DSSR

Discussion Paper No. 35

# Topic Modeling of Market Responses for Large-Scale Transaction Data

Tsukasa Ishigaki
Nobuhiko Terui
Tadahiko Sato
Greg M. Allenby

February 23, 2015

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

# Topic Modeling of Market Responses for Large-Scale Transaction Data

by

Tsukasa Ishigaki
Tohoku University

Nobuhiko Terui
Tohoku University

Tadahiko Sato
University of Tsukuba

Greg M. Allenby
Ohio State University

February 23, 2015

# Topic Modeling of Market Responses for Large-Scale Transaction Data

**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 cube, 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 simulation study. We find the model to be highly scalable and useful for identifying effective marketing variables for consumers, including infrequent purchasers.

**Keywords:** Cross-category Analysis, Data Cube, Hierarchical Bayes Moldel, Market Response, Panel Data, Personalization, Topic Model, Variational Bayes Inference

† This paper is a revised version of a discussion paper by T. Ishigaki, N. Terui, T. Sato and G. Allenby, A Large-Scale Marketing Model using Variational Bayes Inference for Sparse Transaction Data, Data Science and Service Research Discussion Paper, No. 18, 2014

# 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, 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 changing over occasions requires the use of model-based inference as opposed to methods applied directly to the marginal data (Anderson 2003).

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. Thus, instead of attempting to build an economic model of choice across dozens of product categories, explicitly modeling the presence of substitutes, complements and an inter-related set of budget constraints, 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). These models make the simplifying assumption that votes, words or purchases are outcomes of latent probabilities that describe the occurrence of events. We extend these models so that the latent probabilities are a function of a brand's own marketing variables.

The topic model is a generalization of a finite mixture model in which each data point is associated with a draw from a mixing distribution (The and Jordon 2010). Models of voting blocs (Spirling and Quinn 2010) 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 vector of observations and not to the entire set of observations (e.g., series of votes a legislator or set of words by an author) of the panelist. In our analysis of household purchases, we allow the vector of observed purchases across all product categories on an occasion to be related to a different latent context (topic, or issue). 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 purchase 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 (i.e., at least an order of magnitude) of offerings across multiple product categories.

Our model identifies the latent state, or topic assignment, for each consumer at each point in time, providing information about the array of products a consumer will likely purchase. We do not make a-priori assumptions about substitute and complementary goods in the spirit of market basket analysis in data mining. Our model takes an exploratory approach to analysis and does not test assumptions of the form of the utility function across hundreds of offerings. However, our model does include marketing variables so that their effects on choice can be measured and used in prediction.

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 model and simulation studies that verify the precision of VB estimate and scalability of the model. In section 4, we first discuss the joint segmentation of consumers and items for cross-category analysis, and propose a method of decompressing the information obtained in the reduced-dimensional space to make marketing decisions in the original large-scale original space. Section 5 applies the model to customer purchases in a general merchandise store. Discussion and concluding remarks are offered in Section 6.

# 2 Model Development

The analysis of large-scale transactional data is challenging because of the sparsity of observed purchases. Most consumers do not purchase in most product categories on most shopping trips, and when a purchase is recorded in one category it is frequently

for just one offering. The actual sample size of transactional data is much smaller than the data space reflected by a data cube with dimensions corresponding to the number of consumers, number of products and time. In this situation, standard random-effect model specifications break down because of the high frequency of non-purchase for almost every brand.

In choice models, maximum likelihood estimates of brand intercepts are driven to negative infinity if the brand is never purchased by the household, and since most households do not ever purchase most brands, standard random-effect models result in excessive negative shrinkage of the intercepts. Similarly, the prevalence of non-purchase makes it appear that consumers are not price sensitive because they do not react to competitive price discounts, when in fact they may be making a quick trip to the store and may not even be exposed to prices in many categories. The analysis of large-scale transactional data must therefore employ additional assumptions about heterogeneity and price responsiveness not typically made in the analysis of revealed preference data. We relate consumer purchases to latent segments as is done in models for text analysis that greatly reduces the dimensionality of the model. Response parameters are then introduced in the reduced dimensional space by connecting each choice to their own marketing variables with a hierarchical probit model. We do not attempt to model cross-price and cross-merchandising effects because of the large number of brands under study.

## 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 that is common in topic models in the context of consumer purchases. We seek the probability $p(i|c)$ that consumer $c$ purchases item $i$. We assume the dataset includes $C$ consumers and $I$ product items through $T$ periods. However, the probabilities cannot be accurately calculated because of data sparseness. The topic model calculates $p(i|c)$ by introducing a latent class $z \in \{1 \dots Z\}$

whose dimension is significantly smaller than the number of consumers and items.

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

$$\begin{bmatrix} p\left(i=1|c=1\right) & \cdots & p\left(i=1|c=C\right) \\ \vdots & \ddots & \vdots \\ p\left(i=I|c=1\right) & \cdots & p\left(i=I|c=C\right) \end{bmatrix} = \sum_{z=1}^{Z} \begin{bmatrix} p\left(1|z\right) \\ \vdots \\ p\left(I|z\right) \end{bmatrix} \begin{bmatrix} p\left(z|1\right) & \cdots & p\left(z|C\right) \end{bmatrix} \tag{1}$$

More specifically, we decompose a large probability matrix of size $I \times C$ 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 $p(i|c)$ in the $I \times Z$ matrix. The voting blocs model supposes a Bernoulli distribution for the probability $p(i|c)$. LDA assumes a categorical (i.e., multinomial) distribution for the probability matrix.

The latent classes $z$ serve to define types of purchase baskets across the $I$ products. The first term on the right side of (1) defines a vector of choice probabilities for each item under study, assuming that the purchase occasion is of type $z$. Items with high probability are likely to be jointly present in the basket, so our model identifies likely bundles of goods purchased for different types of shopping trips. The second term is the probability that a consumer's purchases are of type $z$. Our model does not model heterogeneity in a traditional manner, where there is a common set of response parameters for all purchases of an individual. We instead assume that each purchase belongs to one of $Z$ types, and that respondents can also be characterized in terms of the probability their purchases are of these types.

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 pur-

chase 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 Market Response 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 $u_{cit} \leq 0$ 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:

$$p\left(u_{cit} > 0\right) = \sum_{z=1}^{Z} p\left(u_{it} > 0 | z\right) p\left(z|c\right) \tag{2}$$

We denote the utility associated with the latent class $z$ as $u_{it}^{(z)}$, and then the choice probability can be represented as $p(u_{it} > 0|z) = p(u_{it}^{(z)} > 0)$. Assuming a linear Gaussian structure on the utility $u_{it}^{(z)}$ for marketing variables, the right hand side of (1) can be represented as:

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

where $\boldsymbol{\beta}_{zi}^{T} = (\beta_{zi1}, \ldots, \beta_{ziM})$ is a response coefficient vector of latent class $z$ with respect to item $i$, $\boldsymbol{x}_{it}^{T} = (x_{it1}, \ldots, x_{itM})$ is a vector of $M$ marketing variable for item $i$ at time $t$, and

$F(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. In our empirical study, $x_{it}$ includes price and promotional variables.

We next set a categorical distribution $\theta_{cz}$ for the probability $p(z|c)$ that consumer $c$ belongs to the latent class $z$. The categorical distribution is multinomial with parameters $\boldsymbol{\theta}_c$. The $\boldsymbol{\theta}_c$ is specified so that the selection probability of consumer $c$ with respect to item $i$ is conditionally independent if the latent class $z$ is given. That is, all of the information about respondent heterogeneity of purchases is conveyed through the latent classes. Then, the right hand side of (1) is represented by:

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

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

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

where the prior distributions for $\boldsymbol{\mu}_i$ and $V_i$ follow an M-dimensional multivariable normal distribution $N_M(\tilde{\boldsymbol{\mu}}, \tilde{\sigma}^2 V_i)$ and an inverse-Wishart distribution $IW(\tilde{W},\tilde{w})$, where $\tilde{\boldsymbol{\mu}}$, $\tilde{\sigma}^2$, $\tilde{W}$ and $\tilde{w}$ are parameters specified by the analyst. We assume that the $M$-dimensional coefficient vector $\boldsymbol{\beta}_{zi}$ for each segment, $z$, is a draw from a distribution with mean and covariance that is item-specific.

