donald B Rubin. 2013. *Bayesian Data Analysis*. Third edition ed. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Gelman, Andrew, Iain Pardoe. 2006. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. *Technometrics* **48**(2) 241–251.

Green, Peter J. 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82**(4) 711–732.

Green, Peter J, Krzysztof Łatuszyński, Marcelo Pereyra, Christian P Robert. 2015. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing* **25**(4) 835–862.

Gross, Justin H, Daniel Manrique-Vallier. 2014. A Mixed-Membership Approach to the Assessment of Political Ideology from Survey Responses. *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall/CRC, 119–139.

Horsky, Dan, Sanjog Misra, Paul Nelson. 2006. Observed and Unobserved Preference Heterogeneity in Brand-Choice Models. *Marketing Science* **25**(4) 322–335.

Johnson, Valen E, James H Albert. 2006. *Ordinal Data Modeling*. Springer Science & Business Media.

Joutard, Cyrille, Edoardo M Airoldi, Stephen E Fienberg, Tanzy M Love. 2007. Discovery of Latent Patterns with Hierarchical Bayesian Mixed-Membership Models and the Issue of Model Choice. *Data Mining Patterns: New Methods and Applications*. IGI Global, Hershey, PA, USA, 1–36.

Kamakura, Wagner A, Gary J Russell. 1989. A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. *Journal of Marketing Research* **26**(4) 379–390.

Lee, Sik-Yum. 2007. *Structural Equation Modeling: A Bayesian Approach*, vol. 711. John Wiley & Sons.

Lee, Thomas Y, Eric T Bradlow. 2011. Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research* **48**(5) 881–894.

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, Martin R Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science* **15**(2) 173–191.

Manton, Kenneth G, Max A Woodbury, H Dennis Tolley. 1994. *Statistical Application Using Fuzzy Sets*. Wiley, New York.

Marini, Margaret Mooney, Xiaoli Li, Pi-Ling Fan. 1996. Characterizing Latent Structure: Factor Analytic and Grade of Membership Models. *Sociological Methodology* **26** 133–164.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, Moshe Fresko. 2012. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science* **31**(3) 521–543.

Newton, Michael A, Adrian E Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(1) 3–48.

Rossi, Peter E, Greg M Allenby. 2003. Bayesian Statistics and Marketing. *Marketing Science* **22**(3) 304–328.

Rossi, Peter E, Greg M Allenby, Robert E McCulloch. 2005. *Bayesian Statistics and Marketing*. J. Wiley and Sons.

Rossi, Peter E, Robert E McCulloch, Greg M Allenby. 1996. The Value of Purchase History Data in Target Marketing. *Marketing Science* **15**(4) 321–340.

Spiegelhalter, David J, Nicola G Best, Bradley P Carlin, Angelika Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4) 583–639.

Stewart, David W. 1981. The Application and Misapplication of Factor Analysis in Marketing Research. *Journal of Marketing Research* **18**(1) 51–62.

Tanner, Martin A, Wing Hung Wong. 1987. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82**(398) 528–540.

Tirunillai, Seshadri, Gerard J Tellis. 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research* **51**(4) 463–479.

Woodbury, Max A, Jonathan Clive, Arthur Garson Jr. 1978. Mathematical Typology: A Grade of Membership Technique for Obtaining Disease Definition. *Computers and Biomedical Research* **11**(3) 277–298.