# A Large-Scale Marketing Model using Variational Bayes Inference for Sparse Transaction Data

DSSR

Discussion Paper No. 18

**A Large-Scale Marketing Model using Variational Bayes Inference for Sparse Transaction Data**

Tsukasa Ishigaki
Nobuhiko Terui
Tadahiko Sato
Greg M. Allenby

January , 2014

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

# A Large-Scale Marketing Model using Variational Bayes Inference for Sparse Transaction Data

Tsukasa Ishigaki*

Nobuhiko Terui*

Tadahiko Sato**

Greg M. Allenby***

January, 2014

*Graduate School of Economics and Management, Tohoku University, Sendai 980-8576, Japan

** Graduate School of Business Science, University of Tsukuba, Tokyo 112-0012, Japan

*** Fisher College of Business, Ohio State University, Columbus, OH, U.S.A.

Ishigaki acknowledges the financial support of KAKENHI Grant-in-Aid for Young Scientists (A) 24683012. Terui acknowledges the financial support of the Japanese Ministry of Education Scientific Research Grants (A) 25245054.

**Abstract**

Large-scale databases in marketing track multiple consumers across multiple product categories. A challenge in modeling these data is the resulting size of the data matrix, which often has thousands of consumers and thousands of choice alternatives with prices and merchandising variables changing over time. We develop a heterogeneous topic model for these data, and employ variational Bayes techniques for estimation that are shown to be accurate in a Monte Carlo simulation study. We find the model to be highly scalable and useful for identifying effective marketing variables for different consumers, and for predicting the choices of infrequent purchasers.

*Key words and phrases:*

database marketing, dimension reduction, Bayesian analysis, choice models, topic model, targeting

## 1. Introduction

Modern analytic techniques in marketing are continuously confronted with the necessity of extracting relevant information from large volumes of data by identifying important drivers of consumer behavior. It is common for datasets to record household purchases of products that are orders of magnitude larger than what current models of behavior are currently capable. Existing models of choice and demand, for example, are typically limited to less than twenty or so product alternatives that are tracked across possibly hundreds of consumers (see Rossi et al. 2005 and Chintagunta and Nair 2011).

Increasing the number of products analyzed is problematic because of potential complexities in the structure of demand and the accompanying increase in the required number of model parameters. Increasing the number of respondents is also problematic because of computational constraints arising from respondent heterogeneity that is found to be important in describing demand and deriving policy implications. While a variety of dimension-reducing techniques have been studied in the fields of statistics and data-mining, the presence of heterogeneous consumers and heterogeneous purchase environments with prices and other variables change over occasions requires the use of model-based inference as opposed to methods applied directly to the marginal data (Chintagunta and Nair 2011).

Naik et al. (2008) discusses three solutions to the challenges in massive data analysis: increasing computer power, employing alternative approaches for data analysis, and using scalable estimation methods. In this paper, we combine the second and third options to obtain improved inferences about consumer behavior in large datasets. We extend the voting bloc model of Spirling and Quinn (2010) and Grimmer (2011) that are a variation of topic models used to conduct large-scale analysis of text data (Blei et al. 2003).

The topic model is a generalization of a finite mixture model (Kamakura and Russell 1989) in which each data point is associated with a draw from a mixing distribution (Teh and Jordon 2010). Models of voting blocs (Spirling and Quinn 2010), for example, track the votes of legislators (aye or nay) across multiple bills, with each bill associated with a potentially different concern or issue. Similarly, the latent Dirichlet allocation (LDA) model of Blei et al. (2003) allocates words within documents to a small number of latent topics whose patterns are meaningful and interpretable. Each vote and each word is associated with a potentially different issue or topic, and hence the mixing distribution is applied to the individual datum. In our

analysis of household purchases, we allow every purchase (and every non-purchase) in every product category to be related to a potentially different latent context (topic, or issue) for which the good is purchased. This allows us to view a consumer's purchases as responding to different needs or occasions (e.g., family dinner, snacks, etc.), and allows us to identify the ensemble of goods that collectively define latent product segments across a large number of products.

We obtain a scalable estimation method by employing variational Bayes (VB) inference as in Jordan et al. (1999) and Bishop (2006), instead of the standard Markov chain Monte Carlo (MCMC) inference. MCMC methods can incur large computational cost in large-scale problems. VB inference approximates a posterior distribution of target by variational optimization in a computationally efficient manner.

Our approach combines variational Bayes (VB) methods, as in Jordan et al. (1999) and Bishop (2006), with a topic-like probit model to obtain a computationally feasible model of consumer purchases that is scalable to large databases. Individual-level inference is possible in our model, where we can identify the marketing variables that are effective for specific individuals and the products for which they are effective. Our model is therefore similar to adaptive personalization systems proposed by Ansari and Mela (2003), Rust and Chung (2006), Chung et al. (2009) and Braun and McAuliffe (2010). However, it is different in that our model structure facilitates analysis of a much larger array of product categories.

In the next section, we propose a model for consumer purchases in multiple product categories. Section 3 describes a variational Bayes inference scheme for the models and a simulation study that verifies the scalability. The prediction performance of the proposed models is presented in Section 4. Section 5 applies the model to actual customer purchases in a general merchandise store. Discussion and concluding remarks are presented in Section 6.

## 2. Model

Estimating parameters of choice model for a large number of consumers and products is often computationally infeasible. In addition, the actual sample size of transactional data is often much smaller than the data space reflected by a data cube with dimensions corresponding to the number of consumers, number of products and time, making fixed-effect estimates of model parameters with heterogeneity unsuitable. We address this challenge by relating consumer purchases to latent segments (similar to topics and blocs) that greatly reduces the

dimensionality of the model. Response parameters are then introduced in the reduced dimensional space by connecting each choice to marketing variables with a hierarchical probit model.

**2.1 Dimensional Reduction by Topic Models**

Dimensional reduction is an important technique in massive data analysis. Here we briefly introduce the idea of introducing a latent variable to common in topic models in the context of consumer purchases. We seek the probability $p(i \mid c)$ that consumer *c* purchases item *i* . Dataset includes *C* consumers and *I* product items through *T* periods. However, the probabilities cannot be directly calculated because of computational difficulty imposed by the large-scale setting and data sparseness. The topic model calculates $p(i \mid c)$ by introducing a latent class $z \in \{1, \cdots, Z\}$ whose dimension is significantly smaller than the number of consumers and items ( $Z \ll C, I$ ).

The latent variable is used to represent the sparse data matrix as a finite mixture of vectors commonly found in topic models:

$$\begin{bmatrix} p(i=1 \mid c=1) & \cdots & p(i=1 \mid c=C) \\ \vdots & \ddots & \vdots \\ p(i=I \mid c=1) & \cdots & p(i=I \mid c=C) \end{bmatrix} = \sum_{z=1}^{Z} \begin{bmatrix} p(1 \mid z) \\ \vdots \\ p(I \mid z) \end{bmatrix} \begin{bmatrix} p(z \mid 1) \cdots p(z \mid C) \end{bmatrix}. \qquad (1)$$

More specifically, we decompose a large probability matrix of size $C \times I$ to two small probability matrices of sizes $I \times Z$ and $Z \times C$ based on the property of conditional independence.

The main difference between voting blocs model and LDA is assumed distributions for probabilities $p(i \mid z)$ in the $I \times Z$ size probability matrix. The voting blocs model supposes a Bernoulli distribution for the probability $p(i \mid z)$. LDA assumes a categorical distribution for the probability matrix.

In the analysis of purchase behavior using topic models for large consumer transaction data, Iwata et al. (2009) extracted dynamic patterns between purchased product items and consumer interests. Ishigaki et al. (2010) fused heterogeneous transaction data and consumer lifestyle questionnaire data, while Iwata et al. (2012) identified consumer purchase patterns by using a topic model with price information on the purchased products. These approaches identify patterns among consumers and product items. The labeled LDA proposed by Ramage et al. (2009), and the supervised LDA of Blei and McAuliffe (2007) extend the topic models

by incorporating additional data in the analysis. However, none of these approaches are suitable for relating marketing variables to individual consumer choices as explanation variables. In the following sections, we construct a model that links marketing variables with consumers and products.

**2.2 A Reduced Dimensional Choice Model**

Let $y_{cit}$ denote consumer *c*'s purchase record of product *i* at time *t*, assigning $y_{cit} = 1$ if consumer *c* purchased the item, and $y_{cit} = 0$ otherwise. Denote $u_{cit}$ as the utility of consumer *c*'s purchase record of product *i* at time *t*. We assume a binary probit model with $u_{cit} > 0$ if $y_{cit} = 1$, and $u_{cit} \le 0$ if $y_{cit} = 0$. We couple the topic model in (1) with the binary choice probability as in a voting bloc model to obtain the choice probability:

$$p(u_{cit} > 0) = \sum_{z=1}^{Z} p(u_{cit} > 0 \mid z)\, p(z \mid c). \tag{2}$$

We denote the utility associated with the latent class *z* as $u_{cit}^{(z)}$, and then the choice probability can be represented as $p\left(u_{cit} > 0 \mid z\right) = p\left(u_{cit}^{(z)} > 0\right)$. Assuming a linear Gaussian structure on the segment utility $u_{cit}^{(z)}$ with marketing variables as $u_{cit}^{(z)} \sim N\left(\mathbf{x}_{it}^{T}\boldsymbol{\beta}_{zi}, 1\right)$, the right hand side of (1) is represented as,

$$\sum_{z=1}^{Z} \begin{bmatrix} F(\mathbf{x}_{1t}^{T}\boldsymbol{\beta}_{z1}) \\ \vdots \\ F(\mathbf{x}_{It}^{T}\boldsymbol{\beta}_{zI}) \end{bmatrix} \left[ p(z \mid 1) \cdots p(z \mid C) \right] \tag{3}$$

where $\boldsymbol{\beta}_{zi} = \left[\beta_{zi1}, \cdots, \beta_{ziM}\right]^{T}$ is a response coefficient vector of latent class *z* with respect to item *i*, $\mathbf{x}_{it} = \left[x_{it1}, \cdots, x_{itM}\right]^{T}$ is a vector of *M* marketing variable for item *i* at time *t*, and $F(\bullet)$ is the cumulative distribution function (CDF) of the standard normal distribution. In our empirical study, $\mathbf{x}_{it}$ includes price and promotional variables.