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

$$\boldsymbol{\theta}_c \sim Dirichlet\left(\tilde{\boldsymbol{\gamma}}\right) \tag{6}$$

The likelihood is given as:

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

where $p(y_{cit}|x_{it}, \beta_{zi}, z)$ 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 $c$ at least once during the period $t = 1, \ldots, T$, that is, $T_c \in \left\{ t | \sum_{i=1}^{I} y_{cit} > 0 \right\}$ and $I_c \in \left\{ i | \sum_{t=1}^{T} y_{cit} > 0 \right\}$.

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 $z_{cit} \in \{1, \ldots, z \ldots, Z\}$ 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:

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

where $p(z_{cit} = z|\boldsymbol{\theta}_c)$ denotes a categorical distribution when $\boldsymbol{\theta}_c$ is given. Hereinafter, $(z_{cit} = z)$ is denoted as $z_{cit}$ to simplify notation.

Our model for large-scale transaction data is different from related standard models in two respects. First, the likelihood is defined over brands and time periods in which purchases are observed to take place at least once as indicated by the variables $T_c$ and $I_c$. It is composed of not only purchase but also non-purchase occations for identifying market response parameter. In this sense, our model differs from topic models used in text analysis where the likelihood is formed using the words present in a corpus, not the words that are not present. Second, heterogeneity is introduced at the observation-level, allowing the different transactions of a household to reflect different latent states, $z$ at

every $(c, i, t)$, which is denoted by $z_{cit}$. It provides us with useful information for characterizing respondents and brands, and predicting their purchases. This differs from the traditional latent class model where the likelihood of all household purchases contributes to inferences about a respondent's latent class membership ($z$) and parameters ($\beta$).

The posterior distribution of parameters including latent variables of states $\{z_{cit}\}$ and augmented utilities $\{u_{cit}^{(z)}\}$ of probit model is then given by:

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

# 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 $\boldsymbol{D}$ is given, the log marginal likelihood $\log p(\boldsymbol{D})$ of the target distribution is decomposed into two components as:

$$\log p\left(\boldsymbol{D}\right) = L\left(q\right) + KL\left(q \,\|p\right) \tag{10}$$

$$L\left(q\right) = \int q\left(\boldsymbol{\Theta}\right) \log\left\{p\left(\boldsymbol{D}, \boldsymbol{\Theta}\right) q(\boldsymbol{\Theta})^{-1}\right\} dZ \tag{11}$$

$$KL\left(q \,\|p\right) = -\int q\left(\boldsymbol{\Theta}\right) \log\left\{p\left(\boldsymbol{\Theta}|\boldsymbol{D}\right) q(\boldsymbol{\Theta})^{-1}\right\} dZ \tag{12}$$

$L(q)$ is called variational lower bound in VB inference, and $KL\left(q \,\|p\right)$ is the Kullback-Leibler divergence of the target and variational distributions. As is well known, $KL\left(q \,\|p\right)$ 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 $KL\left(q \,\|p\right)$ is minimized. However, it is difficult to evaluate the value of $KL\left(q \,\|p\right)$ because the expression involves a posterior distribution of $p(\boldsymbol{\Theta}|\boldsymbol{D})$.

In contrast, $L(q)$ involves a joint distribution $p(\boldsymbol{D}, \boldsymbol{\Theta})$ 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 $KL\left(q \,\|p\right)$ 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, the parameters are called variational parameters $q\left(\boldsymbol{\Theta}\right) = \prod_{j=1}^{J} q_j\left(\boldsymbol{\Theta}^{(j)^*}\right)$ and can be maximized by the following updating algorithm (Jordan et al., 1999):

$$\begin{aligned}\boldsymbol{\Theta}^{(j)*\{new\}} &\leftarrow \underset{\boldsymbol{\Theta}^{(j)*}}{\arg\max}\, L\left(\prod\nolimits_j^J q_j\left(\boldsymbol{\Theta}^{(j)*}\right)\right) \\ &\propto \exp\left(\boldsymbol{E}_{k\neq j}\left[\log p\left(\boldsymbol{D}, \boldsymbol{\Theta}\right)\right]\right). \end{aligned} \tag{13}$$

The $E_{k\neq j}[\ \ ]$ are the expectation value associated with $q_j$ distributions over all parameters $\boldsymbol{\Theta}^{(j)^*}$, where $k \neq j$. 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 $q_j(\boldsymbol{\Theta}^{(j)^*})$ (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 proposed model. The parameters and variational parameters are denoted as
$\boldsymbol{\Theta} = \left\{ \{\boldsymbol{\theta}_c\}, \{z_{cit}\}, \left\{u_{cit}^{(z)}\right\}, \{\boldsymbol{\beta}_{zi}\}, \{\boldsymbol{\mu}_i\}, \{V_i\} \right\}$ and
$\boldsymbol{\Theta}^* = \left\{ \{\boldsymbol{\theta}_c^*\}, \{\boldsymbol{z}_{cit}^*\}, \left\{u_{cit}^{(z)*}\right\}, \{\boldsymbol{\beta}_{zi}^*\}, \left\{V_{iz}^{\beta*}\right\}, \{\boldsymbol{\mu}_i^*\}, \{\sigma_i^{\mu*}\}, \{w_i^*\}, \{W_i^*\} \right\}$ respectively, while the variational distributions are configured as

$$
\begin{aligned}
& q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*, \{\boldsymbol{x}_{it}\}, \{y_{cit}\}\right) \\
&= \left[\prod_{c=1}^{C} q_c\left(\boldsymbol{\theta}_c \mid \boldsymbol{\theta}_c^*\right)\right] \left[\prod_{c=1}^{C} \prod_{i \in I_c} \prod_{t \in T_c} q_z\left(z_{cit} \mid \boldsymbol{z}_{cit}^*\right)\right] \left[\prod_{c=1}^{C} \prod_{i \in I_c} \prod_{t \in T_c} q_u\left(u_{cit}^{(z)} \mid u_{cit}^{(z)*}, \boldsymbol{z}_{cit}^*, \boldsymbol{\beta}_{zi}^* . \boldsymbol{x}_{it}\right)\right] \\
&\quad \left[\prod_{i=1}^{I} \prod_{z=1}^{Z} q_\beta\left(\boldsymbol{\beta}_{zi} \mid \boldsymbol{\beta}_{zi}^*, V_{zi}^{\beta*}\right)\right] \left[\prod_{i=1}^{I} q_{\mu,V}\left(\boldsymbol{\mu}_i, V_i \mid \boldsymbol{\mu}_i^*, \sigma_i^{\mu*}, w_i^*, W_i^*\right)\right] \qquad (14)
\end{aligned}
$$

where $q_c$ is a Dirichlet distribution with variational parameter $\boldsymbol{\theta}_c^*$, $q_z$ is a categorical distribution with variational parameter $\boldsymbol{z}_{cit}^*$, $q_u$ is a truncated normal distribution with parameter $z_{cit}$ and variational parameter $u_{cit}^{(z)*}$, $q_\beta$ is an $M$-dimensional multivariable normal distribution with two variational parameters (mean vector $\boldsymbol{\beta}_{zi}^*$ and covariance matrix $V_{zi}^{\beta*}$), and $q_{\mu,V}$ is a multivariable normal–inverse Wishart distribution with variational parameters $\boldsymbol{\mu}_i^*$, $\sigma_i^{\mu*}$, $w_i^*$, $W_i^*$. 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.

## 3.2 Simulation Study

In this subsection we examine the performance of the proposed VB estimator. In addition to computational time, VB has another advantage over MCMC in that it is not prone to the label switching problem encountered in MCMC estimation (see Puolamaki and Kaski, 2009).

We examine the precision of the estimates and computational time separately. The first simulation evaluates the recovery of true parameter values by VB, and the second simulation examines scalability. We compare the computational times of VB to MCMC, ignoring label switching problem encountered with MCMC estimation. 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.

### 3.2.1 Simulation Dataset

In this simulation study, purchase records are generated by simulation using marketing variables. The marketing variables are extracted from a real customer database of a general merchandise store. The marketing variables vector is composed of price ($\bar{P}_{it}$), display ($D_{it}$), and feature ($F_{it}$); that is, $\boldsymbol{x}_{it}^T = [1\ \bar{P}_{it}\ D_{it}\ F_{it}]$ . $\bar{P}_{it}$ is the discount rate 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. Here, the value of $\bar{P}_{it}$ is normalized into interval $[0, 1]$ in order to conform the scale of discount rate to the scale of dummy variables.

