DSSR

Nobuhiko Terui
and
Yinxing Li

Data Science and Service Research
Discussion Paper

Measuring Large-Scale Market Responses from Aggregated Sales
- Regression Model for High-Dimensional Sparse Data -

Nobuhiko Terui[1]

and

Yinxing Li

February, 2017

Measuring Large-Scale Market Responses from Aggregated Sales
- Regression Model for High-Dimensional Sparse Data –

Abstract

In this article, we propose a regression model for high-dimensional sparse data from store-level aggregated POS systems. The modeling procedure comprises two sub-models—topic model and hierarchical factor regression model—that are applied sequentially not only for accommodating high dimensionality and sparseness but also for managerial interpretation.

First, the topic model is applied unusually to aggregated data to decompose the daily aggregated sales volume of a product into sub-sales for several topics by allocating each unit sale ("word" in text analysis) in a day ("document") into one of topics based on joint purchase information. This stage reduces the dimension of data inside topics because topic distribution is not uniform and product sales are allocated mostly into smaller numbers of topics. Next, the market response regression model in the topic is estimated by using information about other items in the same topic. That is, we construct a topic-wise market response function by using explanatory variables not only of itself but also of other items belonging to the same topic. Additional reduction of dimensionality remains necessary for each topic, and we propose a hierarchical factor regression model based on canonical correlation analysis for original high-dimensional sample spaces. Then we discuss feature selection based on a credible interval of parameters' posterior density.

Empirical study shows that (i) our model has the advantage of managerial implications obtained from topic-hierarchical factor regression defined according to their contexts, and (ii) it offers better fit than does conventional category regressions in-sample as well as out of sample.

Keywords: Topic Model, Hierarchical Factor Regression, Dimension Reduction, High Dimensional Sparse Data, Feature Selection

1. Introduction

Disaggregated store data of scanner panel records have been analyzed using many models and from many different perspectives. For example, many choice models have been proposed based on the theories of microeconomics and consumer behavior to understand customers and explore the effectiveness of the strategies suggested by said models.

The active use of daily aggregated store data—POS data that are accumulated automatically at customer check out points—is important for most merchandisers, even for those without a membership system. Most traditional methods of analyzing POS data specify market response function after the range of products is limited to a specific category and the number of product is smaller than the number of records (days). This category-based approach is useful when applied to products from well-recognized categories depending on the validity of the assumed categories. However, it cannot be applied to all products in a store, particularly to products that are purchased infrequently over observational periods. It is possible to lose useful information from store data when using the category-based approach.

By contrast, for this purpose, it is well known that scanning entire databases creates room for possibly discovering unexpected hidden patterns of joint purchases, which would bring new insights for marketing management by understanding the market baskets or shopping contexts of their customers. The POS data in a store contain records of the number of sales, prices, and promotions for 8,000 products. The direct use of these variables as covariates to explain sales is prohibited because the data is sparse and includes many zeros, and the size of the covariate matrix in the market response function is intractably large, that is, the entire data of a POS is Big Data. Even when the model can be estimated, overfitting occurs due to the so-called "N < P" problem, where N and P, respectively, refer to the numbers of samples and covariates. Thus, we need to generate smaller-sized datasets in some ways, for example, by decomposing larger ones into several datasets or reducing the dimension of the data matrix.

In this study, we relax these restrictions and do not assume category by applying entire products. The proposed model is composed of two sub-models. The first sub-model uses the topic model to reduce the dimension of the original data space by decomposing it into a prespecified number of sub-datasets and uncovers a hidden structure by aggregating

individual product purchases with store POS data, to which end we apply the disaggregated data model (topic model) to aggregated data. The second module solves the N<P problem by reducing the dimension of the covariate space, to which end we propose a hierarchical factor regression model interpretable as a Bayesian canonical correlation model to estimate the market structure in a lower-dimensional space between the dependent variable and the covariates, where market response functions are estimated regularly given that there are not many zeros in the reduced dimensional space. Finally, the market structure in the high-dimensional original data space is recovered by converting the estimated structure in a reduced-dimensional space to the original space. An overview of the proposed model is shown in Figure 1.

Figure 1: Overview of Model

In the next section, we apply the topic submodel to aggregated sales data and generate sub-datasets, each of which contain purchases pertaining to each topic. We interpret topic as shopping context in our study. These datasets already have smaller dimensions compared to those of the original data space because it is likely that products sales are not allocated to every topic evenly.

In Section 3, conditional on the datasets for topics, we use the second submodel to reduce dimensionality further to ensure that it is feasible to estimate topic-wise market response functions for individual topics. More specifically, topic-wise market response functions for products are estimated by using covariates related to variables of the products that belong to a given topic. This market response structure is estimated in a reduced-dimensional space and is then converted to the original space to extract sets of operational covariates that affect sales. In Section 4, we report an empirical study performed using POS data. The concluding remarks are given in Section 5.

2. Dimension Reduction Using Topic Model

2.1. Decomposing Aggregated Sales into Sub-sales by Shopping Context

Consumers have reasons when they purchase products in their shopping trips. Their

motivations, in other words, contexts, are buried in aggregated sales. For example, out of 50 sales of a chocolate, 15 could be consumers purchasing for themselves, 25 for gifting purchased jointly with a card, and 10 for cooking purchased jointly with flour. The decomposition of total sales into several contexts, that is, topics, leads to better understanding of the market and helps design efficient marketing strategies by targeting sub-sales in various contexts and conducting marketing to meet heterogeneous needs in the corresponding topics. We employ the topic model, which has been applied successfully to text analysis of natural language processing, for accommodating the latent topics in aggregated sales data.