We next set a categorical distribution $C_{cz}$ for the probability $p(z \mid c)$ that consumer *c* belongs to the latent class *z*. The categorical distribution is multinomial with parameters $\mathbf{C}_c$. The $\mathbf{C}_c$ is determined so that the selection probability of consumer *c* with respect to item *i* is *conditionally independent* if the latent class *z* is given. Then, the right hand side of (1) is represented by:

$$\sum_{z=1}^{Z} \begin{bmatrix} F(\mathbf{x}_{1t}^{T}\boldsymbol{\beta}_{z1}) \\ \vdots \\ F(\mathbf{x}_{It}^{T}\boldsymbol{\beta}_{zI}) \end{bmatrix} \left[ C_{1z} \cdots C_{Cz} \right] \tag{4}$$

Finally, segment-level heterogeneity is introduced through a hierarchical model with a random effect for response coefficient $\boldsymbol{\beta}_{zi}$

$$\boldsymbol{\beta}_{zi} \sim N_M\left(\boldsymbol{\mu}_i, V_i\right), \tag{5}$$

where the prior distributions for $\boldsymbol{\mu}_i$ and $V_i$ follow an *M*-dimensional multivariable normal distribution $N_M\left(\tilde{\boldsymbol{\mu}}, \tilde{\sigma}_\mu V_i\right)$ and an inverse-Wishart distribution $\mathrm{IW}\left(\tilde{W}, \tilde{w}\right)$, respectively. $\tilde{\boldsymbol{\mu}}$, $\tilde{\sigma}_\mu$, $\tilde{W}$ and $\tilde{w}$ are hyperparameters. That is, we assume that the *M*-dimensional coefficient vector $\beta_{zi}$ for each segment, *z*, is a draw from a distribution with mean and covariance that is item-specific.

We specify a prior distribution for $\mathbf{C}_c$, assuming the Dirichlet distribution as the natural conjugate prior distribution of categorical distribution:

$$\mathbf{C}_c \sim \mathrm{Dirichlet}\left(\tilde{\boldsymbol{\gamma}}\right), \tag{6}$$

where $\tilde{\boldsymbol{\gamma}}$ is a hyperparameter vector of the Dirichlet distribution.

The likelihood is given as

$$l\left(\{y_{cit}\} \mid \{\mathbf{C}_c\}, \{\boldsymbol{\beta}_{zi}\}, \{\mathbf{x}_{it}\}\right) = \prod_{c=1}^{C} \prod_{i \in I_c} \prod_{t \in T_c} \sum_{z=1}^{Z} \left[ C_{cz}\, p\left(y_{cit} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_{zi}, z\right) \right], \tag{7}$$

where $p\left(y_{cit} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_{zi}, z\right)$ denotes the kernel of the binary probit model conditional on *z*, $T_c$ denotes a subset of *t* in which consumer *c* purchased any item in a store, and $I_c$ is a subset of items *i* purchased by consumer *c* during the period $t = 1, \cdots, T$, that is, $T_c \in \left\{ t \mid \sum_{i=1}^{I} y_{cit} > 0 \right\}$ and $I_c \in \left\{ i \mid \sum_{t=1}^{T} y_{cit} > 0 \right\}$.

Equation (7) is difficult to use directly because the likelihood includes summations over latent class *z*. Instead, we employ a data augmentation approach by Tanner (1987) with respect to latent variable *z*. We introduce variables $z_{cit} \in \{1, \cdots, Z\}$ denoting the label of the latent class for each consumer *c*, each purchased item *i*, and each purchasing event *t*. Conditioning on the $z_{cit}$ for each purchasing transaction, as in the LDA of Blei et al. (2003), the likelihood in (7) simplifies to:

$$l\left(\{y_{cit}\} \mid \{\mathbf{C}_c\}, \{z_{cit}\}, \{\boldsymbol{\beta}_{zi}\}, \{\mathbf{x}_{it}\}\right) = \prod_{c=1}^{C} \prod_{i \in I_c} \prod_{t \in T_c} p\left(z_{cit} \mid \mathbf{C}_c\right) p\left(y_{cit} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_{zi}, z_{cit}\right), \tag{8}$$

where $p\left(z_{cit} \mid \mathbf{C}_c\right)$ denotes a categorical distribution when $\mathbf{C}_c$ is given.

Our model for massive data analysis is different to the LDA model in that it only deals with the presence of products appearing in the purchase basket of the consumer, and does not deal

with non-purchase of product. This is different to model encountered in the analysis of text data where it is the presence of words, and not their absence, that characterizes the latent topics. The co-occurrence of the products selected during a shopping trip is what gives meaning to segments, as modified by the marketing variables.

The posterior distribution of parameters including latent variables $\{z_{cit}\},\{u_{cit}^{(z)}\}$ is then given by

$$
\begin{aligned}
& p\left(\{\mathbf{C}_c\},\{z_{cit}\},\{u_{cit}^{(z)}\},\{\boldsymbol{\beta}_{zi}\},\{\boldsymbol{\mu}_i\},\{V_i\}\,|\,\{\mathbf{x}_{it}\},\{y_{cit}\}\right) \\
&= p\left(\{\mathbf{C}_c\}\,|\,\{z_{cit}\}\right) \\
&\times p\left(\{z_{cit}\}\,|\,\{\mathbf{C}_c,\boldsymbol{\beta}_{zi},\mathbf{x}_{it},y_{cit}\}\right) \\
&\times p\left(\{u_{cit}^{(z)}\}\,|\,\{\boldsymbol{\beta}_{zi},z_{cit},\mathbf{x}_{it},y_{cit}\}\right) \\
&\times p\left(\{\boldsymbol{\mu}_i,V_i\}\,|\,\{\boldsymbol{\beta}_{zi}\}\right) \\
&\times p\left(\{\boldsymbol{\beta}_{zi}\}\,|\,\{u_{cit}^{(z)},\boldsymbol{\mu}_i,V_i,\mathbf{x}_{it}\}\right) \\
&\propto p\left(\{\mathbf{C}_c\},\{z_{cit}\},\{u_{cit}^{(z)}\},\{\boldsymbol{\beta}_{zi}\},\{\boldsymbol{\mu}_i\},\{V_i\},\{\mathbf{x}_{it}\},\{y_{cit}\}\right) \\
&= \left[\prod_{c=1}^{C} p\left(\mathbf{C}_c\right)\right]\left[\prod_{i=1}^{I} p\left(\boldsymbol{\mu}_i,V_i\right)\prod_{z=1}^{Z} p\left(\boldsymbol{\beta}_{zi}\,|\,\boldsymbol{\mu}_i,V_i\right)\right] \\
&\left[\prod_{c=1}^{C}\prod_{i\in I_c}\prod_{t\in T_c} p\left(z_{cit}\,|\,\mathbf{C}_c\right) p\left(y_{cit}\,|\,u_{cit}^{(z)},\boldsymbol{\beta}_{zi},z_{cit},\mathbf{x}_{it}\right)\right]
\end{aligned}
\tag{9}
$$

## 3. Variational Bayes Inference

We introduce VB inference in order to achieve computational feasibility for large-scale transaction data. VB inference approximates the posterior, or target distribution in a Bayesian model. The advantage of this method over MCMC is low computational cost. VB also takes advantage of parameters that can be decomposed into several mutually independent groups. This is necessary for our analysis using a large database.

The target and approximate distributions are denoted as *p* and *q*, respectively. The latter is called the *variational* distribution. Distributions *p* and *q* share a parameter set $\boldsymbol{\theta}$. In general, when the data **D** is given, the log marginal likelihood $p(\mathbf{D})$ of the target distribution is decomposed into two components as

$$
\begin{aligned}
\log p(\mathbf{D}) &= L(q) + KL(q \parallel p), \\
L(q) &= \int q(\boldsymbol{\theta}) \log \left\{ p(\mathbf{D}, \boldsymbol{\theta}) q(\boldsymbol{\theta})^{-1} \right\} dZ, \\
\mathrm{KL}(q \parallel p) &= -\int q(\boldsymbol{\theta}) \log \left\{ p(\boldsymbol{\theta} \mid \mathbf{D}) q(\boldsymbol{\theta})^{-1} \right\} dZ.
\end{aligned} \tag{10}
$$

$L(q)$ is called the *variational lower bound* in VB inference, and $\mathrm{KL}(q \parallel p)$ is the Kullback–Leibler divergence of the target and variational distributions. As is well known, $\mathrm{KL}(q \parallel p)$ is zero if $p$ and $q$ are the same distribution. Therefore, a reasonable solution to estimating the posterior distribution $p$ is the variational distribution $q$ for which $\mathrm{KL}(q \parallel p)$ is minimized. However, it is difficult to evaluate the value of $\mathrm{KL}(q \parallel p)$ because the expression involves a posterior distribution of $p(\boldsymbol{\theta} \mid \mathbf{D})$.

In contrast, $L(q)$ involves a joint distribution $p(\mathbf{D}, \boldsymbol{\theta})$ that is easily evaluated in many cases because it is obtained as the product of the prior and the likelihood in Bayesian models. We note that maximizing $L(q)$ is equivalent to minimizing $\mathrm{KL}(q \parallel p)$ because the log marginal likelihood of the target distribution is constant for a given dataset. In this situation, assuming that the distribution $q$ and parameter set $\boldsymbol{\theta}$ are decomposable for some groups as $q(\boldsymbol{\theta}) = \prod_j^J q_j\left(\boldsymbol{\theta}^{(j)*}\right)$, where the parameters $\boldsymbol{\theta}^{(j)*}$ are called *variational parameters*, $L(q)$ can be maximized by the following updating algorithm (Jordan et al., 1999):

$$
\begin{aligned}
\boldsymbol{\theta}^{(j)*\{new\}} &\leftarrow \arg\max_{\boldsymbol{\theta}^{(j)*}} L\left(\prod_j^J q_j\left(\boldsymbol{\theta}^{(j)*}\right)\right) \\
&\propto \exp\left(\mathbf{E}_{i \neq j}\left[\log p\left(\mathbf{D}, \boldsymbol{\theta}\right)\right]\right).
\end{aligned} \tag{11}
$$

The $\mathbf{E}_{i \neq j}[\ ]$ are the expectation value associated with $q_j$ distributions over all parameters $\boldsymbol{\theta}^{(j)*}$, where $i \neq j$. The variational parameters are updated for each variational parameter set $\boldsymbol{\theta}^{(j)*}$ until convergence of the algorithm. The initial variational parameters are proper random values. The VB is guaranteed to converge after several iterations because $L(q)$ is convex with respect to each $q_j\left(\boldsymbol{\theta}^{(j)*}\right)$ (Bishop, 2006). The variational lower bound monotonically increases as the iteration proceeds; therefore, convergence can be confirmed by checking the value of $L(q)$ at each iteration.