We assume that any customers belong to one of three segments characterized by response coefficients for marketing variables. First segment (Segment 1) has a response coefficient $\bar{\boldsymbol{\beta}}_1 = [-0.5, 1, 0, 0]^T$, that is, customers in the segment sensitively respond to discount of product items and are not affected from display or feature. Similarly, we employ $\bar{\boldsymbol{\beta}}_2 = [-0.5, 0, 1, 0]^T$ and $\bar{\boldsymbol{\beta}}_3 = [-0.5, 0, 0, 1]^T$ as response coefficient vectors for second (Segment 2) and third segments (Segment 3) that are influenced from display and

feature promotion only, respectively. The three vectors are set as true values of response parameter. This setting means that any product items have the same properties on the response to marketing promotions for a simplification of analysis. The verification or check of parameter estimation will be too complicated if we employ different coefficient vector for each product items.

Next, we make coefficient vectors of individual customers. Here, we suppose that each segment consists of 100 customers and 50 product items are in a store. The individual coefficients vectors $\bar{\boldsymbol{\alpha}}_{ci}$ are generated by followings; $\bar{\boldsymbol{\alpha}}_{ci} \sim N(\bar{\boldsymbol{\beta}}_1, \sigma \boldsymbol{I})$ $(c = 1, \cdots, 100)$, $\bar{\boldsymbol{\alpha}}_{ci} \sim N(\bar{\boldsymbol{\beta}}_2, \sigma \boldsymbol{I})$ $(c = 101, \cdots, 200)$ and $\bar{\boldsymbol{\alpha}}_{ci} \sim N(\bar{\boldsymbol{\beta}}_3, \sigma \boldsymbol{I})$ $(c = 201, \cdots, 300)$, and $\sigma$ is set as 0.1. Then, the utilities for 30 days are simulated by $\bar{u}_{cit} = \boldsymbol{x}_{it}^T \bar{\boldsymbol{\alpha}}_{ci} + \bar{\epsilon}_{cit}$ $(\bar{\epsilon}_{cit} \sim N(0, 1))$ and the purchased records $\{\bar{y}_{cit}\}$ are generated as $\bar{y}_{cit} = 1$ if $\bar{u}_{cit} > 0$ and $\bar{y}_{cit} = 0$ otherwise.

### 3.2.2 Precision of Estimates

In this subsection, we examine how well the parameters of $\bar{\boldsymbol{\beta}}_1$, $\bar{\boldsymbol{\beta}}_2$ and $\bar{\boldsymbol{\beta}}_3$ are recovered in the proposed model with VB. Here, we generate ten simulation datasets by the procedures above, and we set hyper-parameters as $\tilde{\boldsymbol{\gamma}} = [0.1, \cdots, 0.1]^T$, $\tilde{\boldsymbol{\mu}} = [0, \cdots, 0]^T$, $\tilde{\sigma}^2 = 1$, $\tilde{W} = \boldsymbol{I}_M$ and $\tilde{w} = 10$ and appropriate initial values. $\boldsymbol{I}_M$ is the identity matrix of size $M$. In VB estimation, the iterations are terminated when the variational lower bound improves by less than $10^{-5}\%$ of the current value in two consecutive iterations (the variational lower bound is described in Appendix B). These settings for the hyper-parameters and the stopping rule of the VB iterations are adopted in all empirical studies hereafter.

Table 1 displays the means and standard deviations of estimates using the ten simulation dataset. The numbers in Table 1 are calculated as $50^{-1} \sum_{i=1}^{I} \hat{\boldsymbol{\beta}}_{zi}$ ($\hat{\boldsymbol{\beta}}_{zi}$ represents a estimated posterior mean of $\boldsymbol{\beta}_{zi}$.). The results indicate that the VB estimates are close to true values for all parameters in every segment.

Table 1: Estimates of Simulation Data

### 3.2.3 Scalability

Scalability is investigated for: $C = \{1000, 5000, 10000\}$, $I = \{100, 500, 1000\}$, $T = 30$ and $Z = \{5, 10, 20\}$. Thus, 27 different scenarios were explored in the scalability study. The MCMC estimator is described in Appendix C, and we forecast the simulation times for 6,000 MCMC samples from ten samples for computational feasibility. We note that the selection of 6,000 MCMC samples is consistent with the simulation study of Braun and Mcauliffe (2010). The simulated data is the same as used in above, and the results reported below were calculated in identical computational environment (64-bit version of Python 2.7.5 with Numpy, implemented on a 3.5 GHz processor (Quad-Core Xeon; Intel Corp.) with 64 GB memory).

Table 2 reports computation time in hours for the VB and MCMC estimators. For both algorithms, the computational cost increases linearly with the size of the dataset specified in terms of the number of consumers, items, and latent classes. In all scenarios, the times of MCMC computations exceed those of VB. The VB algorithm is approximately 20 to 50 times more efficient than MCMC, depending on the scenario. The time of computation using large-scale data (C = 10000, I = 1000) by MCMC is estimated to be over 450 hours, and thus we recognize that MCMC is not applicable for our problem. The results of the simulation show that VB estimates are reliable in precision and computationally feasible for analysis. In contrast, MCMC becomes increasingly prohibitive as the number of customers and choice alternatives increases.

Table 2: Simulation Time by VB and MCMC

# 4 Joint Segmentation and Personalization

The variational estimates $\hat{\boldsymbol{\beta}}^*_{zi}$, $\hat{u}^{(z)*}_{cit}$, $\hat{\theta}_c$, and $\hat{z}^{(z)*}_{cit}$ can be transformed into statistics that are relevant for segmentation and targeting using the $M$ dimensional vector of probability

$\hat{z}_{cit}^{(z)*}, z =, 1, ..., Z$ at each point of data cube:

$$q_z \left( z_{cit} \mid \hat{\boldsymbol{z}}_{cit}^{*} \right) \tag{15}$$

Given the variational Bayes estimates $\hat{\Lambda} = \{\hat{\boldsymbol{\beta}}_{zi}^{*}, \hat{u}_{cit}^{(z)*}, \hat{\theta}_c\}$, we obtain the probability of customer segment membership by aggregating over products ($i$) and time ($t$):

$$p\left(c \in z|\hat{\Lambda}\right) = \frac{\sum_{i \in I} \sum_{t \in T_c} \hat{z}_{cit}^{(z)*} \times I(y_{cit} = 1)}{\sum_{z_k=1}^{Z} \sum_{i \in I} \sum_{t \in T_c} \hat{z}_{cit}^{(z)*} \times I(y_{cit} = 1)} \tag{16}$$

and aggregating over customers ($c$) and time ($t$) yields the probability of product segment membership:

$$p\left(i \in z|\hat{\Lambda}\right) = \frac{\sum_{c=1}^{C} \sum_{t \in T_c} \hat{z}_{cit}^{(z)*} \times I(y_{cit} = 1)}{\sum_{z_k=1}^{Z} \sum_{c}^{C} \sum_{t \in T_c} \hat{z}_{cit}^{(z)*} \times I(y_{cit} = 1)} \tag{17}$$

where $I(\cdot)$ is the indicator function equal to one if the augment holds and zero otherwise. We take the sums over the instances of purchase because we believe that non-purchase can occur for many reasons other than non-membership (e.g., having large household inventory of the product). Our estimates of customer and product latent membership are driven by customer actions and not their inactions.

We can also construct market response estimates for each respondent and each product from $\hat{\Lambda} = \{\hat{\boldsymbol{\beta}}_{zi}^{*}, \hat{u}_{cit}^{(z)*}, \hat{\theta}_c\}$ by projecting the estimates of latent utility on marketing variables. That is, the estimates are obtained from an auxiliary regression of latent utility $\hat{U}_{ci}^{(k)*}$ stacked by $\hat{u}_{cit}^{(k)*}$ with the state $k$ = argmax $\hat{z}_{cit}^{(z)*}$ changing over time on the corresponding marketing variables $X_{ci}$ constituted by $\boldsymbol{x}_{it}$ ($t \in T_c$).

$$\hat{\boldsymbol{\alpha}}_{ci} = \left( {X_{ci}}^T X_{ci} \right)^{-1} {X_{ci}}^T \hat{U}_{ci}^{(k)*}. \tag{18}$$

The estimates above provide a bridge between the granularity of the model, where heterogeneity is introduced at each point in the data cube, and managerial inferences and

decisions that are made across products (e.g., which customers to reward), across customers (e.g., which products to promote) and over time. In addition, the standard $t$ test in the standard linear regression models can be used for testing significance of estimates.