The topic model is a reduced-dimensional model that has been applied successfully to text analysis for determining the frequency of "words" in "documents." We employ the latent Dirichlet allocation (LDA) model developed by Blei et al. (2003), which is well-established in natural language processing and used in a variety of disciplines. The LDA model is a generative model that allows for sets of observations to be explained by unobserved groups, explaining why some parts of the data are similar. It is based on the assumption that each document can be viewed as a mixture of various latent topics, where the topics follow a multinomial distribution over words. Let $w_{i|d}$ denote the *i-th* word in document *d* and $z_{i|d}$ the latent topic of the *i-th* word in document *d*. The model assumes that the vocabulary (*v*) distribution of $w_{i|d}$ in topic *k* follows a multinomial distribution ($w_{i|d} \sim Multi(\phi_{i|z_d})$) and $z_{i|d}$ follows a multinomial distribution ($z_{i|d} \sim Multi(\theta_d)$) in document *d*. Then, the model describes the probability that vocabulary *v* appears in document *d* and is represented as the sum of the products of topic distribution and vocabulary distribution in *K* possible ways:

$$p(v \mid d) = \sum_{k=1}^{K} p(v \mid k) p(k \mid d) = \sum_{k=1}^{K} \phi_{v|k} \theta_{k|d} \ . \tag{2}$$

The most common method of estimating the parameters $\theta_{k|d}$ and $\phi_{v|k}$ is Bayesian inference by using data augmentation of latent parameter **z** by way of full conditional posterior density to evaluate the posterior density of these parameters. When there is a large volume of text data like, as our study, Gibbs sampling of the parameters requires a

considerable amount of time. Therefore, we employ a collapsed Gibbs sampling, which analytically uses the natural conjugate of the prior distribution to integrate $\theta_{k|d}$ and $\phi_{v|k}$. $\{\phi_{v|k}, v = 1, \dots, V\}$ is the vocabulary distribution in topic k, and $\{\theta_{k|d}, k = 1, \dots, K\}$ is the topic distribution in document d.

We first apply the topic model to POS data so that one unit of an item purchased on each day in a store is allocated to one of the topics (segments) based on information about joint purchase with other items. That is, the number of sales of each product at time t is decomposed into several topics, and datasets are constructed for topic-wise market response functions.

In this study, one unit of an item purchased in each day in a store is allocated to one of the topics (segments) based on information about joint purchase with other products. The vocabulary v of text analysis corresponds to a product "j" and the document d means the day t in our study. Let Y_{jt} denote the number of sales of product "j" on day "t" and $\mathbf{Y}_t = \left(Y_{1t}, Y_{2t}, \dots, Y_{n_t}\right)'$ denote the vector of the total number of sales on the same day. In the context of text analysis, $\mathbf{Y}_t$ corresponds to the frequency vector of n_t types of vocabularies. The frequency of Y_{jt} of the j-th vocabulary is allocated to each topic proportional to the probability that each word from that vocabulary, i.e., unit sale of j-th product, belongs to each topic $p(j \mid t) = \sum_{k=1}^{K} p(j \mid k) p(k \mid t) = \sum_{k=1}^{K} \phi_{j,k} \theta_{k,t}$. Based on the allocation probability $\phi_{j,k} \theta_{k,t}$, we divide Y_{jt} into sub-sales in topic k, denoted by $Y_{jt}^{(k)}$, according to the allocation

$$Y_{jt}^{(k)} = Y_{jt} \times E\left(\phi_{j,k} \theta_{k,t}\right) \tag{3}$$

so that the aggregated sales are represented by the sum of topic-based sub-sales

$$Y_{jt} = \sum_{k=1}^{K} Y_{jt}^{(k)} \tag{4}$$

2.2. Topic-wise Market Response Function

We have "k Bags" of product sales according to topics and we build market response functions for the respective topics. Contrary to the usual market response functions, we

incorporate variables pertaining to other products in the same topic in the form of covariates because the topics are extracted from joint purchase information with other products. Then, we define the market response function of product item j in topic k, $Y_{jt}^{(k)}$, not only by its marketing variables such as price and promotions, but also other items' sales and their marketing variables:

$$Y_{jt}^{(k)} = \alpha_{0j}^{(k)} + \boldsymbol{\alpha}_j^{(k)}{}'\mathbf{X}_{jt} + \sum_{m \neq j} \beta_m^{(k)} Y_{mt}^{(k)} + \sum_{m \neq j} \boldsymbol{\gamma}_m^{(k)}{}'\mathbf{X}_{mt}^{(k)} + \varepsilon_{jt}^{(k)}, \qquad t = 1,...,T_j^{(k)} \quad (5)$$

where $\mathbf{X}_{jt}$ is the vector of the marketing variables of item j, $Y_{mt}^{(k)}$ is the number of sales of different items ($m \neq j$) allocated similarly to topic k, and $\mathbf{X}_{mt}^{(k)}$ is the vector of their marketing mix variables.

The forecast is defined by the sum of predictors constructed in each topic:

$$\hat{Y}_{it} = \hat{Y}_{it}^{(1)} + \hat{Y}_{it}^{(2)} + \cdots + \hat{Y}_{it}^{(K)} \qquad (6)$$

This step reduces the dimensions of the variable because the distribution of allocation probability $\left(\phi_{j,k}\theta_{k,t}\right)$ is usually not uniform but is concentrated on a few topics.

3. Dimension-Reduced Model for High-dimensional Market Responses

3.1 Hierarchical Factor Regression

Next, we consider the regression model for each topic between high-dimensional variables $\mathbf{Y}^{(k)}$ and $\mathbf{X}^{(k)}$ in topic k. We delete the superscript (k) for ease of reading hereafter.