### 3.1 VB for the Proposed Model

We introduce the variational distributions and parameters for the modes of proposed model. The parameters and variational parameters are denoted as

$\boldsymbol{\theta} = \left\{\{\mathbf{C}_c\}, \{z_{cit}\}, \{u_{cit}^{(z)}\}, \{\boldsymbol{\beta}_{zi}\}, \{\boldsymbol{\mu}_i\}, \{V_i\}\right\}$ and

$\boldsymbol{\theta}^* = \left\{\left\{\mathbf{C}_{cit}^*\right\}, \left\{\boldsymbol{\gamma}_c^*\right\}, \left\{\boldsymbol{\beta}_{cit}^{(z)*}\right\}, \left\{\boldsymbol{\mu}_{iz}^*\right\}, \left\{V_{iz}^*\right\}, \left\{\boldsymbol{\mu}_i^{\mu*}\right\}, \left\{\sigma_i^{\mu*}\right\}, \left\{w_i^*\right\}, \left\{W_i^*\right\}\right\}$ respectively, while the variational distributions are configured as

$$
\begin{aligned}
& q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*, \{\mathbf{x}_{it}\}, \{y_{cit}\}\right) \\
& = \left[\prod_{c=1}^{C} q_c\left(\mathbf{C}_c \mid \boldsymbol{\gamma}_c^*\right)\right]\left[\prod_{c=1}^{C}\prod_{i\in I_c}\prod_{t\in T_c} q_z\left(z_{cit} \mid \mathbf{C}_{cit}^*\right)\right]\left[\prod_{c=1}^{C}\prod_{i\in I_c}\prod_{t\in T_c} q_u\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{cit}^{(z)*}, \mathbf{x}_{it}, z_{cit}\right)\right] \\
& \left[\prod_{i=1}^{I}\prod_{z=1}^{Z} q_\beta\left(\boldsymbol{\beta}_{zi} \mid \boldsymbol{\mu}_{zi}^*, V_{zi}^*\right)\right]\left[\prod_{i=1}^{I} q_{\mu,V}\left(\boldsymbol{\mu}_i, V_i \mid \boldsymbol{\mu}_i^{\mu*}, \sigma_i^{\mu*}, w_i^*, W_i^*\right)\right]
\end{aligned}
\tag{12}
$$

where $q_c$ is a Dirichlet distribution with variational parameter $\boldsymbol{\gamma}_c^*$, $q_c$ is a categorical distribution with variational parameter $\mathbf{C}_{cit}^*$, $q_u$ is a truncated normal distribution with parameter $z_{cit}$ and variational parameter $\boldsymbol{\beta}_{cit}^{(z)*}$, $q_\beta$ is an *M*-dimensional multivariable normal distribution with two variational parameters (mean vector $\boldsymbol{\mu}_{zi}^*$ and covariance matrix $V_{zi}^*$), and $q_{\mu,V}$ is a multivariable normal–inverse Wishart distribution with variational parameters $\boldsymbol{\mu}_i^{\mu*}, \sigma_i^{\mu*}, w_i^*, W_i^*$. Here, to realize effective variational inference, we assume that all variational parameters are independent. The update equation and the derivations of the variational parameters are detailed in Appendix A.

## 4. Simulation Study

In this section we examine the performance of the proposed VB estimator relative to MCMC using simulated data. We show that MCMC becomes too computationally demanding as the size of the dataset increases, and that VB provides a computationally efficient and accurate approximation to the posterior with good predictive properties.

The simulation datasets are generated as follows:

a) We determine the number of consumers, items and time period for the dataset.

b) Consumers randomly are assigned to a segment. Consumers assigned to same segment have a same *product set* of items, which are a subset of items that consumers evaluate when making a purchase decision. The number of consumer segments and the number of items in a product set is set to $0.02\times C$ and $0.2\times I$, respectively, where *C* and *I* are varied in our analysis.

c) Items in a product set are randomly assigned from all available items, with a lower bounds on the number of items needed in the set.

d) The marketing variables are comprised of $\mathbf{x}_{it} = [1, P_i]^T$, where $P_i$ is a price discount rate generated from a uniform distribution with interval [0.1 1.0].

e) Each consumer randomly visits the simulated shop on five occasions during data period *T*, and purchases *N* items with probability proportional to the discount from production set.

Computational time and predictive results are calculated using data simulated from the above steps. The computational results reported below were calculated in same computational environment (64-bit version of Python 2.7.5, implemented on a 3.5 GHz processor (Quad-Core Xeon; Intel Corp.) with 64 GB memory).

**4.1 Scalability**

The scalability is investigated under the condition as follows: $C$ = {1000, 5000, 10000}, $I$ = {100, 500, 1000}, $T$ = 30, $Z$ = {5, 10, 20} and $N$ = 10. Thus, 27 different scenarios were explored in the simulation study. The simulation times in hours are shown in Table 1. Here, we set hyperparameters as $\tilde{\boldsymbol{\gamma}} = [0.1, \cdots, 0.1]^T$, $\tilde{\boldsymbol{\mu}} = [0, \cdots, 0]^T$, $\tilde{\sigma}_\mu = 1$, $\tilde{W}^\beta = \mathbf{I}_M$, $\tilde{w}^\beta = 10$ in VB and MCMC. $\mathbf{I}_M$ means identity matrix of size *M*. In VB method, iterations are terminated when the variational lower bound improves by less than $10^{-5}$% of current value in two consecutive iterations. The MCMC method uses Gibbs sampler, and its simulation times of 6,000 MCMC samples are estimated from those of 10 samples, as is to be computationally infeasible. We also note that the selection of 6,000 MCMC samples is consistent with the simulation study of Braun and Mcauliffe (2010). Three kinds of settings for hyperparameters, stopping rule of VB iteration and the number of MCMC samples, defined above, are adopted in all empirical studies hereafter.

Table 1 shows the computational time for VB and MCMC methods. In both algorithms, the cost increases approximately linearly with the size of the dataset specified in term of the number of consumers, items, and latent classes. In all scenarios, the computational time of MCMC exceeds that of VB. The VB algorithm is around 20 to 50 times more efficient than MCMC, depending on the scenario. The time of estimation using large-scale data ($C$ = 10000, $I$ = 1000) by MCMC is estimated over 450 h, and thus we recognize that MCMC is not applicable for our problem.

Table 1 : Simulation time by VB and MCMC

### 4.2 Data Sparseness and RMSE of Prediction

The predictive performance of the models is investigated by simulation, focusing on whether the model can adapt to sparse data or not. Sparse data are commonly encountered in datasets containing many items. In this paper, we define the data density rate (DDR) as

$$DDR \equiv \left(C \cdot I \cdot T\right)^{-1} \sum_{c=1}^{C} \sum_{i=1}^{I} \sum_{t=1}^{T} y_{cit} \tag{13}$$

The DDR specifies the rate of $y_{cit} = 1$ events in the data space. The $N$ controls values of DDR in the simulation dataset. Here we generated datasets for $C$ = 500, $I$ = 100, $T$ = 100, $Z$ = 20, and $N$ = {2, 4, 6, 8, 10}; this specification of $N$ implies that DDR = {0.1%, 0.2%, 0.3%, 0.4%, 0.5%}. DDRs of actual scan-panel datasets are, to our knowledge, always below 1 %.

The prediction performance is measured by the root mean square error (RMSE), given by

$$RSME \equiv \sqrt{\left( I \cdot \sum_{c=1}^{C} \left|T_c\right| \right)^{-1} \sum_{c=1}^{C} \sum_{i=1}^{I} \sum_{t \in T_c} \left\{ y_{cit} - p(y_{cit} = 1) \right\}^2} \tag{14}$$

where $\left|T_c\right|$ denotes the number of elements $T_c$ for each consumer; that is, the number of store visits within the specified data period. The $p(y_{cit} = 1)$ is calculated by $\sum_{z=1}^{Z} C_{cz} F(\mathbf{x}_{it}^{T} \boldsymbol{\beta}_{zi})$.

The hold-out and hold-in samples are generated by the same procedure. Using the Gibbs sampler, we generate 6,000 samples of each parameter, where the first 5,000 are discarded as burn-in samples.

Table 2 shows the average RMSE obtained in three simulation trials of DDR for each of four methods, namely Random, Homogeneous, VB and MCMC. In this simulation, we set five levels of DDR from 0.1% to 0.5%. “Random” means the situation that consumers are permitted completely random choice of purchased items and shop visits, and its RMSE is approximately 0.577. “Homogeneous” implies the model with $Z$ = 1. We first observe that the proposed models works well compared to “Random”. Second, heterogeneity significantly improves the predictive performance. Third, the performances of VB and MCMC are comparable. These properties hold throughout every setting of DDRs.

Table 2 RMSE as a function of DDR in simulations

## 5 Empirical Application Using Customer Database

In this section, we apply the proposed model to a real large-scale customer dataset. The results of the simulation study indicate that the VB estimator provides a close approximation to the MCMC with a large improvement in computational speed. We use VB in the empirical application in this section and report on estimation results and their managerial implications.

### 5.1 Data and Variables

A customer database from a general merchandise store, recorded from April 1 to June 30 in 2002, is used in our analysis. A customer identifier, price, display, and feature were recorded for a given purchase occasion. The dataset contains 162,775 transactions involving 1,647 consumers and 1,004 items. The 1,004 selected items were displayed or featured at least once in the data period. The DDR of the scan-panel data is 0.31 %.

The marketing variables are price ( $P_{it}$ ), display ( $D_{it}$ ), and feature ( $F_{it}$ ); that is, $\mathbf{x}_{it} = \begin{bmatrix} 1 & P_{it} & D_{it} & F_{it} \end{bmatrix}^T$ . $P_{it}$ is the price relative to the maximum price of item *i* in the observational period. Display and feature are binary entries, equal to one if the item *i* is displayed or featured at time *t*, and zero otherwise.