# 5 Empirical Analysis

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 variables were recorded for each purchase occasion. The dataset contains 94,297 transactions involving 1,650 consumers and 500 items. The items were chosen by being displayed and featured at least once in the data period. The marketing variables are price ($P_{it}$), display ($D_{it}$), and feature ($F_{it}$); that is, $x_{it}^T = [1\ P_{it}\ D_{it}\ F_{it}]$ . $P_{it}$ is the price relative to the maximum price of item $i$ in the observational period. The display and feature are binary entries, equal to one if the item $i$ is displayed or featured at time $t$, and zero otherwise.

## 5.1 Cross-category analysis

Our model of purchase behavior allows for observation-level heterogeneity that acknowledges that each purchase occasion can be viewed as the building-block for analysis. Some occasions are associated with large trips to the store while other occasions may have be more focused on a specific set of offerings. Moreover, consumers may exhibit behavior consistent with multiple occasions, or topics, over time. While it may be desirable for firms to classify goods and respondents to segments for the purpose of understanding different types of customers and goods, our model is capable of conducting analysis at a more disaggregate level. Alternatively, our model can be used to associated both offerings and customers to latent topics, or segments, for understanding and managing market basket purchases.

We illustrate such cross-category analysis using a $z = 10$ topic solution. Conditioning

on the number of segments is common practice in the machine learning literature. We tried, but were not successful in estimating $z$ as part of our model (see appendix B) and leave this as an area for future research..

Table 3: Joint Segmentation for Cross-category Analysis

Table 3 displays the result of the joint segmentation of products and consumers using equations (16) and (17). The five products with highest probability and their average levels of marketing activity are shown for each segment. The first column reports the brand name, the second column reports the product category associated with the offering, and the remaining columns display the average level of marketing activity, i.e., the average price rate, average display rate, and average feature rate. The title of each segment contains the numbers of items and customers jointly classified into the same segment. The segments are interpreted as follows.

The first segment has 31 consumers and 9 items are assigned to it. This segment contains beverages across different categories with small discount rates and low rates of feature advertising. The second segment is characterized as being composed of the identical brands in the desert category. The items are infrequently discounted and have a higher rate of display than the first segment. Segments 3 through 7 have relatively fewer consumers and items, and they exhibit greater variation in the level of marketing activity. In particular, Segment 5 contains two offerings in both the ice cream and dressing categories with the same brand names, both with relatively high rates of display and feature activity. Segment 6 contains contains mainly items from the drink category and is similar in marketing activity with segment 5. Segment 7 is also comprised of drink items with higher marketing level as well as other items with lower level of activities. The items in segment 8 are comprised of variety of product categories with relatively higher level of display. Segment 9 is the largest cluster with 946 consumers and 332 items. It is

characterized as having the highest level of display activity. Segments 8 and 10 contain the less discounting and more displayed items, and the former is double and triple sized in consumers and items.

The potential use of this information is in managing cross-category behavior. Knowing the products typically purchased for different types of shopping trips can be used to determine the range of impact of price promotions and merchandising activity. If consumers have a budget for a particular shopping occasion, rather than for a particular product category, then the influence of a price reduction will have a broader effect in traditional models of demand. Our model allows for the identification of the boundary of effects as part of the topic, or latent segment, characterization.

## 5.2 Individual-level parameter estimates

The management of pricing, displays and feature activity within a store involves decisions that cut across time and consumers, and requires knowledge of which product categories are most sensitive to these actions. More recently, targeted coupon delivery systems have allowed for the individual-level customization of prices. Managing these decisions requires a view of the sensitivity of consumers and product categories to these actions.

Individual-level estimates of market response is obtained by using the equation (18) and two sided significance test on each estimate with the level of 5% is conducted by $t$ test for deciding effectiveness of marketing variables in empirical analysis. We note that customers will display variation in their sensitivity to variables such a price across product categories because of varying aspects of the product categories (e.g., necessary versus luxury goods, amount of product differentiation, price expectations) and different purposes of the shopping visit over time (e.g., shopping for one's self or others, large versus small shopping trip, etc.).

We can marginalize $\hat{\boldsymbol{\alpha}}_{ci}$ by either of its arguments, $c$ and $i$, to obtain characterizations of customers and items useful for analysis. The empirical marginal distribution of

consumer parameter estimates is obtained by averaging across the 500 products in our analysis, i.e., $\left\{ \sum_{c=1}^{C} \hat{\boldsymbol{\alpha}}_{ci}/C \right\}$. A histogram of 500 items for each marketing variable are displayed on the left side of Figure 1, and provides information about the general distribution of heterogeneity faced by the firm for actions such as price customization. We find that the individual-level estimates to be plausible in that the price coefficients are negative and the display and feature coefficients are estimated to be positive.

We can also summarize heterogeneity across consumers and examine the distribution of marketing variables for the 500 products in our analysis. The empirical marginal distributions of individual products, averaging over the 1650 consumers, i.e., of $\left\{ \sum_{i=1}^{I} \hat{\boldsymbol{\alpha}}_{ci}/I \right\}$, are depicted on the right of Figure 1. The products that never displayed and featured in the data period have been omitted from the histograms. These estimates are useful for deciding which product categories should receive merchandising support in the form of in-store displays and feature advertising. We find that the estimates are plausible in most product categories with negative price coefficients, and positive display and feature coefficients, but there exists fairly wide variation in the effectiveness of these variables across products. Many product categories appear to be unresponsive to merchandising efforts.

Figure 1: Marginal Distribution of Individual Parameter Estimates:

Figure 2 provides a two dimensional summary of the data and coefficient estimates. Figure 2(a) is a scatter plots of two dimensional data cube with respect to customers ($i$) and products ($c$), aggregated along the time ($t$) dimension. If a customer has never purchased a specific product in the dataset, the coordinate $(i, c)$ is colored "white," and it is "black" if they have purchased the product at least once. We observe that customer-item space is still very sparse.

Figures 2(b)-2(d) show the results of testing with a 5% level of significance level for non-zero individual response coefficients. In figure 2(b), the coordinates with a significant price

coefficient is indicated as "black" and "white" shows that the estimate is insignificant. The effectiveness of display and feature promotions are similarly defined. We find that our model produces many significant price, display and feature coefficients.

Figure 2: Personalized Effective Marketing Variables for All Customers and Items

Figure 3 provides a close up of Figure 2 for 100 products and customers. An interesting aspect of our analysis is that significant coefficients can arise even when a customer has never purchase a product because of the imputation present in the topic model for non-purchases. The topic model greatly reduces the dimensionality of the data cube, as shown in Equation (1), and results in individual-level estimates in a sparse data environment. Our analysis yields coefficient estimates at the individual- and product-level by way of the latent topics that transcend the product categories. Our model enables marketers to develop effective pricing and promotional strategies by recognizing the presence of latent topics, or shopping baskets, present at each point in time in the data cube.

Figure 3: Personalized Effective Marketing Variables for 100 Customers and Items

# 6 Discussion

The unit of analysis in marketing is not a person or a product, but a person embedded within a context of action for which a product might be useful. Consumers find value in the goods and services that help them deal with issues in their lives, which is time specific. It is not surprising that shopping behavior is therefore time specific, with some shopping trips encompassing a large number of purchases and expenditures, while other shopping trips having a much smaller number of items being purchased. We propose a model for dealing with a large number of offerings by recognizing the presence of shopping heterogeneity at each point in time, and employ a topic model for dealing with the many choice alternatives available at retailers.

This paper addresses three challenges in estimating models of demand in large databases: i) the large number of available products, ii) the large number of consumers who purchase these products, and iii) the sparseness of transaction data. 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 produced similar results. We therefore believe that estimation times can be further reduced in practice from those reported in this paper.

Our model allows us to engage in the joint segmentation of consumers and items by using the posterior probability of latent state which is allocated to every point of data

cube. The information on response to marketing efforts in a reduced dimensional space compressed by topic model is decompressed into original space by using variational Bayes inference to obtain the individual response parameters in data cube. We show how the model can be used to produce information useful for personalized marketing for both specific customers and specific products, and effectively deals with data spareness due to infrequent consumer purchases.

Our model assumes the stability of the topic structure over time. However, it is possible that consumer's market response and purchase patterns change over time because of factors such as new trends, state dependence and the arrival of new purchase and delivery technologies. We believe the development of a dynamic topic model for purchase is an interesting extension of our work, and leave this for 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

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

where $\boldsymbol{D} = \{\{\boldsymbol{x}_{it}\}, \{y_{cit}\}\}$.