Now, in case of the multivariate regression model for P_y-dimensional $\mathbf{Y}$ and P_x-dimensional $\mathbf{X}$

$$\mathbf{Y} = \mathbf{F}\mathbf{X} + \mathbf{e}, \qquad (7)$$

we consider the situation that the structural coefficient matrix $\mathbf{F}$ cannot be estimated directly due to high dimensionality of the original variable spaces. We define a class of model for the high-dimensional variables of $\mathbf{Y}$ and $\mathbf{X}$. First, we assume that they, respectively, have dimension-reduction models for defining their marginal distributions

$p(\mathbf{Y}|\mathbf{U},\mathbf{a})$ and $p(\mathbf{X}|\mathbf{V},\mathbf{b})$

$$\mathbf{Y}_i = \mathbf{U}\mathbf{a}_i + \mathbf{\eta}_{yi}, \qquad i = 1,...,N \qquad (8)$$

$$\mathbf{X}_i = \mathbf{V}\mathbf{b}_i + \mathbf{\eta}_{xi}, \qquad i = 1,...,N, \qquad (9)$$

where $\mathbf{a}_i$ is a $f_y(<P_y)$-dimensional vector and $\mathbf{U}$ is a $P_y \times f_y$ matrix, and $\mathbf{b}_i$ is $f_x(<P_x)$-dimensional vector and $\mathbf{V}$ is a $P_x \times f_x$ matrix. By stacking the vectors with respect to i to generate matrix forms, we have

$$\mathbf{Y} = \mathbf{U}\mathbf{a} + \mathbf{\eta}_y \qquad (10)$$

$$\mathbf{X} = \mathbf{V}\mathbf{b} + \mathbf{\eta}_x, \qquad (11)$$

where the augmented vectors are $\mathbf{Y} : P_y \times N$, $\mathbf{a} : f_y \times N$, $\mathbf{\eta}_y : f_y \times N$, $\mathbf{X} : P_x \times N$, $\mathbf{b} : f_x \times N$, and $\mathbf{\eta}_y : f_y \times N$, and we assume that $\mathbf{\eta}_y \sim N(0, \Sigma_y)$ and $\mathbf{\eta}_x \sim N(0, \Sigma_x)$.

As for the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$, we assume that they are conditionally independent of the common parameter H. That is,

$$p(\mathbf{Y},\mathbf{X}|\mathbf{U},\mathbf{a},\mathbf{V},\mathbf{b}) = p(\mathbf{Y}|\mathbf{U},\mathbf{a},\mathbf{H}) p(\mathbf{X}|\mathbf{V},\mathbf{b},\mathbf{H}) p(\mathbf{H}|\mathbf{a},\mathbf{b}). \qquad (12)$$

More specifically, in the structure equation (1) between $\mathbf{Y}$ and $\mathbf{X}$ in the original space, e is the error with zero mean and assumedly independent of $\mathbf{X}$.

Instead of dealing with (7), following Brynjarsdotter and Berliner (2014), we assume that there is a relationship in the reduced-dimensional space in terms of the hierarchical multivariate regression model:

$$\mathbf{a} = \mathbf{H}\mathbf{b} + \mathbf{\varepsilon} \qquad (13)$$

where the regression coefficient matrix H is of the order $f_y \times f_x$ and the error matrix $\mathbf{\varepsilon}$ is of the order $f_y \times N$ whose columns are assumed to independently follow $N(0, \sigma^2 I)$.

This reduced-dimensional space is called "crystallized space" in Brynjarsdotter and Berliner (2014). The motivation of the model is that the projection of $\mathbf{X}$ onto $\mathbf{Y}$ is infeasible, and direct inference on (7) is not applicable. They estimated $\mathbf{H}$ by using relationship (13) for forecasting $\mathbf{Y}$ as $\hat{\mathbf{Y}} = \mathbf{U}\hat{\mathbf{H}}\hat{\mathbf{b}} = \mathbf{U}\hat{\mathbf{H}}\mathbf{V}'\mathbf{X}$ under the restriction of

orthogonality on $\mathbf{V}$ in spatio-temporal modeling, but the estimation of structure $\mathbf{F}$ was out of the scope of their study. This model assumes the presence of a relationship $\mathbf{H}$ between $\mathbf{Y}$ and $\mathbf{X}$, and they are independent conditionally of $\mathbf{H}$. This model can be interpreted as a Bayesian multivariate canonical correlation model.

In the above, $\mathbf{H}$ plays the intermediate roles of $\mathbf{a}$ and $\mathbf{b}$ in the form of regression, and $\mathbf{Y}$ and $\mathbf{X}$ are independent conditionally of $\mathbf{H}$. That is, by considering that $\mathbf{a}$ and $\mathbf{b}$ are interpreted as the principal components or factors of $\mathbf{Y}$ and $\mathbf{X}$, they are determined by using their relations by way of $\mathbf{H}$.

3.2 Recovering Structure in High Dimensional Space

When $\mathbf{X}$ is given, structural equation (7) defines the conditional distribution. Under the assumption of conditional expectation of error term as $E_{|x}[\mathbf{e}] = 0$, the conditional probability measure of $\mathbf{Y}$ on $\mathbf{X}$ yields

$$E_{|x}[\mathbf{Y}] = \mathbf{FX} + E_{|x}[\mathbf{e}] = \mathbf{FX} \tag{14}$$

Then, by taking expectation with respect to the probability measure of $\mathbf{X}$,

$$E_x\{E_{|x}[\mathbf{Y}]\} = \mathbf{F}E_x[\mathbf{X}] \quad \text{i.e. } \boldsymbol{\mu}_y = \mathbf{F}\boldsymbol{\mu}_x \tag{15}$$

In terms of variables in the reduced-dimensional space, $E_x\{E_{|x}[\mathbf{Y}]\} = E[\mathbf{Y}] = \mathbf{Ua}$ and $E_x[\mathbf{X}] = \mathbf{Vb}$, (15) induces the following relationship

$$\mathbf{Ua} = \mathbf{FVb}. \tag{16}$$

In turn, the matrix $\mathbf{F}$ of the structure connecting the original variables in high-dimensional sample space is obtained by

$$\mathbf{F} = \mathbf{Uab'V'}(\mathbf{Vbb'V'})^{-1} \tag{17}$$

The joint prior density of factor model parameters $p(\mathbf{U}|\mathbf{a},\Sigma_y)p(\Sigma_y)$, $p(\mathbf{V}|\mathbf{b},\Sigma_x)p(\Sigma_x)$, $p(\mathbf{a}|\Lambda_a)p(\Lambda_a)$, and $p(\mathbf{b}|\Lambda_b)p(\Lambda_b)$ are also specified as normal-inverted gamma conjugate priors for Gibbs sampling, as is applied to standard factor models, e.g., by Lee (2007). Then, the posterior distributions of $p(\mathbf{U},\mathbf{a}|\mathbf{H},\mathbf{Y})$ and $p(\mathbf{V},\mathbf{b}|\mathbf{H},\mathbf{X})$ are derived by using the procedure of Bayesian factor models. Under the

assumption of normal-inverted gamma prior distribution for $p\left(\mathbf{H}\,|\,\Lambda_{h}\right)p\left(\Lambda_{h}\right)$, $p\left(\mathbf{H}\,|\,\mathbf{a},\mathbf{b},\sigma^{2}\right)$ is also available analytically, as is the posterior density of coefficient parameters in the normal linear regression model.