### 5.2 Model Comparisons

The proposed models are compared in terms of RMSE. The parameters are estimated with the number of segment $Z = \{2, 3, 4, 5, 10, 20\}$. The hold-out sample comprises records from July 1 to September 30, 2002. The RMSEs for comparable models are shown in Table 3. We observe that our proposed models (*Z* greater than two) have smaller value of RMSEs than "Random" and "Homogeneous" models. The models with *Z* greater than five have the same RMSEs and thus we understand that the model with $Z = 5$ is appropriate for the empirical analysis below.

The comparison of RMSE of (i) all customers with that of (ii) infrequent customers provides useful information of the performance of our models. The largest number of purchases by one customer in our data is 390 items and 88 visits to the store, and we define infrequent customers as those with fewer than five purchases and three visits to the store. The predictive performance of infrequent consumers slightly decreases compared to that of all consumers,

however, the RMSE is almost equivalent to that all customers. Thus, our models present a tolerable prediction performance for even infrequent consumers.

Table 3 : RMSEs for real customer database – all and infrequent customers

**5.3 Segment-Level Parameter Estimates**

Table 4 displays the parameter estimates for "Price", "Display" and "Feature" and "Intercept." The rows of the tables correspond to products, and the columns to the segments. The rows are ordered in terms of the differences of estimates among the segments. The first row of each table is for the product estimates with maximum variation, and the last row of each table corresponds to estimates with minimum variation. The products positioned near the top of the table have larger heterogeneity in the response parameters, while those at bottom of the table have relatively similar values among segments. The products are identified in terms of their sales rank, with the product named "No. 1" the most purchased product in our database.

We observe that "Price" coefficients are estimated negative and "Display" and "Feature" are positive for most products. This means that our proposed model produces economically reasonable estimates. The coefficient estimates also indicate the effectiveness of the marketing mix variables for each product. In the "Price" table, for example, product No. 205 has the highest rank and consumers in segment 3 do not respond to variation in price for this product. For product No. 111, segment 1 is the least price sensitive and for product No. 153 the price insensitive segment is segment 5. Similar results are found with the display and feature portions of the table, where we see that the most responsive segment is product-specific. We also find that the lower ranked products in each table show nearly uniform response in each segment. The results imply that marketers can perform effective promotions to specific segment for higher ranked products, however, homogeneous promotions to any segments are enough for lower ranked products.

Table 4: Characteristics of $\boldsymbol{\beta}_{zi}$ for the five segments of consumers

Next, we extract relative preferences of product category for five segments. First, we define the relative preference score of segment *z* for product *i* by using estimated intercepts for

product $i$ as $RP_{zi} \equiv \beta_{zi0} - Z^{-1}\sum_{z=1}^{Z}\beta_{zi0}$. The consumers in segment $z$ with the highest value of $RP_{zi}$ relatively prefer product $i$ than consumers in other four segments. Thus, we observe preference of five segments by ordering of the score with respect to each segment.

For visualization of segment's preferences, we count the number of product category in top 100 products in order of $RP_{zi}$ with respect to each segment. Table 5 shows name and number of product categories containing over three kinds of products in the top 100 products for each segment. It discloses consumer's preferences of purchased products for estimated segments.

Table 5: Relative preferences of purchased product category for five segments

**5.4 Individual Level Parameter Estimates**

Individual level estimates of market response is obtained by taking expectation of segment level estimates with respect to $C_{cz}$

$$\boldsymbol{\alpha}_{ci} \equiv \sum_{z=1}^{Z} C_{cz}\boldsymbol{\beta}_{zi} . \tag{15}$$

We characterize individual consumer in terms of her estimated response coefficient $\boldsymbol{\alpha}_{ci}$. The empirical marginal distribution of individual consumer parameter estimates taking average $\left\{ I^{-1}\sum_{i=1}^{I}\boldsymbol{\alpha}_{ci} \right\}_{i=1,\cdots,I}$ of 1,647 products for each marketing variable are displayed by histograms in Figure 1(a). On the other hand, the empirical marginal distributions of individual product, taking average over 1004 consumers, i.e., of $\left\{ C^{-1}\sum_{c=1}^{C}\boldsymbol{\alpha}_{ci} \right\}_{c=1,\cdots,C}$, are depicted in Figure 1(b). The products that never displayed and featured in the data period have been omitted.

The marginal distributions provide reasonable individual consumer estimates since almost intercept's and price's coefficients are negative and almost display's and feature's coefficient are positive. The distributions of individual product estimates show that the feature's distribution has a sharp peak around zero, implying that promotions by display and feature are effective for many products, however, there are a lot of products that they are not

effective for the feature.

Table 6 shows the results of testing the significance of estimated response coefficients for individual consumers and products by using the 95% HPD (highest probability density) region. Owing to the space constraints, the results are shown for a portion of the dataset (five consumers and five products). "*" signifies that the 95% HPD region does not contain zero; if the HPD region includes zero (i.e., if the coefficient is insignificant), this square is left blank. We call this graph "Customer-Promotion Diagram". For example, in the "Price" diagram, the purchase of No.253 by consumers (a) and (b) is highly influenced by price; however, the price of No. 318 affects the purchasing behavior of consumers (a) only. We also report that discounting No.18 will not promote consumer (b) to purchase it. This analysis informs retailers and marketers which promotion of specified product is effective to any specific individual consumer. Thus, our proposed model enables marketers to develop effective pricing and promotional strategies for targeted consumers and products.

Table 6: Personalized effective marketing variables for individual consumers and products

Figure 1 Marginal distribution of parameter estimates of individual consumers and products

**5.5 Precision of Approximation to Posterior Density**

We examine the precision of VB by comparing estimates to those obtained with MCMC, as is assumed that MCMC is a more correct than VB as long as there are sufficient iterations to fully characterize the posterior distribution. For comparison purposes, we reduced the size of data so that MCMC computations terminate within one day. We extracted 500 customers randomly from our dataset, and choose the top 100 products in terms of sales volume.

Table 7 shows estimates of response parameters for VB methods. The vector of estimates for 100 products are displayed in row according to the order of number of purchases, that is, first row specifies the estimates for the most purchased product. The numbers in the table are sample mean of segment level estimates. "-" mean the product which has never displayed or featured in the record. From this result, we observe that price coefficients are reasonably estimated in the sense most of product have negative values. The same thing holds for the coefficients of display and features. That is, our proposed models perform well.

Table 7 : Estimated parameters by VB and MCMC

## 6. Discussion

This paper addresses two challenges in estimating models of demand in large databases: i) the large number of available products and ii) the large number of consumers who purchase these products. Existing models in marketing and methods of estimation tend to focus on a narrow set of products and a subset of consumers to understand the richness of the competitive environment within a product category among a random sample of consumers. This goal, however, is often at odds with the goals of practitioners who want to score existing datasets to identify a wide set of customers and products to allocate promotional budgets and increase sales.

We propose a descriptive model of demand based on the idea of topic models where products purchased by consumers take the place of words used by authors in creating documents. We allow for a product's purchase probability to be affected by price, display and feature advertising variables, but do not treat purchases to arise from a process of constrained utility maximization. The advantage of this approach is that it allows us to side-step complications associated with competitive effects and model a much larger set of products than that possible with existing economic models. By retaining prices and other marketing variables in our model we can still predict the effect of these variables on own-sales. This tradeoff is inevitable in the analysis of large-scale databases where purchases are tracked across thousands of products. The proposed model links the characteristics of consumer segments to marketing variables, and it is applicable to both segment-level and individual-level marketing across a large set of products.

The scalability and predictive performance of the proposed models were confirmed through a simulation study involving variational Bayes inference. In our analysis, we imposed a fairly conservative convergence criteria for VB of $10^{-5}$%, but also found that coarser thresholds (for instance, $10^{-3}$%) produced similar results. We therefore believe that estimation times can be further reduced in practice from those reported in this paper.

Finally, we employed the RMSE criterion for choosing the number of segments. In the VB framework, the variational lower bound is used for this criterion, as is shown in, for

example, Corduneanu and Bishop (2001). The variational lower bound in our models is somewhat sensitive to changes in the number of segments, *Z*. We identify this as a future research.

**Appendix A: Derivation of VB Algorithm for Proposed Model**

This appendix details the variational inference of proposed model. The update procedure derives from the analytical calculation of equation (13). The update equation for each variational parameter is obtained from the following expectation values

$$\begin{aligned}\mathbf{E}_{\neq q_j}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] &\equiv \mathbf{E}_{i\neq j}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] \\ &= \int \log p\left(\mathbf{D},\boldsymbol{\theta}\right)\prod_{i\neq j} q_i\left(\boldsymbol{\theta}^{(i)*}\right)d\boldsymbol{\theta}^{(i)*},\end{aligned} \tag{A1}$$

where $\mathbf{D}=\left\{\left\{\mathbf{x}_{it}\right\},\left\{y_{cit}\right\}\right\}$.

The update procedures of variational parameters $\mathbf{C}^*_{cit}$, $\boldsymbol{\gamma}^*_c$, $\boldsymbol{\beta}^{(z)*}_{cit}$, $\boldsymbol{\mu}^*_{iz}$, $V^*_{iz}$, $\boldsymbol{\mu}^{\mu*}_i$, $\sigma^{\mu*}_i$, $w^*_i$, and $W^*_i$ are presented below.