The update procedures of variational parameters $\boldsymbol{\theta}_c^*$, $\boldsymbol{z}_{cit}^*$, $u_{cit}^{(z)*}$, $\boldsymbol{\beta}_{iz}^*$, $V_{iz}^{\beta*}$, $\boldsymbol{\mu}_i^*$, $\sigma_i^{\mu*}$, $w_i^*$, and $W_i^*$ are presented below.

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

The Dirichlet and categorical distributions are of the following forms:

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

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

is then calculated for each $c$ as

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

where, $z^*_{citz}$ is a element of $\boldsymbol{z}^*_{cit}$. 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 $\tilde{\gamma}_z + \sum_{i\in I_c}\sum_{t\in T_c} z^*_{citz}$. Therefore,

$$\boldsymbol{\theta}^*_c \leftarrow \tilde{\boldsymbol{\gamma}} + \sum_{i\in I_c}\sum_{t\in T_c}\boldsymbol{z}^*_{cit} \tag{A4}$$

## A.2 Optimization of $\boldsymbol{z}^*_{cit}$

Here we denote a digamma function as $\Psi\left(\cdot\right)$, which will be useful for later discussion, and summarize the property of truncated normal distribution in the probit model. $u^{(z)}_{cit}$ follows a normal distribution with mean $\boldsymbol{x}^T_{it}\boldsymbol{\beta}_{zi}$ and variance 1. Moreover, $u^{(z)}_{cit}$ must satisfy $y_{cit} = 1$ if $u_{cit} > 0$ and $y_{cit} = 0$ if $u_{cit} \leq 0$. Therefore, $u^{(z)}_{cit}$ is generated from a truncated normal distribution as

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

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

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

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

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

where $\varphi_{cit}^{(z)} \equiv (-1)^{(1-y_{cit})} f\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\right) / \Omega_{cit}^{(z)*}$ and $\Omega_{cit}^{(z)*} \equiv \left\{F\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\right)\right\}^{y_{cit}} \left\{1 - F\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\right)\right\}^{(1-y_{cit})}$. Thus, the expected value $\boldsymbol{E}_{\neq q_z}\left[\log p\left(\boldsymbol{D}, \boldsymbol{\Theta}\right)\right]$ is given as

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

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

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

To solve Equation (A8) for $z_{citz}^*$, 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 zeroth-order Taylor expansion in terms of the CDF of normal distribution and the logarithm function. Such bold approximation is 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

$$\begin{aligned}
&\boldsymbol{E}_{q_\beta}\left[\log \Omega_{cit}^{(z)}\right] \approx \text{const},\\
&\boldsymbol{E}_{q_u}\left[\left(u_{cit}^{(z)}\right)^2\right] = \boldsymbol{V}\left[u_{cit}^{(z)}\right] + \left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^* + \varphi_{cit}^{(z)}\right)^2,\\
&\boldsymbol{E}_{q_u,q_\beta}\left[u_{cit}^{(z)}\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}\right] = \left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^* + \varphi_{cit}^{(z)}\right)\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\right) + \boldsymbol{x}_{it}^T V_{zi}^{\beta *}\boldsymbol{x}_{it},\\
&E_{q_\beta}\left[\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}\right)^2\right] = \boldsymbol{x}_{it}^T V_{zi}^{\beta *}\boldsymbol{x}_{it} + \left(\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\right)^2.
\end{aligned} \tag{A10}$$

Finally, $z_{citz}^*$ is updated as

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

where

$$\rho_{citz} = \Psi\left(\theta_{cz}^*\right) - \Psi\left(\sum\nolimits_{z-1}^{Z}\theta_{cz}^*\right) + \frac{1}{2}\boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^*\varphi_{cit}^{(z)} + \frac{1}{2}\boldsymbol{x}_{it}^T V_{zi}^{\beta *}\boldsymbol{x}_{it}. \tag{A12}$$

## A.3 Optimization of $u_{cit}^{(z)*}$

Similar to Equations (A3) and (A9), the expected value that optimizes $u_{cit}^{(z)*}$ is

$$\boldsymbol{E}_{\neq q_u}\left[\log p\left(\boldsymbol{D},\boldsymbol{\Theta}\right)\right] = \boldsymbol{E}_{q_z,q_\beta}\left[\log p\left(u_{cit}^{(z)} \mid \boldsymbol{\beta}_{zi}, z_{cit}, \boldsymbol{x}_{it}, y_{cit}\right)\right] + \text{const.} \tag{A13}$$

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

$$u_{cit}^{(z)*} \leftarrow \boldsymbol{x}_{it}^T\boldsymbol{\beta}_{zi}^* + \varphi_{cit}^{(z)}\ . \tag{A14}$$

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

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).

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

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

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

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 quation (A9). $\boldsymbol{\beta}_{zi}^*$ and $V_{zi}^{\beta *}$ are then arithmetically updated as

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

where

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

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

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

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

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

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

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

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

Next, we optimize $w_i^*$ and $W_i^*$ using Equation (A15) and the relationship $\log q\left(V_i\right) =$

$\log q(\boldsymbol{\mu}_i, V_i) - \log q(\boldsymbol{\mu}_i \mid V_i)$.

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

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

$$\begin{aligned} W_i^* &\leftarrow \tilde{W} + \sum\nolimits_{z=1}^{Z} V_{zi}^{\beta *} + \tilde{\sigma}_\mu^{-1}\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}} + \sum\nolimits_{z=1}^{Z} \boldsymbol{\beta}_{zi}^* \boldsymbol{\beta}_{zi}^{*T} - \left(\tilde{\sigma}_\mu^{-1} + Z\right) \boldsymbol{\mu}_i^* \boldsymbol{\mu}_i^{*T}, \\ w_i^* &\leftarrow \tilde{w} + Z. \end{aligned} \tag{A22}$$

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

# Appendix B: Variational Lower Bound of Proposed Model

The variational lower bound $L(\boldsymbol{\Theta}^*)$ is given by

$$\begin{aligned} L(\boldsymbol{\Theta}^*) &= \int \left[ q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*) \log \frac{p(\boldsymbol{\Theta}, \{\boldsymbol{x}_{it}\}, \{y_{cit}\})}{q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*)} \right] d\boldsymbol{\Theta} = \boldsymbol{E}_{q_{\boldsymbol{\Theta}_\beta}} \left[ \log \frac{p(\boldsymbol{\Theta}, \{\boldsymbol{x}_{it}\}, \{y_{cit}\})}{q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*)} \right] \\ &= L_\theta^{(p)} + L_z^{(p)} + L_u^{(p)} + L_\beta^{(p)} + L_{\mu,V}^{(p)} - L_\theta^{(q)} - L_z^{(q)} - L_u^{(q)} - L_\beta^{(q)} - L_{\mu,V}^{(q)}, \end{aligned}$$

where, each component of $L(\boldsymbol{\Theta}^*)$ is expectation of variables of proposed model. The expectations excepting $L_u^{(p)}$ and $L_u^{(q)}$ are followings;

$$\begin{aligned} L_\theta^{(p)} &= \boldsymbol{E}_{q_c}[\log p(\{\boldsymbol{\theta}_c\})] \\ &= \sum_{c=1}^{C} \left[ \log \Gamma\left( \sum\nolimits_{z=1}^{Z} \tilde{\gamma}_z \right) - \sum_{z=1}^{Z} \log\Gamma(\tilde{\gamma}_z) + \sum_{z=1}^{Z} (\tilde{\gamma}_z - 1) \left\{ \Psi(\theta_{cz}^*) - \Psi\left( \sum\nolimits_{z=1}^{Z} \theta_{cz}^* \right) \right\} \right], \end{aligned}$$

$$
\begin{aligned}
L_z^{(p)} &= \boldsymbol{E}_{q_z,q_c}\left[\log p\left(\{z_{cit}\} \mid \{\boldsymbol{\theta}_c\}\right)\right] \\
&= \sum_{c=1}^{C}\sum_{i\in I_c}\sum_{t\in T_c}\sum_{z=1}^{Z} z_{citz}^{*}\left\{\Psi\left(\theta_{cz}^{*}\right) - \Psi\left(\sum\nolimits_{z=1}^{Z}\theta_{cz}^{*}\right)\right\},
\end{aligned}
$$