Then, the joint posterior density of all types of parameters in models (7)–(17) is represented by

$$
\begin{aligned}
p\left(\mathbf{U},\mathbf{a},\mathbf{V},\mathbf{b},\mathbf{H},\sigma^{2},\mathbf{F},\Sigma_{y},\Sigma_{x},\Lambda_{h}\,|\,\mathbf{Y},\mathbf{X}\right) &\propto \\
p\left(\mathbf{U},\mathbf{a}\,|\,\mathbf{H},\mathbf{Y}\right)&p\left(\mathbf{V},\mathbf{b}\,|\,\mathbf{H},\mathbf{X}\right)p\left(\mathbf{H}\,|\,\mathbf{a},\mathbf{b},\sigma^{2}\right)p\left(\mathbf{F}\,|\,\mathbf{U},\mathbf{a},\mathbf{V},\mathbf{b}\right) \\
&\times p\left(\sigma^{2}\,|\,\mathbf{a},\mathbf{b},\mathbf{H}\right)p\left(\mathbf{U}\,|\,\mathbf{a},\Sigma_{y}\right)p\left(\Sigma_{y}\right)p\left(\mathbf{V}\,|\,\mathbf{b},\Sigma_{x}\right)p\left(\Sigma_{x}\right) \\
&\times p\left(\mathbf{a}\,|\,\Lambda_{a}\right)p\left(\Lambda_{a}\right)p\left(\mathbf{b}\,|\,\Lambda_{b}\right)p\left(\Lambda_{b}\right)p\left(\mathbf{H}\,|\,\Lambda_{h}\right)p\left(\Lambda_{h}\right)p\left(\sigma^{2}\right)
\end{aligned}
\tag{18}
$$

We note that $p\left(\mathbf{F}\,|\,\mathbf{U},\mathbf{a},\mathbf{V},\mathbf{b}\right)$ is a degenerated density restricted as $\phi\left(\mathbf{U},\mathbf{a},\mathbf{V},\mathbf{b},\mathbf{F}\right)=0$ in (17). In order to recover $\mathbf{F}$ in the original high-dimensional space, we use the marginal posterior density $p\left(\mathbf{F}\,|\,\mathbf{Y},\mathbf{X}\right)$, which is marginalized numerically in terms of the MCMC procedure given by (18). The steps of the MCMC procedure are given in the appendix.

We propose a doubly reduced dimensional regression model by applying sequentially the topic model first and the hierarchical factor regression model thereafter. We call this *topic-hierarchical factor regression* in the following.

4. Empirical Application
4.1 Data

We applied the model to daily POS (point of sales) data of a store recorded between May 6, 2002, and May 6, 2003. The dataset contains information about 7,912 items over 363 days, and a total 3,720,419 purchases were recorded. POS data contain daily numbers of sale for purchased items and their price, and three types of marketing promotional variables, namely, two types of display and a feature. There are no records of marketing variables for items that were not purchased.

4.2 Topic Extraction

When we applied the first topic model, we set the number of topics $k = 10$. The LDA model was estimated by using the collapsed Gibbs sampler under the conjugate Dirichlet prior

distributions of hyper parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for topic distribution and vocabulary distribution, respectively.

$$\theta_d \sim Dir(\boldsymbol{\alpha})(d = 1, ..., M), \text{ and } \boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}) \ (k = 1, ..., K)$$

Following Griffiths and Steyvers (2004), we set $\boldsymbol{\alpha} = 50/k$ and $\boldsymbol{\beta} = 0.1$ for every element of their vectors.

The results of the topic model are given in Table 1 and shown in Figure 1. Table 1 lists the items with the top 20 highest probabilities, $\hat{\phi}_{j,k}$ for k = 1, …, 10, where the item is denoted as the product number in braces and the category name is shown. For example, the first item in Topic 1, "[6453]milk" indicates a product with identification number 6453 in the product items.

Table 1: Categories and Items in Topics

4.3 Topic-Hierarchical Factor Regression for a Milk Product

We applied the model to the sale of a brand of milk (JANcode:4902705065161), denoted Y, for developing the market response model. Thus, we consider the regression model with a univariate dependent variable. Y contains 21,482 total sales in 344 days. The first 324 days were used for estimation and last 20 days were used to validate the estimate.

Figure 1: Sales of a Milk Product (JAN: 4902705065161)

Figure 1 shows the time series plot of Y. The product is sold regularly with significant spikes on some days, which could be caused by store promotional activity, and the level of sales remains high over the summer season from July (60th day) to August (70th day).

Figure 2: Averaged Topic Distribution for a Product

Figure 2 shows the topic distribution of a product averaged over entire days, specifically defined by

$$\hat{\phi}_{j,k}\hat{\theta}_{k|\bullet} = \frac{1}{N}\sum_{t=1}^{N}\hat{\phi}_{j,k}\hat{\theta}_{k|t}, k=1,...,K, \qquad (19)$$

where $N = 344$ and $K = 10$. We can observe that the topics have nearly equal weights, except for the most frequent topic at $k = 4$ and the least frequent topic at $k = 8$.

Figure 3: Topic Sales Decomposition

Figure 3 shows the time series of topic-wise sales, $Y_{jt}^{(k)}, k=1,...,10$, which were generated by using the averaged topic distributions. The time series patterns exhibit seasonality with ranges from one to three months, although a few of the ranges contain overlapping periods. Topic 8 is rather exceptional with regularly sold pattern of relatively small numbers. We note that the allocated numbers are not integers in general according to formula (3).