**A.1 Optimization of $\boldsymbol{\gamma}^*_c$**

The Dirichlet and categorical distributions are of the following forms:

$$\begin{aligned}\text{Dirichlet}\left(\mathbf{C}_c \mid \tilde{\boldsymbol{\gamma}}\right) &= \frac{\prod_{z=1}^{Z}\Gamma\left(\tilde{\gamma}_z\right)}{\Gamma\left(\sum_{z=1}^{Z}\tilde{\gamma}_z\right)}\prod_{z=1}^{Z}C_{cz}^{\tilde{\gamma}_z-1} \\ \text{Categorical}\left(z_{cit}\mid \mathbf{C}_c\right) &= \prod_{z=1}^{Z}C_{cz}^{\delta\left(z_{cit}=z\right)}\end{aligned} \tag{A2}$$

where $\Gamma(\cdot)$ is the gamma function and $\delta\left(z_{cit}=z\right)$ is the Dirac delta function defined as $\delta\left(z_{cit}=z\right)=1$ if $z_{cit}=z$, and $\delta\left(z_{cit}=z\right)=0$ otherwise. The expectation value $\mathbf{E}_{\neq q_c}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right]$ is then calculated for each $c$ as

$$\begin{aligned}&\mathbf{E}_{\neq q_c}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] = \log p\left(\mathbf{C}_c\right)+\mathbf{E}_{q_z}\left[\log p\left(\left\{z_{cit}\right\}\mid \mathbf{C}_c\right)\right]+\text{const.} \\ &= \log\Gamma\left(\sum_{z-1}^{Z}\tilde{\gamma}_z\right)-\sum_{z=1}^{Z}\log\Gamma\left(\tilde{\gamma}_z\right)+\sum_{z=1}^{Z}\left[\left(\tilde{\gamma}_z+\sum_{i\in I_c}\sum_{t\in T_c}C^*_{citz}-1\right)\right]\log C_{cz}+\text{const.}\end{aligned} \tag{A3}$$

Here and hereafter, *const.* denotes any terms not included in the relevant parameters. The second line of the above equations describes a log-Dirichlet function with parameter $\tilde{\gamma}_z+\sum_{i\in I_c}\sum_{t\in T_c}C^*_{citz}$. Therefore,

$$\boldsymbol{\gamma}^{*(\text{new})}_c \leftarrow \tilde{\boldsymbol{\gamma}}+\sum_{i\in I_c}\sum_{t\in T_c}C^*_{citz} \tag{A4}$$

**A2. Optimization of $C^*_{citz}$**

Here we denote a digamma function as $\Psi(\cdot)$, which will be useful for later discussion, and summarize the property of truncated normal distribution in the probit model. $u_{cit}^{(z)}$

follows a normal distribution with mean $\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}$ and variance 1. Moreover, $u_{cit}^{(z)}$ must satisfy $y_{cit}=1$ if $u_{cit}>0$ and $y_{cit}=0$ if $u_{cit}\le 0$. Therefore, $u_{cit}^{(z)}$ is generated from a truncated normal distribution as

$$u_{cit}^{(z)} \sim \begin{cases} TN_{(0,\infty)}\left(\mathbf{x}_{it}^T\boldsymbol{\beta}_{iz},1\right) & \text{if } y_{cit}=1 \\ TN_{(-\infty,0)}\left(\mathbf{x}_{it}^T\boldsymbol{\beta}_{iz},1\right) & \text{if } y_{cit}=0 \end{cases}. \tag{A5}$$

where $TN_{(n_1,n_2)}(\cdot,\cdot)$ denotes a normal distribution truncated from $n_1$ to $n_2$. The distribution of $u_{cit}^{(z)}$ is therefore expressed as

$$p\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{zi}, z_{cit}, \mathbf{x}_{it}, y_{cit}\right) = \frac{1}{\Omega_{cit}^{(z)}}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(u_{cit}^{(z)}-\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right)^2\right\} \tag{A6}$$

with $\Omega_{cit}^{(z)} \equiv \left\{1-F\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right)\right\}^{\delta(y_{cit}=1)}\left\{F\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right)\right\}^{\delta(y_{cit}=0)}$. In addition, the expectation value and variance are expressed as

$$\begin{aligned} \mathbf{E}\left[u_{cit}^{(z)}\right] &= \mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*} + \varphi_{cit}^{(z)} \\ \mathrm{Var}\left[u_{cit}^{(z)}\right] &= 1-\mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*}\varphi_{cit}^{(z)} - \left(\varphi_{cit}^{(z)}\right)^2 \end{aligned} \tag{A7}$$

where $\varphi_{cit}^{(z)} \equiv \left(\frac{f\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*}\right)}{1-F\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*}\right)}\right)^{\delta(y_{cit}=1)}\left(-\frac{f\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*}\right)}{F\left(-\mathbf{x}_{it}^T\boldsymbol{\beta}_{cit}^{(z)*}\right)}\right)^{\delta(y_{cit}=0)}$.

Thus, the expected value $\mathbf{E}_{q_z}\left[\log p(\mathbf{D},\boldsymbol{\theta})\right]$ is given as

$$\begin{aligned} \mathbf{E}_{\neq q_z}\left[\log p(\mathbf{D},\boldsymbol{\theta})\right] &= \mathbf{E}_{q_c}\left[\log p\left(z_{cit} \mid \mathbf{C}_c\right)\right] \\ &\quad + \mathbf{E}_{q_u,q_\beta}\left[\log p\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{zi}, z_{cit}, \mathbf{x}_{it}, y_{cit}\right)\right] + \text{const.} \end{aligned} \tag{A8}$$

The first term in the right hand side of Equation (A8) is obtained as $\Psi\left(\gamma_{cz}^*\right) - \Psi\left(\sum_{z-1}^{Z}\gamma_{cz}^*\right)$ (Blei et al. 2003), while the second term is evaluated as

$$\begin{aligned} &\mathbf{E}_{q_u,q_\beta}\left[\log p\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{zi}, z_{cit}, \mathbf{x}_{it}, y_{cit}\right)\right] = \mathbf{E}_{q_u,q_\beta}\left[-\log\sqrt{2\pi}\Omega_{cit}^{(z)} - \frac{1}{2}\left(u_{cit}^{(z)}-\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right)^2\right] \\ &= -\mathbf{E}_{q_\beta}\left[\log\Omega_{cit}^{(z)}\right] - \frac{1}{2}\mathbf{E}_{q_u}\left[\left(u_{cit}^{(z)}\right)^2\right] + \mathbf{E}_{q_u,q_\beta}\left[u_{cit}^{(z)}\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right] - \frac{1}{2}\mathbf{E}_{q_\beta}\left[\left(\mathbf{x}_{it}^T\boldsymbol{\beta}_{zi}\right)^2\right] + \text{const.} \end{aligned} \tag{A9}$$

To solve Equation (A8) for $C_{citz}^*$, we must evaluate the four terms of Equation (A9). The first term includes a CDF from which the expectation value is difficult to obtain analytically. Thus,

we expand the term as a first-order Taylor expansion in terms of the CDF of normal distribution and the logarithm function. In addition, we assume that the expectation of the third term can be approximated by a linear approximation. Such bold approximations are standard strategies for adapting topic models with VB to practical computation (for examples, zeroth-order Taylor approximation by Asuncion et al. (2009) and Sato and Nakagawa (2012), and zeroth and first order delta approximation by Braun and McAuliffe (2010)). The four expectation values in Equation (A9) are then written as

$$\begin{aligned}
&\mathbf{E}_{q_\beta}\left[\log \Omega_{cit}^{(z)}\right] \approx -\frac{1}{2} + (-1)^{\delta(y_{cit}=0)} \frac{\mathbf{x}_{it}^T \boldsymbol{\mu}_{zi}^*}{\sqrt{2\pi}}, \\
&\mathbf{E}_{q_u}\left[\left(u_{cit}^{(z)}\right)^2\right] = \mathrm{var}\left[u_{cit}^{(z)}\right] + \left(\mathbf{x}_{it}^T \boldsymbol{\beta}_{cit}^{(z)*} + \varphi_{cit}^{(z)}\right)^2, \\
&\mathbf{E}_{q_u, q_\beta}\left[u_{cit}^{(z)} \mathbf{x}_{it}^T \boldsymbol{\beta}_{zi}\right] \approx \left(\mathbf{x}_{it}^T \boldsymbol{\beta}_{cit}^{(z)*} + \varphi_{cit}^{(z)}\right)\left(\mathbf{x}_{it}^T \boldsymbol{\mu}_{zi}^*\right), \\
&E_{q_\beta}\left[\left(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi}\right)^2\right] = \mathbf{x}_{it}^T V_{zi}^{\beta *} \mathbf{x}_{it} + \left(\mathbf{x}_{it}^T \boldsymbol{\mu}_{zi}^*\right)^2.
\end{aligned} \tag{A10}$$

In appendix A.3, we find that $\boldsymbol{\mu}_{zi}^*$ is updated to the optimized value of $\boldsymbol{\beta}_{cit}^{(z)*}$. Finally, $C_{citz}^*$ is updated as

$$C_{citz}^* \leftarrow \frac{\exp\left(\rho_{citz}\right)}{\sum_{z=1}^{Z} \exp\left(\rho_{citz}\right)}, \tag{A11}$$

where

$$\rho_{citz} = \Psi\left(\gamma_{cz}^*\right) - \Psi\left(\sum\nolimits_{z-1}^{Z} \gamma_{cz}^*\right) - \mathbf{E}_{q_\beta}\left[\log \Omega_{cit}^{(z)}\right] + \frac{1}{2} \mathbf{x}_{it}^T \boldsymbol{\mu}_{zi}^* \varphi_{cit}^{(z)} + \mathbf{x}_{it}^T V_{zi}^{\beta *} \mathbf{x}_{it}. \tag{A12}$$

### A.3 Optimization of $\boldsymbol{\beta}_{cit}^{(z)*}$

Similar to equations (A3) and (A9), the expected value that optimizes $\boldsymbol{\beta}_{cit}^{(z)*}$ is

$$\begin{aligned}
\mathbf{E}_{\neq q_u}\left[\log p\left(\mathbf{D}, \boldsymbol{\theta}\right)\right] &= \mathbf{E}_{q_z, q_\beta}\left[\log p\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{zi}, z_{cit}, \mathbf{x}_{it}, y_{cit}\right)\right] + \text{const.} \\
&\propto \exp\left\{-\frac{1}{2C_{citz}^*}\left(u_{cit}^{(z)} - \mathbf{x}_{it}^T \boldsymbol{\mu}_{zi}^*\right)^2\right\} + \text{const.}
\end{aligned} \tag{A13}$$

Here we seek the mean vector of the truncated normal distribution of $u_{cit}^{(z)}$. Therefore, the update equation becomes

$$\boldsymbol{\beta}_{cit}^{(z)*} \leftarrow \boldsymbol{\mu}_{zi}^*. \tag{A14}$$

**A.4 Optimization of $\boldsymbol{\mu}_{zi}^{*}$ and $V_{zi}^{*}$**

First, we derive an inverse Wishart distribution function and adopt some well-known properties of multivariable normal and inverse Wishart distributions (Anderson 2003, Bishop 2006).