$$
\begin{aligned}
L_\beta^{(p)} &= \boldsymbol{E}_{q_\beta,q_\mu,q_{V^\beta}}\left[\log p\left(\{\boldsymbol{\beta}_{zi}\} \mid \{\boldsymbol{\mu}_i, V_i\}\right)\right] \\
&= -\frac{1}{2}\sum_{i=1}^{I}\sum_{z=1}^{Z}\left[\begin{array}{l} M\log 2\pi + \sum_{m=1}^{M}\Psi\left(\frac{w_i^{*}+1-m}{2}\right) + M\log 2 + \log\left|W_i^{*-1}\right| \\ +\left(\boldsymbol{\mu}_{zi}^{*}-\boldsymbol{\mu}_i^{\mu *}\right)^T w_i^{*}\left(W_i^{*}\right)^{-1}\left(\boldsymbol{\mu}_{zi}^{*}-\boldsymbol{\mu}_i^{\mu *}\right) + tr\left\{w_i^{*}\left(W_i^{*}\right)^{-1}V_{zi}^{\beta *}\right\} + \sigma_i^{\mu *}\end{array}\right],
\end{aligned}
$$

$$
\begin{aligned}
L_{\mu,V}^{(p)} &= \boldsymbol{E}_{q_\mu,q_{V^\beta}}\left[\log p\left(\{\boldsymbol{\mu}_i, V_i\}\right)\right] \\
&= -\frac{1}{2}\sum_{i=1}^{I}\left[\begin{array}{l} M\log 2\pi + \tilde{\sigma}_\mu^{-1}\left[\left(\boldsymbol{\mu}_i^{\mu *}-\tilde{\boldsymbol{\mu}}^{\mu}\right)^T w_i^{*}\left(W_i^{*}\right)^{-1}\left(\boldsymbol{\mu}_i^{\mu *}-\tilde{\boldsymbol{\mu}}^{\mu}\right) + \sigma_i^{\mu *}\right] \\ -\tilde{w}\log\left|\tilde{W}\right| + \log 2 + 2\log\Gamma\left(\frac{\tilde{w}}{2}\right) + tr\left\{\tilde{W}\left(W_i^{*}\right)^{-1}\right\} \\ +(\tilde{w}+M+2)\left\{\sum_{m=1}^{M}\Psi\left(\frac{w_i^{*}+1-m}{2}\right) + M\log 2 + \log\left|\left(W_i^{*}\right)^{-1}\right|\right\}\end{array}\right],
\end{aligned}
$$

$$
\begin{aligned}
L_\theta^{(q)} &= \boldsymbol{E}_{q_c}\left[\log q_c\left(\{\boldsymbol{\theta}_c\} \mid \{\boldsymbol{\theta}_c^{*}\}\right)\right] \\
&= \sum_{c=1}^{C}\left[\log\Gamma\left(\sum\nolimits_{z-1}^{Z}\theta_{cz}^{*}\right) - \sum_{z=1}^{Z}\log\Gamma\left(\theta_{cz}^{*}\right) + \sum_{z=1}^{Z}\left(\theta_{cz}^{*}-1\right)\left\{\Psi\left(\theta_{cz}^{*}\right) - \Psi\left(\sum\nolimits_{z-1}^{Z}\theta_{cz}^{*}\right)\right\}\right],
\end{aligned}
$$

$$
\begin{aligned}
L_z^{(q)} &= \boldsymbol{E}_{q_z}\left[\log q_z\left(\{z_{cit}\} \mid \{\boldsymbol{z}_{cit}^{*}\}\right)\right] \\
&= \sum_{c=1}^{C}\sum_{i\in I_c}\sum_{t\in T_c}\sum_{z=1}^{Z} z_{citz}^{*}\log z_{citz}^{*},
\end{aligned}
$$

$$L_{\beta}^{(q)} = \boldsymbol{E}_{q_\beta}\left[\log q_\beta\left(\{\boldsymbol{\beta}_{zi}\} \mid \left\{\boldsymbol{\mu}_{zi}^{*}, V_{zi}^{\beta *}\right\}\right)\right]$$
$$= -\frac{1}{2}\sum_{i=1}^{I}\sum_{z=1}^{Z}\left\{M\log\left(2\pi e\right) + \log|V_{zi}^{*}|\right\}$$

and

$$L_{\mu,V}^{(q)} = \boldsymbol{E}_{q_\mu, q_{V^\beta}}\left[\log q_{\mu,V^\beta}\left(\{\boldsymbol{\mu}_i, V_i\} \mid \{\boldsymbol{\mu}_i^{\mu *}, \sigma_i^{\mu *}, w_i^{*}, W_i^{*}\}\right)\right]$$
$$= -\frac{1}{2}\sum_{i=1}^{I}\left[\begin{array}{l} M\log 2\pi + \log|\sigma_i^{\mu *}| - w_i^{*}\log|W_i^{*}| + w_i^{*}M\log 2 + \frac{1}{2}\log\Gamma\left(\frac{w_i^{*}}{2}\right) \\ +\left(w_i^{*}+M+2\right)\left\{\sum_{m=1}^{M}\Psi\left(\frac{w_i^{*}+1-m}{2}\right) + M\log 2 + \log\left|(W_i^{*})^{-1}\right|\right\} \\ +w_i^{*}+1 \end{array}\right].$$

## B.1 Derivation of $\boldsymbol{L_u^{(p)} - L_u^{(q)}}$

The entropy of $u_{cit}^{(z)}$ is given by

$$Entropy = -\frac{1}{2}\left\{\boldsymbol{E}\left[\xi^2\right] - 2\boldsymbol{x}_{it}^{T}\boldsymbol{\mu}_{zi}^{*}\boldsymbol{E}\left[\xi\right] + \left(\boldsymbol{x}_{it}^{T}\boldsymbol{\mu}_{zi}^{*}\right)^2 + \log(2\pi)\right\} - \log\Omega_{cit}^{(z)*},$$

where, $\xi$ is a random variable of the distribution (Grimmer 2010 b). Therefore,

$$\begin{aligned} L_u^{(p)} - L_u^{(q)} =& \boldsymbol{E}_{q_u, q_\beta, q_z}\left[\log p\left(\left\{u_{cit}^{(z)}\right\} \mid \{\boldsymbol{\beta}_{zi}, z_{cit}, \boldsymbol{x}_{it}, y_{cit}\}\right)\right] \\ & - \boldsymbol{E}_{q_u}\left[\log q_u\left(\left\{u_{cit}^{(z)}\right\} \mid \left\{u_{cit}^{(z)*}, \boldsymbol{x}_{it}, y_{cit}\right\}\right)\right] \\ =& -\frac{1}{2}\sum_{i=1}^{I}\left[Tr\left\{X_i X_i\left(\boldsymbol{\mu}_{zi}^{*}\boldsymbol{\mu}_{zi}^{*T} + V_{zi}^{\beta *}\right)\right\}\right] \\ & + \sum_{c=1}^{C}\sum_{i\in I_c}\sum_{t\in T_c}\left\{\frac{1}{2}\theta_{citz}^{*}\left(\boldsymbol{x}_{it}^{T}\boldsymbol{\mu}_{zi}^{*}\right)^2 + \theta_{citz}^{*}\log\Omega_{cit}^{(z)*}\right\}. \end{aligned}$$

The value of $L\left(\boldsymbol{\Theta}^{*}\right)$ is calculated by summation of the above ten expectations.

# Appendix C: Gibbs Sampler

The joint posterior distribution, assuming conditional independence between variables, provides the full conditional posterior distributions:

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

## C.1 Sampling of $\boldsymbol{\theta}_c$

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

$$p\left(\boldsymbol{\theta}_c \mid -\right) = p\left(\boldsymbol{\theta}_c\right) p\left(z_{cit} \mid \boldsymbol{\theta}_c\right) = \text{Diriclet}\left(\boldsymbol{n}_c + \tilde{\gamma}\right) \tag{C2}$$

## C.2 Sampling of $z_{cit} \mid -$

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

$$\Pr\left\{z_{cit} = z \mid \boldsymbol{\theta}_c, \{\boldsymbol{x}_{it}\}, \{\boldsymbol{\beta}_{zi}\}, \{y_{cit}\}\right\} = \frac{\theta_{cz}\Omega_{cit}^{(z)}}{\sum_{j=1}^{Z}\theta_{cz}\Omega_{cit}^{(j)}}, \tag{C3}$$

## C.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).

## C.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

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

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

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

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

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] Anderson, T. W. (2003). *An Introduction to Multivarite Statistical Analysis*, Wiley, U.S.A

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

[3] 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.

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

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

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

[7] Blei, D., and McAuliffe. J. (2007). "Supervised Topic Models," *Proceedings of Neural Information Processing System,* 3, 993-1022.

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

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

[10] 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.