To be specific, Topic 1 refers to the sales between September and October, and the sales between January and March are classified in Topic 2. Topic 4 refers to the sales in the summer vacation period between July and August and the period between January and February, and Topic 10 refers to the sales at the start of the fiscal year in April. The sales in summer and early autumn are covered by Topic 7, and so on.

Combined time series plots of topics with averaged topic distribution and LDA topic model decompose the sales over the entire year into clearly distinguished submarkets defined by the seasons, and among all, summer is most important season for the target product as mostly four times sales as other topics.

4.4 Model Comparison

In empirical analysis, we set a univariate Y. We then defined the number of reduced-dimensional space of X as $f_x = 20$ for every topic. According to the topic-hierarchical factor regression model, the estimated dependent variable in topic k is calculated by evaluating the posterior means of related parameters and covariates

$$\hat{Y}_t^{(k)} = E\left(\mathbf{F}^{(k)}\mathbf{X}_t^{(k)}\right) = E\left(\mathbf{U}^{(k)}\mathbf{a}^{(k)}\mathbf{b}^{(k)}{}'\mathbf{V}^{(k)}{}'\left(\mathbf{V}^{(k)}\mathbf{b}^{(k)}\mathbf{b}^{(k)}{}'\mathbf{V}^{(k)}{}'\right)^{-1}\mathbf{X}^{(k)}{}_t\right) \quad (17)$$

and the posterior density of forecast is defined as $p\left(\hat{Y}_t \mid Data\right) = p\left(\sum_{k=1}^{10} \hat{Y}_t^{(k)} \mid Data\right)$.

In the above, the covariate $\mathbf{X}_t^{(k)}$ in topic k contains four marketing variables, in addition to the joint sales of other products and their marketing variables.

The model fit is evaluated by the root mean square error $RMSE = \sqrt{\sum_{s=1}^{S}\left(Y_s - E\left(\hat{Y}_s\right)\right)^2 / S}$ for in-sample and out-of-sample data, where the posterior mean $E\left(\hat{Y}_s\right)$ is used as a point forecast. Table 2 shows the RMSEs of in-sample and out-of-sample forecasts for various models.

The first model is *topic-hierarchical factor regression*. The structural regression model defined by (7) in the original space contains 39,560 parameters, obtained as 7,912 item numbers multiplied by five types of covariates, that is, the target item's marketing variables and the sale of jointly purchased other items and their marketing variables.

Table 2: Model Fit and Comparisons

The model's performance can be improved by extracting effective covariates to reduce the number of model parameters and increase the degrees of freedom. Then, the second proposed model contains a reduced number of covariates after eliminating covariates with insignificant parameter estimates. We evaluated the posterior density $p\left(\mathbf{F} \mid Data\right)$ of the regression coefficient, and we can apply the 95% credible interval test for determining the significance of coefficients to increase the number degrees of freedom. Specifically, we selected covariates with estimated coefficient parameters based on the criterion that 95% of the central region of posterior density does not include zero.

Table 2 shows the total number of effective parameters, and Table 3 shows the number of covariates for the respective models in the 10 topics. In total, 2,131 parameters remained as effective variables. The number of parameters decreased dramatically to approximately 5.4% of the number of parameters in the original space. The numbers of selected covariates in the respective topics were almost proportional to the volumes of data allocated to the topics.

Then, we re-estimate the model by eliminating insignificant variables following the same procedure as before.   We call this model *effective topic-hierarchical factor regression*.

The third alternative model uses only the top 20 items with the highest topic probability in each topic, called *topic regression*.   This selection was followed by studies on topic model analysis in natural language processing, adopted usually in machine learning studies.   The estimation process is the same, except the 2nd reduced-dimensional factor model is not used. The fourth alternative model assumes a category a priori for the target item and uses only items from the same category.   This model has been used conventionally in marketing, and it is called *category regression*.   This is the benchmark model.   This category contains 48 items. In addition to four marketing variables, the model contains 47 times 5 covariates of other items sales and their marketing variables, so 239 is the total number of covariates.

Table 2 shows the results of model fit in terms of RMSE.   First, the conventionally used category regression model performs the worst in in-sample and out-of-sample forecasts.   This means that shopping contexts expressed by the topics are likely to have important roles, and the narrowly defined category-based analysis loses useful information hidden in purchase records. The comparison of (i) topic-factor and (iii) topic regression models indicates rather mixed results in the in-sample and the out-of–sample forecasts. Although both models use information of all items when topics are extracted, (ii) uses only 20 variables in regression, and it might lose information in forecasting, even though it is adequate to explain in-sample data.

Finally, the comparison of (i) the topic-factor and (ii) the efficient topic-factor models clearly supports (ii), in particular, for out-of-sample forecasting.

4.5 Efficient Topic-Factor Reduced Dimensional Regression

Model (ii) is supported as the best, and thus, we report the estimation results of that model in greater detail in this section.

Figure 5: Estimated Sales in Topics

Table 4(1), (2): Feature Selection for Topics

Figure 5 shows the in-sample fit for the respective topics. In addition to fit for total sale in Figure 4, we observe that the model performs well for every topic. Table 4(1) and (2) indicate the results of feature selection in terms of 95% credible interval for recovered posterior density. The significant estimates of "sales", "price", "promotion 1," and "promotion 2" are tabulated up to ten in the order of absolute magnitude. The estimated coefficients are given after the name of category and product ID number in parentheses. The coefficient means elasticity for log-transformed sales and price, and it holds approximately for other binary promotional variables. We observed the following:

(i)   The sales variables of other products are identified as effective covariates across topics. This is reasonable because we modeled the regression by using joint purchase patterns.

(ii)   "price" variables are chosen as effective in topic 6 and in many cases in topic 10. All of them are cross elastic, and self-elasticity is not significant. The price of this targeted product does not change much. Conventional category regression shares this result for self-elasticity.

(iii)   "promotions" are effective for topics 5, 6, 7, 9, and 10.