$$\begin{aligned}
&\mathrm{IW}\left(\tilde{W},\tilde{w}\right)=\frac{\left|\tilde{W}\right|^{\tilde{w}/2}}{2^{\tilde{w}M}\,\Gamma(\tilde{w}/2)}\left|V_i\right|^{-\frac{\tilde{w}+M+1}{2}}\exp\left\{-\frac{1}{2}tr\left(\tilde{W}V_i^{-1}\right)\right\},\\
&\mathbf{E}_{q_V}\left[\log\left|V_i\right|\right]=\sum_{m=1}^{M}\Psi\left(\frac{w_i^{*}+1-m}{2}\right)+M\log 2+\log\left|W_i^{*-1}\right|,\\
&\mathbf{E}_{q_V}\left[V_i^{-1}\right]=w_i^{*}W_i^{*-1},\\
&E_{q_\mu,q_V}\left[\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i\right)V_i^{-1}\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i\right)^{T}\right]=\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i^{\mu^*}\right)^{T}w_i^{*}W_i^{*-1}\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i^{\mu^*}\right)+\sigma_i^{\mu^*}.
\end{aligned}\tag{A15}$$

We obtain the optimization procedures of $\boldsymbol{\mu}_{iz}^{*}$ and $V_{iz}^{*}$ by the following expected value:

$$\begin{aligned}
\mathbf{E}_{\neq q_\beta}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right]&=\mathbf{E}_{q_\mu,q_V}\left[\log p\left(\boldsymbol{\beta}_{zi}\mid\boldsymbol{\mu}_i,V_i\right)\right]\\
&\quad+\mathbf{E}_{q_u,q_z}\left[\log p\left(\left\{u_{cit}^{(z)}\right\}\mid\boldsymbol{\beta}_{zi},\left\{z_{cit},\mathbf{x}_{it},y_{cit}\right\}\right)\right]+\text{const.}\\
&=-\frac{1}{2}\mathbf{E}_{q_\mu,q_V}\left[\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i\right)^{T}V_i^{-1}\left(\boldsymbol{\beta}_{zi}-\boldsymbol{\mu}_i\right)\right]\\
&\quad-\frac{1}{2}\sum_{c=1}^{C}\sum_{t\in T_c}\mathbf{E}_{q_u,q_z}\left[\left(u_{cit}^{(z)}-\mathbf{x}_{it}^{T}\boldsymbol{\beta}_{zi}\right)^{2}\right]+\text{const.}
\end{aligned}\tag{A16}$$

The first and second terms of the second line are given by the last and third lines of equation (A10), while the third and fourth terms are given by Equations (A2) and (A3), respectively, derived in a manner similar to (A9). $\boldsymbol{\mu}_{iz}^{*}$ and $V_{zi}^{*}$ are then arithmetically updated as

$$\begin{aligned}
\boldsymbol{\mu}_{zi}^{*}&\leftarrow\left\{w_i^{*}W_i^{*-1}+X_{zi}X_i^{T}\right\}^{-1}\left\{w_i^{*}W_i^{*-1}\boldsymbol{\mu}_i^{\mu^*}+X_{zi}\overline{\mathbf{u}}_{zi}\right\}\\
V_{zi}^{*}&\leftarrow\left\{w_i^{*}W_i^{*-1}+X_{zi}X_i^{T}\right\}^{-1}
\end{aligned}\tag{A17}$$

where

$$\overline{\mathbf{u}}_{zi}\equiv\left[\left\{\mathbf{E}\left[u_{cit}^{(z)}\right]\right\}_{c=1,\cdots C,t\in T_c}\right]^{T},\ X_i\equiv\left[\left\{\mathbf{x}_{it}\right\}_{c=1,\cdots C,t\in T_c}\right],\ X_{zi}\equiv\left[\left\{C_{citz}^{*}\mathbf{x}_{it}\right\}_{c=1,\cdots C,t\in T_c}\right].$$

The $\overline{\mathbf{u}}_{zi}$ is vector and $X_i$ and $X_{zi}$ are matrices. The number of elements in $\overline{\mathbf{u}}_i$, $X_i$ and $X_{zi}$ are decided by the size of the consumer base and by $T_c$.

**A.5 Optimization of $\boldsymbol{\mu}_i^{\mu^*}$, $\sigma_i^{\mu^*}$, $w_i^{*}$, and $W_i^{*}$**

Here we consider a joint distribution of a multivariable normal distribution of $\boldsymbol{\mu}_i$ and an inverse Wishart distribution of $V_i$, and derive the update equations for four types of

variational parameters from this joint distribution. To this end, we require the following expectation value from the joint distribution function:

$$\begin{aligned}
&\mathbf{E}_{\neq q_\mu, q_V}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] = \log p\left(\boldsymbol{\mu}_i, V_i\right) + E_{q_\beta}\left[\log p\left(\{\boldsymbol{\beta}_{zi}\} \mid \boldsymbol{\mu}_i, V_i\right)\right] + \text{const.} \\
&= -\frac{1}{2}\log|V_i| - \frac{1}{2}\tilde{\sigma}_\mu^{-1}\left(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}^\mu\right)^T V_i^{-1}\left(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}^\mu\right) - \frac{\tilde{w}+M+1}{2}\log|V_i| - \frac{1}{2}tr\left\{\tilde{W}V_i^{-1}\right\} \\
&-\frac{1}{2}Z\cdot E_{q_\beta}\left[\log|V_i|\right] - \frac{1}{2}\sum_{z=1}^{Z} E_{q_\beta}\left[\left(\boldsymbol{\mu}_i - \boldsymbol{\beta}_{zi}\right)^T V_i^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\beta}_{zi}\right)\right] + \text{const.}
\end{aligned} \tag{A18}$$

First, we extract from this expectation value all terms linked to multivariable variational parameters $\boldsymbol{\mu}_i^{\mu^*}$ and $\sigma_i^{\mu^*}$; that is

$$\begin{aligned}
\mathbf{E}_{\neq q_\mu}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] = &-\frac{1}{2}\tilde{\sigma}_\mu^{-1}\left(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}^\mu\right)^T V_i^{-1}\left(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}^\mu\right) \\
&-\frac{1}{2}\sum_{z=1}^{Z} E_{q_\beta}\left[\left(\boldsymbol{\mu}_i - \boldsymbol{\beta}_{zi}\right)^T V_i^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\beta}_{zi}\right)\right] + \text{const.}
\end{aligned} \tag{A19}$$

The second term in the above equation is obtained in the same manner as (A15). The multivariable normal distribution function is then constructed in a straightforward manner as follows:

$$\begin{aligned}
\boldsymbol{\mu}_i^{\mu^*} &\leftarrow \left(\tilde{\sigma}_\mu^{-1} + Z\right)^{-1}\left(\tilde{\sigma}_\mu^{-1}\tilde{\boldsymbol{\mu}}^\mu + \sum_{z=1}^{Z}\boldsymbol{\mu}_{zi}^*\right), \\
\sigma_i^{\mu^*} &\leftarrow \left(\tilde{\sigma}_\mu^{-1} + Z\right)^{-1}.
\end{aligned} \tag{A20}$$

Next, we optimize $w_i^*$ and $W_i^*$ using Equation (A15) and the relationship $\log q\left(V_i\right) = \log q\left(\boldsymbol{\mu}_i, V_i\right) - \log q\left(\boldsymbol{\mu}_i \mid V_i\right)$.

$$\mathbf{E}_{\neq q_V}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] = \mathbf{E}_{\neq q_\mu, q_V}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] - \mathbf{E}_{\neq q_\mu}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right] \tag{A21}$$

The expectation value $\mathbf{E}_{\neq q_V}\left[\log p\left(\mathbf{D},\boldsymbol{\theta}\right)\right]$ is calculated in a straightforward manner by using (A16) and (A17). Finally, we obtain the update equations for $w_i^*$ and $W_i^*$ as

$$\begin{aligned}
W_i^* &\leftarrow \tilde{W} + \sum_{z=1}^{Z} V_{zi}^* + \tilde{\sigma}_\mu^{-1}\tilde{\boldsymbol{\mu}}^\mu\tilde{\boldsymbol{\mu}}^{\mu T} + \sum_{z=1}^{Z}\boldsymbol{\mu}_{zi}^*\boldsymbol{\mu}_{zi}^{*T} - \left(\tilde{\sigma}_\mu^{-1} + Z\right)\boldsymbol{\mu}_i^{\mu^*}\boldsymbol{\mu}_i^{\mu^* T}, \\
V_i^* &\leftarrow \tilde{w} + Z.
\end{aligned} \tag{A(22)}$$

Notice that $\sigma_i^{\mu^*}$ and $w_i^*$ are constant if the hyperparameters and the number latent class are given.

**Appendix B: Gibbs Sampler**

The joint posterior distribution, assuming conditional independence between variables,

provides the full conditional posterior distributions:

$$\begin{aligned}
&\mathbf{C}_c \mid -\sim p\left(\mathbf{C}_c \mid z_{cit}\right)\\
&z_{cit} \mid -\sim p\left(z_{cit} \mid \mathbf{C}_c, \{\boldsymbol{\beta}_{zi}\}, \{\mathbf{x}_{it}\}, \{y_{cit}\}\right)\\
&u_{cit}^{(z)} \mid -\sim p\left(u_{cit}^{(z)} \mid z_{cit}, \boldsymbol{\beta}_{zi}, \mathbf{x}_{it}, y_{cit}\right)\\
&\boldsymbol{\beta}_{zi} \mid -\sim p\left(\boldsymbol{\beta}_{zi} \mid \{u_{cit}^{(z)}\}, \boldsymbol{\mu}_i, V_i, \{\mathbf{x}_{it}\}\right)\\
&\boldsymbol{\mu}_i \mid -\sim p\left(\boldsymbol{\mu}_i \mid \{\boldsymbol{\beta}_{zi}\}, V_i\right)\\
&V_i \mid -\sim p\left(V_i \mid \{\boldsymbol{\beta}_{zi}\}, \boldsymbol{\mu}_i\right)
\end{aligned} \tag{C1}$$

where *TN* denotes a truncated normal distribution.