[11] Corduneanu, A., and Bishop, C.M. (2001). "Variational Bayesian Model Selection for Mixture Distributions. In: Jaakkola, T., Richardson, T. (Eds.)", *Articial Intelligence and Statistics,* Morgan Kaufmann: Los Altos, CA, 2734.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] Puolamaki, K. and S. Kaski (2009). "Bayesian solutions to the label switching problem," In Advances in Intelligent Data Analysis VIII, 381-392, Springer

[18] 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.

[19] 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.

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

[21] Rust, R.T. and Chung, T.S. (2005). "Marketing Models of Service and Relationships," *Marketing Science,* 25, 560-580.

[22] 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.

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

[24] 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.

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

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

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

Table 1: Estimates of Simulation Data

| | Estimates (Posterior mean) | | | |
|---|---|---|---|---|
| | Intercept | Discount | Display | Feature |
| Segment 1 | -0.45 (0.03) | 0.89 (0.05) | -0.03 (0.05) | 0.04 (0.01) |
| Segment 2 | -0.50 (0.01) | 0.07 (0.04) | 0.91 (0.02) | 0.02 (0.02) |
| Segment 3 | -0.51 (0.01) | 0.04 (0.03) | 0.00 (0.03) | 0.93 (0.03) |

Simulated data ($C = 300$, $I = 50$, $T = 30$).

Table 2: Simulation Time by VB and MCMC

| | | VB | | | MCMC | | |
|---|---|---|---|---|---|---|---|
| | Z | 5 | 10 | 20 | 5 | 10 | 20 |
| | $C = 1000$ | | | | | | |
| | 100 | 0.6 | 0.8 | 1.1 | 5.3 | 7.1 | 14.2 |
| | 500 | 1.4 | 1.7 | 2.3 | 21.7 | 29.6 | 41.7 |
| | 1000 | 2.0 | 2.2 | 2.7 | 49.0 | 54.6 | 62.4 |
| | $C = 5000$ | | | | | | |
| $I$ | 100 | 2.1 | 2.3 | 3.0 | 23.4 | 30.3 | 46.8 |
| | 500 | 2.3 | 3.2 | 5.2 | 65.5 | 81.2 | 104.1 |
| | 1000 | 4.4 | 5.2 | 8.2 | 128.7 | 144.0 | 166.2 |
| | $C$= 10000 | | | | | | |
| | 100 | 3.5 | 4.2 | 5.7 | 49.4 | 67.9 | 102.5 |
| | 500 | 5.3 | 7.0 | 10.4 | 213.3 | 261.0 | 343.0 |
| | 1000 | 8.9 | 12.6 | 17.2 | 430.1 | 482.7 | 580.8 |

The number means hour.

Table 3: Joint Segmentation for Cross Category Analysis

| Segment 1 (C=31,I=9) | | | | | Segment 2 (C=114, I=28) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Brand | Category | Price | Display | Feature | Brand | Category | Price | Display | Feature |
| No.1 | Drink | .99 | .06 | .06 | No.6 | Desert | .94 | .13 | .06 |
| No.2 | Coffee | .89 | .10 | .02 | No.7 | Drink | .72 | .92 | .24 |
| No.3 | Iced noodle | .77 | .60 | .03 | No.6 | Desert | .94 | .17 | .04 |
| No.4 | Bean paste | .75 | .21 | .05 | No.6 | Desert | .93 | .22 | .05 |
| No.5 | Coke | .89 | .24 | .02 | No.6 | Desert | .93 | .19 | .06 |
| Segment 3 (C=22, I=4) | | | | | Segment 4 (C=28, I=6) | | | | |
| Brand | Category | Price | Display | Feature | Brand | Category | Price | Display | Feature |
| No.8 | Fish sausage | .93 | .08 | .08 | No.13 | Noodle | 89. | .23 | .05 |
| No.9 | Water | .60 | .47 | .04 | No.14 | Food | .90 | .03 | .01 |
| No.10 | Detergent | .69 | .20 | .26 | No.13 | Noodle | .78 | .09 | .11 |
| No.11 | Ice cream | .91 | .02 | .02 | No.15 | Fish sausage | .91 | .01 | .01 |
| No.12 | Water | .87 | .11 | .04 | 1No.6 | Drink | .87 | .11 | .04 |
| Segment 5 (C=24, I=5) | | | | | Segment 6 (C=26, I=6) | | | | |
| Brand | Category | Price | Display | Feature | Brand | Category | Price | Display | Feature |
| No.17 | Soup | .84 | .16 | .09 | No.20 | Drink | .81 | .29 | .17 |
| No.18 | Dressing | .76 | .72 | .09 | No.9 | Drink | .76 | .33 | .02 |
| No.19 | Ice cream | .76 | .57 | .22 | No.11 | Ice cream | .99 | .03 | .03 |
| No.18 | Dressing | .83 | .42 | .15 | No.20 | Drink | .75 | .31 | .17 |
| No.19 | Ice cream | .82 | .14 | .10 | No.21 | Drink | .64 | .73 | .11 |
| Segment 7 (C=67, I=14) | | | | | Segment 8 (C=267, I=68) | | | | |
| Brand | Category | Price | Display | Feature | Brand | Category | Price | Display | Feature |
| No.6 | Desert | .96 | .13 | .06 | No.12 | Cookie | .98 | .29 | .06 |
| No.14 | Food | .90 | .03 | .01 | No.22 | Coffee | .81 | .28 | .08 |
| No.12 | Sugar | .99 | .26 | .05 | No.20 | Ice cream | .89 | .36 | .02 |
| No.22 | Drink | .77 | .63 | .17 | No.23 | Dressing | .74 | .80 | .08 |
| No.20 | Drink | .75 | .52 | .16 | No.15 | Fish sausage | .91 | .01 | .01 |
| Segment 9 (C=946, I=332) | | | | | Segment 10 (C=124, I=28) | | | | |
| Brand | Category | Price | Display | Feature | Brand | Category | Price | Display | Feature |
| No.24 | Cleaner | .85 | .48 | .11 | No.27 | Drink | .99 | .25 | .11 |
| No.21 | Sauce | .74 | .35 | .07 | No.12 | Water | .87 | .26 | .01 |
| No.25 | Snack | .86 | .16 | .09 | No.11 | Ice cream | .99 | .03 | .03 |
| No.26 | Noodle | .68 | .98 | .09 | No.19 | Yoghurt | .88 | .10 | .16 |
| No.9 | Energy drink | .68 | .88 | .06 | No.25 | Curry | .67 | .98 | .08 |