(iv)   Topic 9 has the most number of effective covariates. The sales of foodstuffs are selected as features, and we interpret this fact as milk is purchased not for drinking but for cooking. Promotions of chocolate and other drinks are effective.

(v)   In topic 4, which has the most volume, the sales of water and juice are synchronized with the sale of the target product and are opposite to the sales of chocolates and cookies from June to August. This may be consistent with the seasonal pattern of summer from June to August.

Table4(1): Feature Selection for Topics

Table 4(2): Feature Selection for Topics

5. Concluding Remarks

The volume of high-dimensional data is growing with progress in network technology. These high-dimensional data contain many zeros, that is, sparse in data spaces, as is the case with POS data in our study.   That is, the daily number of product items purchased is considerably smaller than that of items displayed in a store, and many items are recorded as having zero sales and information about their price and promotional variables are not recorded, thus producing yet another data entry with zero value.   POS data are used to find effective relationships between product items and promotional variables to facilitate efficient management.   In these cases, statistical models that accommodating simply high-dimensional data do not work.

We proposed a regression model for high-dimensional and sparse data.   In the proposed model, two types of dimension-reduction models, namely, topic model and canonical correlation model, are used sequentially.   More specifically, the store-level aggregated sales and marketing mix data are analyzed using a model that combines the topic model with the regression model.   First, the topic model decomposes sales of product items into several topics by allocating each unit sale ("word" in text analysis) in a day ("document") into one of topics based on joint purchase information.

Next, conditional to the topic model estimates, the market response function of an item is estimated by using information regarding variables on the items inside a topic.   That is, we construct topic-wise market response functions of an item by using explanatory variables not only of said item but also of the other items belonging to the same topic.   In addition, the explanatory variables include not only marketing mix variables but also the numbers of sales of product items in the same topic.   We can expect unusual findings from these response functions because sets of related items are not considered in advance and they are not obtained by conventional category-based market response functions.

Finally, we forecast the sales of an item by summing up respective predictors constructed

for topics.　　We compared the model performance with that of conventionally predetermined category regression model to show that (i) our model has the advantage of offering managerial implications obtained from topic-wise regressions defined according to their contexts, and (ii) it has better fit than does the category regression model.

　　We have basically kept the number of topics for topic model and the numbers of factors for hierarchical factor regression model as fixed in the empirical application, mostly due to the restrictions on computational times.　　By using R under usual PC environment, it took four days for estimating topic model, and two days for the part of hierarchical factor regression model when numerically calculating inverse of large scale matrix to recover the structure in original space.　　We leave this model selection problem for future research by developing more efficient computation procedures, although we currently recognize that the estimation procedure of second submodel's part could be eased if we employ parallel computing for topic wise market response functions.

Reference

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

Blei, D. M. (2012), "Introduction to Probabilistic Topic models," *Communications of the ACM*, 55, 77-84.

Brynjarsdottir, J. and Berliner, L.K. (2014), "Dimension-reduced modeling of spatio-temporal process," *Journal of American Statistical Association*, vol. 109, 1647-1659.

Iwata,T. and Sawada H.(2013), "Topic model for analyzing purchase data with price information," *Data Mining and Knowledge Discovery*, 26, 559-573.

Lee, S. Y. (2007), *Structural Equation Modeling: A Bayesian Approach*, Wiley, New York.

Hanssen, D.M., L.J. Parsons and R.L. Shultz (2001), *Market Response Models: Econometric and Time Series Analysis,* 2nd ed., Boston, MA: Kluwer Academic Press Inc.

Rossi, P.E., G. Allenby and R. McCulloch(2005), *Bayesian Statistics and Marketing*, Wiley, New York."

Appendix: MCMC Algorithm

1. LDA Topic Model

Under the prior distributions of Dirichlet distributions on hyperparameter α, β $\theta_d \sim Dir(\boldsymbol{\alpha})$ $(d = 1, \ldots, M)$, and $\phi_k \sim Dir(\boldsymbol{\beta})$ $(k = 1, \ldots, K)$, where M is the number of documents and K is the number of topics. Following Griffiths and Steyvers (2004), we set the vectors with $\boldsymbol{\alpha} = 50/K$ and $\boldsymbol{\beta} = 0.1$. Then, we used the collapsed Gibbs sampler to obtain posterior density in the closed form as

$$p\left(z_{d,i} = k \mid w_{d,i} = v, \mathbf{w}^{\backslash d,i}, \mathbf{z}^{\backslash d,i}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{n_{k,v}^{\backslash d,i} + \beta_v}{n_{k,\cdot}^{\backslash d,i} + \sum_{v'} \beta_{v'}} \frac{n_{d,k}^{\backslash d,i} + \alpha_k}{n_d^{\backslash d,i} + \sum_{k'} \alpha_{k'}},$$

where $w_{d,i}$ means word i in document d, $z_{d,i}$ is the latent topic of word i in document d,

θ_d means topic distribution of document d, and ϕ_k is the vocabulary distribution of topic

k. $\mathbf{w}^{\backslash d,i}$ is the set of all words from the text data except word $w_{d,i}$, $\mathbf{z}^{\backslash d,i}$ is the set of all topics

except $z_{d,i}$, n_d is the number of words in document d, $n_{k,v}^{\backslash d,i}$ refers to the frequency of

word v in topic k except for that of word i in document d, and $n_{k,\cdot}^{\backslash d,i} = \sum_v n_{k,v}^{\backslash d,i}$.