**B.1 Sampling of $\mathbf{C}_c$**

The $\mathbf{C}_c$ is generated by a Dirichlet categorical relation. The Dirichlet distribution is a conjugate prior of a categorical distribution. For each consumer *c*, $\mathbf{n}_c = \left[n_{c1}, \cdots, n_{cZ}\right]^T$ denotes the number of generated latent classes $z_c$ by categorical distribution of parameter $\mathbf{C}_c$ in each MCMC step. A Dirichlet categorical relation gives the posterior distribution with respect to $\mathbf{C}_c$ as

$$p\left(\mathbf{C}_c \mid -\right) = p\left(\mathbf{C}_c\right) p\left(z_c \mid \mathbf{C}_c\right) = \text{Diriclet}\left(\mathbf{n}_c + \tilde{\gamma}\right) \tag{C2}$$

**B.2 Sampling of $z_{cit} \mid -$**

The posterior probability of ($z_{cit} = j$) is given as

$$\Pr\left\{z_{cit} = j \mid \mathbf{C}_c, \left\{\mathbf{x}_{it}\right\}, \left\{\boldsymbol{\beta}_{zi}\right\}, \left\{y_{cit}\right\}\right\} = \frac{C_{cj} A_{citj}}{\sum_{z=1}^{Z} C_{cz} A_{citz}}, \tag{C3}$$

where $A_{citz} = F\left(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi}\right)^{y_{cit}} \left\{1 - F\left(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi}\right)\right\}^{1-y_{cit}}$.

**B.3 Sampling of $u_{cit}^{(z)} \mid -$**

The distribution of $u_{cit}^{(z)}$ is described in Appendix A.2. $u_{cit}^{(z)}$ is sampled from a truncated normal distribution in Equation (A5). This well-known sampling approach is called data augmentation (Tanner, 1987).

**B.4 Sampling of $\boldsymbol{\beta}_{zi}$, $\boldsymbol{\mu}_i$, and $V_i$**

The full conditional posterior distribution of $\boldsymbol{\beta}_{iz}$, $\boldsymbol{\mu}_i$, and $V_i$ is derived from a hierarchical linear regression model. In our case, $\boldsymbol{\beta}_{zi}$ for each *i* and each *z* is sampled from

$$\boldsymbol{\beta}_{iz} \sim N_M\left(R^{-1}\left\{\left(\bar{X}_{zi}^T \mathbf{u}_{zi}^{(z)}\right) + V_i^{-1} \boldsymbol{\mu}_i\right\}, R^{-1}\right), \tag{C4}$$

where $R \equiv \bar{X}_{zi}^T \bar{X}_{zi} + V_i^{-1}$, $\mathbf{u}_{zi}^{(z)} \equiv \left[ \left\{ u_{cit}^{(z)} \right\}_{c \in z_c = z, t \in T_c} \right]^T$ and $\bar{X}_{zi} \equiv \left[ \left\{ \mathbf{x}_{it} \right\}_{c \in z_c = z, t \in T_c} \right]^T$.

$\boldsymbol{\mu}_i$ is sampled from

$$\boldsymbol{\mu}_i \sim N_M \left( \left( Z + \tilde{\sigma}_\mu \right)^{-1} \sum_{z=1}^{Z} \boldsymbol{\beta}_{zi}, V_i + \left( Z + \tilde{\sigma}_\mu \right)^{-1} \mathbf{I}_M \right), \tag{C5}$$

for each *i*. Here, the hyperparameters are set to $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$.

Finally, $V_i$ for each *i* is sampled from

$$V_i \sim IW \left( \tilde{w} + Z, \tilde{W} + B^T B \right), \tag{C6}$$

where $B \equiv \sum_{z=1}^{Z} \left( \boldsymbol{\beta}_{zi} - Z^{-1} \sum_{z=1}^{Z} \boldsymbol{\beta}_{zi} \right)$.

## References


[1] Ansari, A., and Mela, C. F. (2003). "E-Customization". *Journal of Marketing Research.* 40, 131-145.

[2] Asuncion, A., Welling, M., Smyth, P. and Teh, Y.W. (2009). "On Smoothing and Inference for Topic Models" *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 27-34.

[3] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, U.S.A.

[4] Blattberg, R.C., Kim, B.D., and Neslin, S.A. (2009). *Database Marketing: Analyzing and Managing Customers*, Springer: PA.

[5] Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). "Latent Dirichlet Allocation," *Journal of Machine Learning Research*. 3, 993-1022.

[6] Blei, D., and McAuliffe. J. (2007). "Supervised Topic Models," *Proceedings of Neural Information Processing System.*

[7] Braun, M., and McAuliffe. J. (2010). "Variational Inference for Large-Scale Models of Discrete Choice," *Journal of the American Statistical Association*, 105, 324-335.

[8] Chintagunta, P.K., and Nair. H.S. (2011). "Discrete-Choice Models of Consumer Demand in Marketing," *Marketing Science*, 30, 977-996.

[9] Chung, T.S., Rust, R., and Wedel. M. (2009). "My Mobile Music: An Adaptive Personalization System for Digital Audio Players," *Marketing Science*, 28, 52-68.

[10] Corduneanu, A., and Bishop, C.M. (2001). "Variational Bayesian Model Selection for Mixture Distributions. In: Jaakkola, T., Richardson, T. (Eds.)", *Artificial Intelligence and Statistics*, Morgan Kaufmann: Los Altos, CA, 27–34.

[11] Grimmer, J. (2011). "An Introduction to Bayesian Inference via Variational Approximations," *Political Analysis*, 19, 32-47.

[12] Ishigaki, T., Takenaka T., and Motomura. Y. (2010). "Category Mining by Heterogeneous Data Fusion Using PdLSI Model in a Retail Service," *Proceeding of IEEE International Conference on Data Mining*, 857-862

[13] Iwata, T., Watanabe, S., Yamada, and T., Ueda, N,. (2009). "Topic Tracking Model for Analyzing Consumer Purchase Behavior," *Proceeding of International Joint Conference on Artificial Intelligence*, 1427-1432.

[14] Iwata, T., and Sawada, H., (2012). “Topic Model for Analyzing Purchase Data with Price Information,” *Data Mining and Knowledge Discovery*, 26, 559-573.

[15] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An Introduction to Variational Methods for Graphical Models,” *Machine Learning*, 37, 183-233.

[16] Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W., Kreulen, J., Lenk, P., Madigan and D.M., Montgomery, A. (2008). “Challenges and opportunities in high-dimensional choice data analyses,” *Marketing Letter*, 19, 201-213.

[17] Ramage, D., Hall, D., Nallapati, R., and Manning, C.D. (2009). “Labeled LDA: a Supervised Topic Model for Credit Attribution in Multi-labeled Corpora,” *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248-256.

[18] Rossi, P.E., Allenby, G.M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons: Chichester, UK.

[19] Rust, R.T. and Chung, T.S. (2005). “Marketing Models of Service and Relationships,” *Marketing Science*, 25, 560–580.

[20] Spirling, A. and Quinn, K. (2010). “Identifying Intraparty Voting Blocs in the U.K. House of Commons”, *Journal of the American Statistical Association*, 105, 447-457.

[21] Sato, I. and Nakagawa, H. (2012). “Rethinking Collapsed Variational Bayes Inference for LDA,” *Proceedings of International Conference on Machine Learning*, 999-1006.

[22] Tanner, M.A. and Wong, W.H. (1987). “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistics Association*, 82, 528-540.

[23] Teh, Y.W. and M. I. Jordan. (2010). “Hierarchical Bayesian nonparametric models with applications,” *Bayesian Nonparametrics: Principles and Practice*, eds N. Hjort, C. Holmes, P. Mueller, and S. Walker, Cambridge University Press, Cambridge, UK:, 2010.

[24] Tsiptsis, K. and Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation*, Wiley: UK.

[25] Wedel, M. and Kamakura, W.A. (1999). *Market Segmentation: Conceptual and Methodological Foundations.* Kluwer Academic Publishers: U.S.A.

Table 1 Simulation time by VB and MCMC

| | | VB | | | MCMC | | |
|---|---|---|---|---|---|---|---|
| | Z | 5 | 10 | 20 | 5 | 10 | 20 |
| *I* | *C* =1000 | | | | | | |
| | 100 | 0.2 | 0.3 | 0.4 | 4.3 | 6.0 | 12.2 |
| | 500 | 0.4 | 0.7 | 1.2 | 17.6 | 22.8 | 32.1 |
| | 1000 | 0.7 | 0.9 | 1.5 | 37.8 | 41.6 | 48.9 |
| | *C* = 5000 | | | | | | |
| | 100 | 0.6 | 0.9 | 1.5 | 17.8 | 23.3 | 35.8 |
| | 500 | 1.3 | 2.1 | 3.9 | 50.5 | 62.6 | 80.9 |
| | 1000 | 2.4 | 4.1 | 6.4 | 99.6 | 111.0 | 128.2 |
| | *C* = 10000 | | | | | | |
| | 100 | 1.5 | 2.1 | 3.7 | 38.8 | 52.3 | 78.5 |
| | 500 | 3.3 | 5.0 | 8.4 | 164.3 | 201.0 | 264.0 |
| | 1000 | 5.9 | 8.6 | 14.7 | 333.3 | 372.7 | 450.2 |

The number means hour.

Table 2 RMSE as a function of DDR in simulations

| | DDR | | | | |
|---|---|---|---|---|---|
| | 0.1% | 0.2% | 0.3% | 0.4% | 0.5% |
| Random | 0.577 | | | | |
| Homogeneous | 0.245 | 0.236 | 0.238 | 0.236 | 0.237 |
| VB | 0.227 | 0.213 | 0.212 | 0.211 | 0.212 |
| MCMC | 0.226 | 0.214 | 0.211 | 0.212 | 0.212 |

Simulated data ($C$ = 500, $I$ = 100).