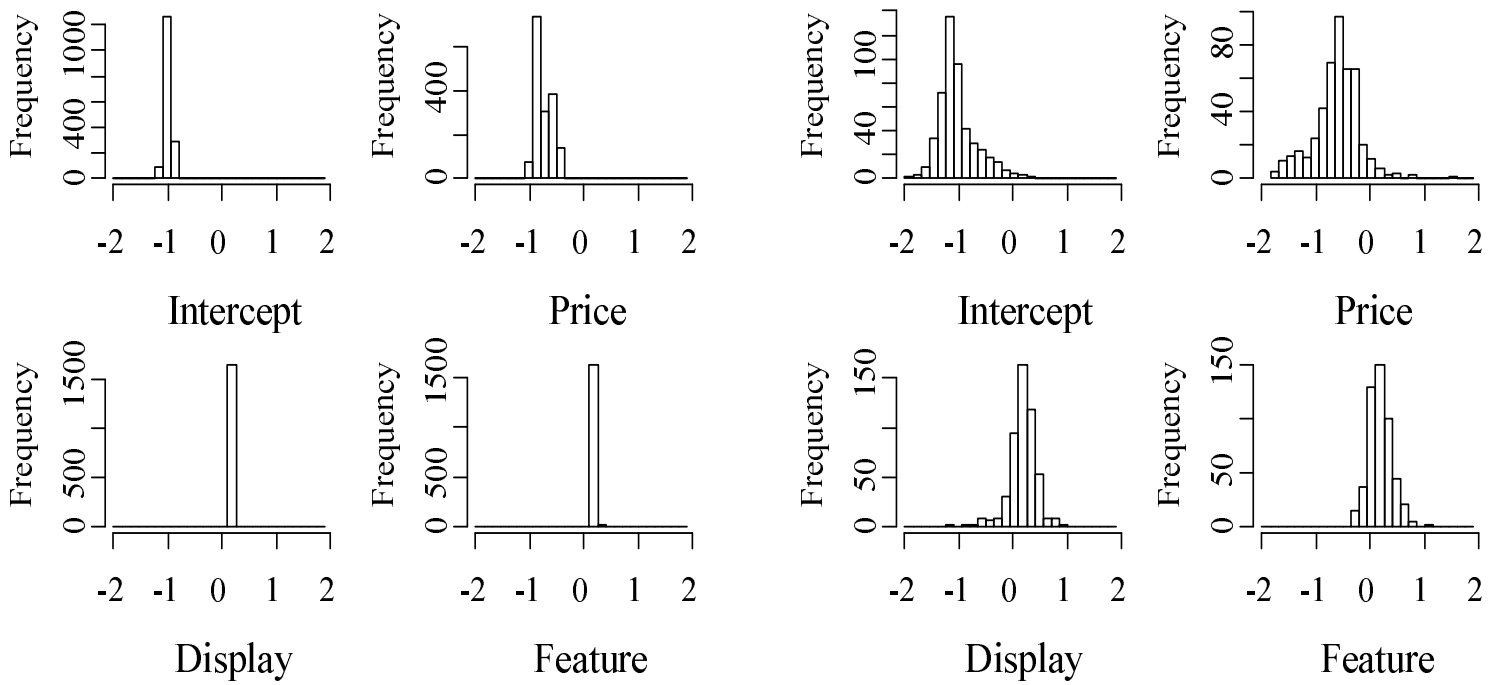


Figure 1: Marginal Distribution of Parameter Estimates of Individual Consumers and Items

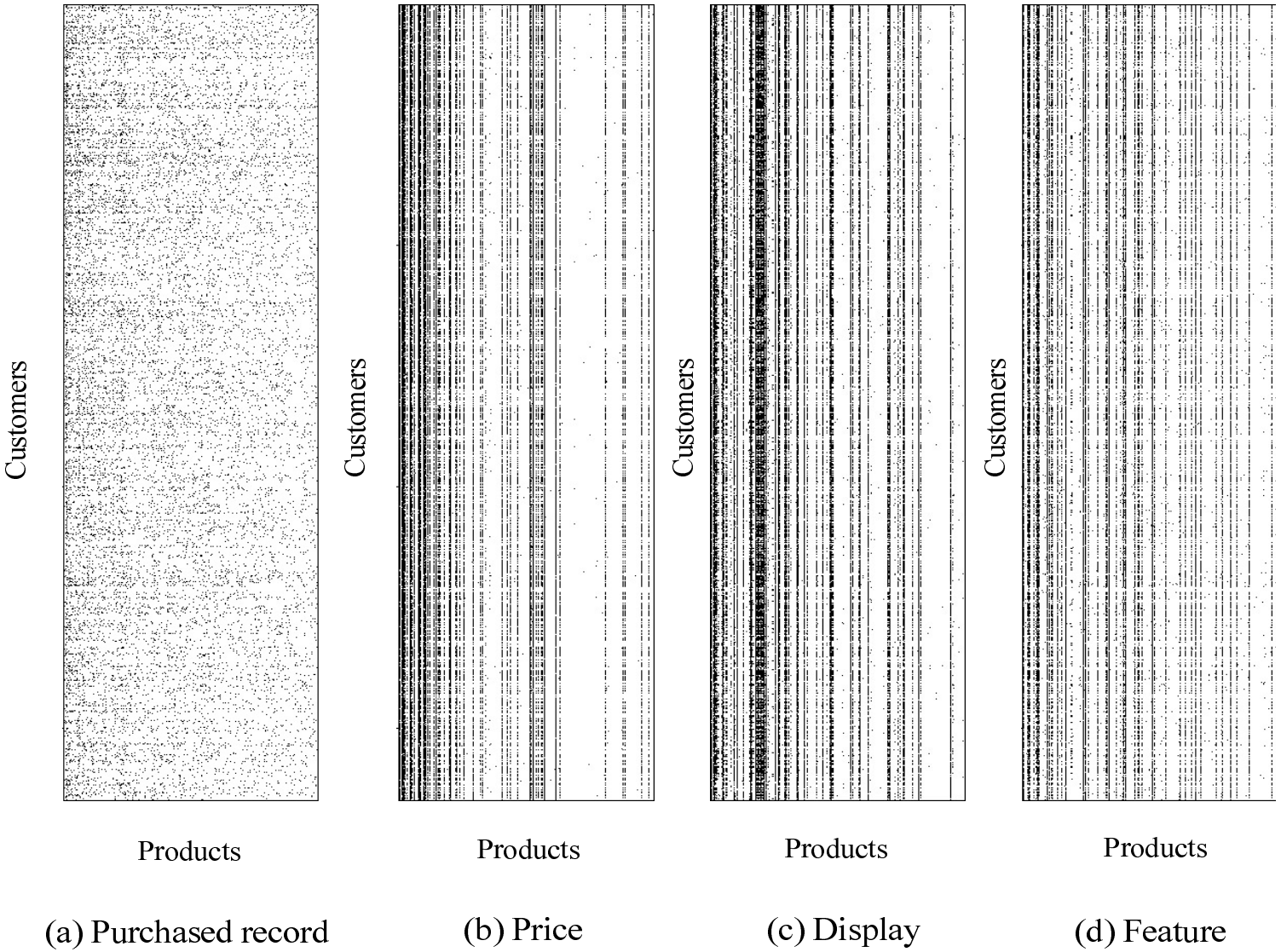


Figure 2: Personalized Effective Marketing Variables for Individual Consumers and Items: All Customers and Items

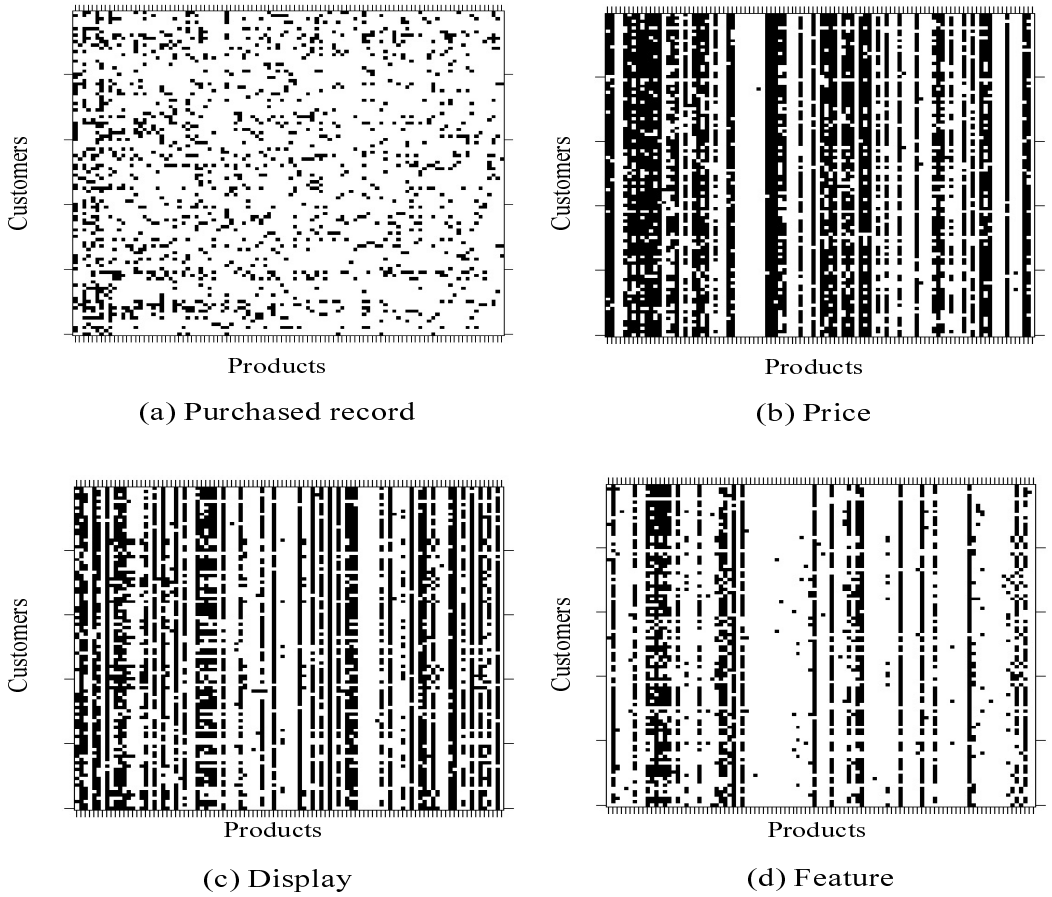


Figure 3: Personalized Effective Marketing Variables for Individual Consumers and Items: 100 Customers;100 Items