2. Hierarchical Factor Regression Model

2.1 Prior Density

Prior	Setting
$\mathbf{a_k} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Phi_0})$, $\mathbf{b_k} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Phi_0})$	$\boldsymbol{\mu_0} \sim 0_{K \times 1}, \boldsymbol{\Phi_0} = I_{K \times K} \times 100$
$\mathbf{Y} \sim N(\mathbf{a_0}, \boldsymbol{\Sigma_y})$, $\mathbf{X} \sim N(\mathbf{b_0}, \boldsymbol{\Sigma_x})$	$\mathbf{a_0} = 0, \mathbf{b_0} = 0$
$\boldsymbol{\Sigma_y} \sim IG(\mathbf{r_0}, \mathbf{s_0})$, $\boldsymbol{\Sigma_x} \sim IG(\mathbf{r_0}, \mathbf{s_0})$	$r_0 = 2, s_0 = 2$
$a \sim N(\boldsymbol{\mu_0}, \sigma_0^2 \boldsymbol{\Lambda_0}^{-1})$	$\boldsymbol{\mu_0} = 0$, $\boldsymbol{\Lambda_0} = I \times 100$
$\Phi x, \Phi y \sim IW_q(\mathbf{R_0}', \rho_0)$	$\mathbf{R_0} = I, \rho_0 = 100$
$\sigma_0^2 \sim IG(\alpha_0, \beta_0)$	$\alpha_0 = 2, \beta_0 = 2$

2.2 Conditional Posterior Distribution for MCMC

(1) $a_i \mid y_i, \mathbf{U}, \mathbf{\Sigma_y}, \mathbf{\Phi_y}$

$$p\big(a_i \mid y_i, \mathbf{U}, \mathbf{\Sigma_y}, \mathbf{\Phi_y}\big) \propto N\Big(\big(\mathbf{\Phi_y^{-1}} + \mathbf{U'}\mathbf{\Sigma_y'}\mathbf{U}\big)^{-1}\mathbf{U'}\mathbf{\Sigma_y'}\mathbf{U}, \big(\big(\mathbf{\Phi_y^{-1}} + \mathbf{U'}\mathbf{\Sigma_y'}\mathbf{U}\big)^{-1}\big)\Big)$$

(2) $\mathbf{U_k} \mid \mathbf{Y}, \mathbf{a}, \mathbf{\Sigma_{yk}^{-1}}$

$$p(\mathbf{U_k} \mid \mathbf{Y}, \mathbf{a}, \mathbf{\Sigma_{yk}^{-1}}) \sim N\Big(\big(\mathbf{a_k^*}\mathbf{a_k^*}'\big)^{-1}\mathbf{a_k^*}\mathbf{Y_k^*}, \ \ \mathbf{\Sigma_{yk}}\big(\mathbf{a_k^*}\mathbf{a_k^*}'\big)^{-1}\Big)$$

(3) $\mathbf{\Sigma_{\psi\kappa}} \mid \mathbf{Y}, \mathbf{a}$

$$p\big(\mathbf{\Sigma_{yk}} \mid \mathbf{Y}, \mathbf{a}\big) \sim IG\Big((n - r_k)/2, s_0 + 2^{-1}\mathbf{Y_k^*}'\big(I_n - \mathbf{a_k^*}'\big(\mathbf{a_k^*}\mathbf{a_k^*}'\big)^{-1}\mathbf{a_k^*}\big)\mathbf{Y_k^*}\Big)$$

(4) $\mathbf{\Phi_y} \mid \mathbf{Y}, \mathbf{U}$

$$p\big(\mathbf{\Phi_y} \mid \mathbf{Y}, \mathbf{U}\big) = IW_q\big(\mathbf{aa'} + \mathbf{R_0'}, n + \rho_0\big)$$

(5) $b_i \mid x_i, \mathbf{V}, \mathbf{\Sigma_x}, \mathbf{\Phi_x}$

$$p\big(b_i \mid x_i, \mathbf{V}, \mathbf{\Sigma_x}, \mathbf{\Phi_x}\big) \propto N\Big(\big(\mathbf{\Phi_x^{-1}} + \mathbf{V'}\mathbf{\Sigma_x'}\mathbf{V}\big)^{-1}\mathbf{V'}\mathbf{\Sigma_x'}\mathbf{V}, \big(\big(\mathbf{\Phi_x^{-1}} + \mathbf{V'}\mathbf{\Sigma_x'}\mathbf{V}\big)^{-1}\big)\Big)$$

(6) $\mathbf{V_k^*} \mid \mathbf{X}, \mathbf{b}, \mathbf{\Sigma_{xk}^{-1}}$

$$p(\mathbf{V_k^*} \mid \mathbf{X}, \mathbf{b}, \mathbf{\Sigma_{xk}^{-1}}) \sim N\Big(\big(\mathbf{b_k^*}\mathbf{b_k^*}'\big)^{-1}\mathbf{b_k^*}\mathbf{X_k^*}, \ \ \mathbf{\Sigma_{xk}}\big(\mathbf{b_k^*}\mathbf{b_k^*}'\big)^{-1}\Big)$$

(7) $\mathbf{\Sigma_{xk}} \mid \mathbf{X}, \mathbf{b}$

$$p\big(\mathbf{\Sigma_{xk}^*} \mid \mathbf{X}, \mathbf{b}\big) \sim IG\Big((n - r_k)/2, 2^{-1}\mathbf{X_k^*}'\big(I_n - \mathbf{b_k^*}'\big(\mathbf{b_k^*}\mathbf{b_k^*}'\big)^{-1}\mathbf{b_k^*}\big)\mathbf{X_k^*}\Big)$$

(8) $\mathbf{\Phi_x} \mid \mathbf{X}, \mathbf{V}$

$$p\big(\mathbf{\Phi_x} \mid \mathbf{X}, \mathbf{V}\big) = IW_q\big(\mathbf{bb'} + \mathbf{R_0'}, n + \rho_0\big)$$

(9) $\mathbf{H} \mid \mathbf{a}, \mathbf{b}, \sigma^2$

$$p\big(\mathbf{H} \mid \mathbf{a}, \mathbf{b}, \sigma^2\big) \propto N\Big(\big(\mathbf{b'b} + \mathbf{\Lambda_0}\big)^{-1}\big(\mathbf{\Lambda_0}\mu_0 + \mathbf{b'a}\big), \big(\mathbf{b'b} + \mathbf{\Lambda_0}\big)^{-1}\Big)$$