Table 3 RMSEs for real customer database – all and infrequent customers

| | # of *Z* | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 10 | 20 |
| Random | 0.577 | | | | | |
| Homogeneous | 0.412 | | | | | |
| All customers | 0.404 | 0.389 | 0.385 | 0.383 | 0.383 | 0.383 |
| Infrequent customers | 0.410 | 0.400 | 0.395 | 0.393 | 0.393 | 0.393 |

Real customer database ($C$ = 1647, $I$ = 1004)

Table 4 Characteristics of $\boldsymbol{\beta}_{zi}$ for the five segments of consumers

| No. | Intercept | | | | |
|---|---|---|---|---|---|
| | segment 1 | segment 2 | segment 3 | segment 4 | segment 5 |
| 111 | -0.82 | 3.56 | 4.17 | 3.76 | 4.09 |
| 205 | 2.38 | 2.48 | -1.60 | 2.68 | 2.57 |
| 153 | 1.44 | 1.25 | 1.48 | 1.52 | -1.46 |
| 253 | -1.42 | -3.19 | -1.43 | -1.48 | -1.52 |
| 120 | -0.46 | -0.50 | -1.34 | -0.41 | -1.85 |
| ... | ... | ... | ... | ... | ... |
| 853 | -0.89 | -0.90 | -0.90 | -0.90 | -0.90 |
| 1002 | -0.73 | -0.72 | -0.73 | -0.73 | -0.73 |
| 822 | -0.03 | -0.03 | -0.03 | -0.02 | -0.02 |
| 479 | -0.13 | -0.13 | -0.13 | -0.12 | -0.13 |
| 166 | 0.05 | 0.06 | 0.06 | 0.06 | 0.06 |

| No. | Price | | | | |
|---|---|---|---|---|---|
| | segment 1 | segment 2 | segment 3 | segment 4 | segment 5 |
| 205 | -4.38 | -4.53 | -0.13 | -4.74 | -4.59 |
| 111 | -1.30 | -5.04 | -5.06 | -5.21 | -5.89 |
| 153 | -3.95 | -3.08 | -4.08 | -4.08 | -0.24 |
| 120 | -2.51 | -2.47 | -3.75 | -2.66 | -0.15 |
| 147 | -2.13 | -2.02 | -2.10 | -1.94 | -4.71 |
| ... | ... | ... | ... | ... | ... |
| 764 | -0.52 | -0.53 | -0.52 | -0.53 | -0.53 |
| 608 | -0.63 | -0.64 | -0.63 | -0.63 | -0.63 |
| 479 | -0.03 | -0.03 | -0.03 | -0.03 | -0.04 |
| 556 | -0.42 | -0.42 | -0.43 | -0.42 | -0.42 |
| 737 | -0.57 | -0.56 | -0.57 | -0.56 | -0.57 |

| No. | Display | | | | |
|---|---|---|---|---|---|
| | segment 1 | segment 2 | segment 3 | segment 4 | segment 5 |
| 195 | 1.26 | 1.32 | 1.29 | 1.31 | 2.74 |
| 18 | 1.35 | 0.14 | 0.11 | 0.14 | 0.12 |
| 182 | 1.28 | 1.18 | 1.21 | 2.27 | 1.22 |
| 225 | 1.57 | 0.62 | 0.61 | 0.63 | 1.33 |
| 47 | 0.66 | 1.59 | 0.66 | 0.62 | 0.58 |
| ... | ... | ... | ... | ... | ... |
| 936 | -0.09 | -0.08 | -0.10 | -0.10 | -0.10 |
| 483 | -0.07 | -0.07 | -0.06 | -0.06 | -0.07 |
| 871 | 0.32 | 0.32 | 0.33 | 0.33 | 0.33 |
| 181 | 0.30 | 0.31 | 0.31 | 0.29 | 0.31 |
| 794 | 0.70 | 0.71 | 0.70 | 0.71 | 0.70 |

| No. | Feature | | | | |
|---|---|---|---|---|---|
| | segment 1 | segment 2 | segment 3 | segment 4 | segment 5 |
| 147 | 1.06 | 0.96 | 1.01 | 0.91 | 2.71 |
| 96 | 2.54 | 0.90 | 0.92 | 0.93 | 0.96 |
| 558 | 1.02 | 1.90 | 2.51 | 0.99 | 1.03 |
| 502 | 1.03 | 1.11 | 1.19 | 2.53 | 1.12 |
| 163 | 0.57 | 0.50 | 1.73 | 0.58 | 0.56 |
| ... | ... | ... | ... | ... | ... |
| 457 | 0.23 | 0.22 | 0.21 | 0.21 | 0.23 |
| 907 | 0.03 | 0.04 | 0.05 | 0.03 | 0.04 |
| 24 | 0.06 | 0.05 | 0.06 | 0.05 | 0.06 |
| 187 | 0.20 | 0.20 | 0.20 | 0.20 | 0.21 |
| 166 | 0.05 | 0.06 | 0.06 | 0.06 | 0.06 |

Table 5 Relative preferences of purchased product category for five segments

| Segment 1 | # | Segment 2 | # | Segment 3 | # |
|---|---|---|---|---|---|
| Dessert | 7 | Dessert | 6 | Yoghurt | 6 |
| Instant noodle | 4 | Dry noodle | 4 | Noodle | 4 |
| Dressing | 4 | Chocolate | 4 | Chocolate | 4 |
| Cookie | 4 | Snak | 4 | Coke | 4 |
| Coffee | 3 | Snacks made from rice | 4 | Dressing | 3 |
| Sauce | 3 | Instant noodle | 3 | Detergent | 3 |
| Yoghurt | 3 | Coke | 3 | Softener | 3 |
| Fish sausage | 3 | Detergent | 3 | Fresh noodle | 3 |
| Dry noodle | 3 | Fish sausage | 3 | | |
| Frozen noodle | 3 | Milk | 3 | | |
| | | Soy sauce | 3 | | |

| Segment 4 | # | Segment 5 | # |
|---|---|---|---|
| Instant noodle | 6 | Yoghurt | 8 |
| Detergent | 4 | Instant noodle | 6 |
| Milk | 4 | Coke | 4 |
| Fresh noodle | 4 | Fish sausage | 4 |
| Japanese tea | 4 | Milk | 4 |
| Coffee | 3 | Tea | 3 |
| Coke | 3 | Snacks made from rice | 3 |
| Dessert | 3 | Japanease Tea | 3 |
| Yoghurt | 3 | Beans paste | 3 |
| Dry noodle | 3 | Detergent | 3 |
| Fizzy drink | 3 | | |

Table 6 Personalized effective marketing variables for individual consumers and products.

| Price | | | | | | |
|---|---|---|---|---|---|---|
| | | Product No. | | | | |
| | | 18 | 110 | 253 | 318 | 742 |
| Customer | (a) | * | * | * | * | * |
| | (b) | * | * | * | | * |
| | (c) | * | * | | | * |
| | (d) | * | * | | | * |
| | (e) | * | * | | * | * |

| Display | | | | | | |
|---|---|---|---|---|---|---|
| | | Product No. | | | | |
| | | 18 | 110 | 253 | 318 | 742 |
| Customer | (a) | * | * | * | | * |
| | (b) | | * | * | | * |
| | (c) | * | * | * | | |
| | (d) | * | * | * | | * |
| | (e) | * | * | * | | |

| Feature | | | | | | |
|---|---|---|---|---|---|---|
| | | Product No. | | | | |
| | | 18 | 110 | 253 | 318 | 742 |
| Customer | (a) | * | * | | * | |
| | (b) | | * | | * | |
| | (c) | | * | | * | |
| | (d) | | * | | * | |
| | (e) | | * | | * | |

Squares marked with * indicate that a consumer (row) is likely to purchase a product (column) based on the marketing variable (price, display, or feature)

Table 7 Estimated parameters by VB and MCMC

| No. | Intercept | | Price | | Display | | Feature | |
|---|---|---|---|---|---|---|---|---|
| | VB | MCMC | VB | MCMC | VB | MCMC | VB | MCMC |
| 1 | 3.29 | 1.35 | -4.91 | -3.42 | - | - | 0.17 | 0.44 |
| 2 | 3.95 | 2.38 | -5.38 | -4.18 | - | - | 0.44 | 0.47 |
| 3 | -0.12 | -0.54 | -0.81 | -0.26 | - | - | 0.71 | 0.96 |
| 4 | 3.61 | 2.67 | -5.33 | -4.26 | - | - | 0.75 | 0.74 |
| 5 | 0.56 | 0.60 | -2.58 | -2.55 | 0.28 | 0.29 | 0.10 | 0.12 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 95 | -0.13 | -0.02 | -0.44 | -0.50 | 0.40 | 0.14 | 0.31 | 0.32 |
| 96 | -0.67 | -0.88 | -2.04 | -1.31 | 0.26 | 0.24 | 0.93 | 0.98 |
| 97 | -0.52 | -0.86 | -1.80 | -1.28 | - | - | 1.10 | 1.21 |
| 99 | 1.30 | 0.34 | -3.08 | -1.97 | 0.17 | 0.12 | - | - |
| 100 | -0.69 | -0.81 | -0.64 | -0.51 | 0.43 | 0.44 | -0.11 | -0.24 |

Real customer database ($C$ = 500, $I$ = 100).

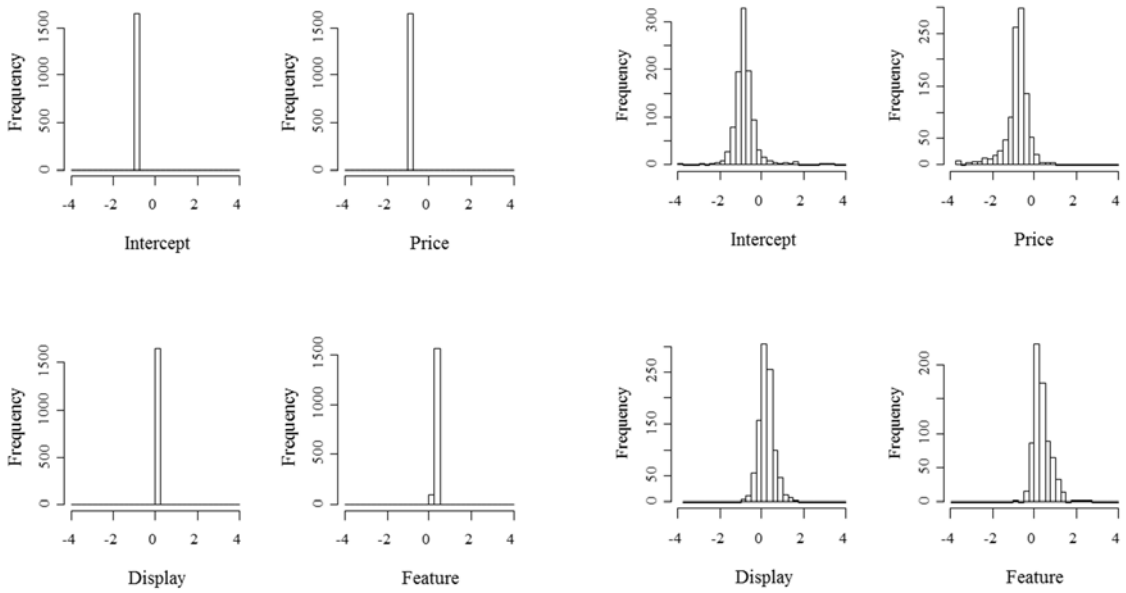


Figure 1 Marginal distribution of parameter estimates of individual consumers and products