(10) $\mathbf{\Lambda_n} \mid \mathbf{\Lambda_0}$

$$\mathbf{\Lambda_n} = \mathbf{b'b} + \mathbf{\Lambda_0}$$

(11) $\sigma^2 \mid a_i, b_i, \mathbf{H}$

$$p\left(\sigma^2 \mid a_i, b_i, \mathbf{H}\right) \propto IG\left(\alpha_n + \frac{n}{2}, \beta_n + \frac{1}{2}\left(\mathbf{a'a} + \boldsymbol{\mu_0}'\boldsymbol{\Lambda_0}\boldsymbol{\mu_0} - \boldsymbol{\mu_n}'\boldsymbol{\Lambda_n}\boldsymbol{\mu_n}\right)\right),$$

where

$$\boldsymbol{\mu_n} = \left(\mathbf{b'b} + \boldsymbol{\Lambda_0}\right)^{-1}\left(\boldsymbol{\Lambda_0}\boldsymbol{\mu_0} + \mathbf{b'b}\hat{\mathbf{H}}\right) = \left(\mathbf{b'b} + \boldsymbol{\Lambda_0}\right)^{-1}\left(\boldsymbol{\Lambda_0}\boldsymbol{\mu_0} + \mathbf{b'a}\right), \boldsymbol{\Lambda_n} = \left(\mathbf{b^T b} + \boldsymbol{\Lambda_0}\right),$$

$$\alpha_n = \alpha_0 + \frac{n}{2}, \text{ and }, \quad \boldsymbol{\beta_n} = \boldsymbol{\beta_0} + \frac{1}{2}\left(\mathbf{a'a} + \boldsymbol{\mu_0}'\boldsymbol{\Lambda_0}\boldsymbol{\mu_0} - \boldsymbol{\mu_n}'\boldsymbol{\Lambda_n}\boldsymbol{\mu_n}\right).$$

Table 1: Categories and Items in Topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[6453]milk	[7262]milk	[4912]milk	[6453]milk	[4808]milk
[3630]milk	[6455]milk	[6453]milk	[6615]water	[4787]yogurt
[6524]energy drink	[3630]milk	[3630]milk	[4928]yogurt	[7262]milk
[6795]cola	[6795]cola	[4787]yogurt	[3630]milk	[3596]milk
[3596]milk	[3596]milk	[6795]cola	[6795]cola	[6795]cola
[6800]carbonated flavor	[6800]carbonated flavor	[3596]milk	[3596]milk	[3630]milk
[4787]yogurt	[5506]cup noodle	[6800]carbonated flavor	[6800]carbonated flavor	[6800]carbonated flavor
[4928]yogurt	[5879]yogurt	[6455]milk	[4787]yogurt	[4928]yogurt
[6622]Janpanese tea and	[4950]yogurt	[4928]yogurt	[4933]yogurt	[3597]milk
[463]yogurt	[3835]cup noodle	[463]yogurt	[3496]cup noodle	[4786]milk
[6798]Chinese tea drink	[463]yogurt	[3712]cola	[3729]Janpanese tea and	[5506]cup noodle
[3496]cup noodle	[3833]cup noodle	[5895]coffee drink	[3238]water	[6798]Chinese tea drink
[453]yogurt	[5887]yogurt	[6798]Chinese tea drink	[1846]coffee drink	[4933]yogurt
[2895]mayonnaise	[6798]Chinese tea drink	[3497]cup noodle	[3706]cola	[463]yogurt
[2162]vegetable juice	[5946]lactic acid bacteria	[3496]cup noodle	[6798]Chinese tea drink	[6524]energy drink
[6787]water	[4997]yogurt	[5946]lactic acid bacteria	[3483]cup noodle	[5946]lactic acid bacteria
[3706]cola	[3735]sports drink	[6833]coffee drink	[463]yogurt	[431]water
[3497]cup noodle	[3605]yogurt	[3605]yogurt	[6622]Janpanese tea and	[1064]yogurt
[5873]milk	[6622]Janpanese tea and	[3164]cup noodle	[5946]lactic acid bacteria	[3288]Janpanese tea and
[5946]lactic acid bacteria	[4915]milk	[3319]sports drink	[3497]cup noodle	[3712]cola

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
[7262]milk	[6453]milk	[408]cup noodle	[4808]milk	[6455]milk
[6795]cola	[3630]milk	[411]cup noodle	[3630]milk	[3630]milk
[3630]milk	[3596]milk	[3630]milk	[7262]milk	[7262]milk
[6800]carbonated flavor	[4787]yogurt	[3635]milk	[3596]milk	[4808]milk
[3596]milk	[6795]cola	[409]cup noodle	[6795]cola	[3596]milk
[4928]yogurt	[6800]carbonated flavor	[3850]cup noodle	[6800]carbonated flavor	[4912]milk
[4787]yogurt	[4928]yogurt	[5756]soy	[3236]Chinese tea drink	[6795]cola
[3735]sports drink	[3497]cup noodle	[2895]mayonnaise	[3735]sports drink	[6800]carbonated flavor
[6624]Janpanese tea and	[463]yogurt	[6746]salad oil	[4787]yogurt	[4787]yogurt
[5887]yogurt	[3496]cup noodle	[3735]sports drink	[5873]milk	[6453]milk
[4808]milk	[6798]Chinese tea drink	[3706]cola	[463]yogurt	[5946]lactic acid bacteria
[451]personal ice (cream)	[5946]lactic acid bacteria	[4243]soup	[3712]cola	[6822]carbonated flavor
[453]yogurt	[4912]milk	[2397]lactic acid bevera	[5879]yogurt	[6557]energy drink
[463]yogurt	[4997]yogurt	[6624]Janpanese tea an	[3706]cola	[4997]yogurt
[3706]cola	[4209]Janpanese tea and	[3497]cup noodle	[3734]sports drink	[463]yogurt
[5506]cup noodle	[3503]cup noodle	[2725]soy	[5895]coffee drink	[5887]yogurt
[6787]water	[3712]cola	[5887]yogurt	[6787]water	[4928]yogurt
[6622]Janpanese tea and	[1064]yogurt	[3496]cup noodle	[6798]Chinese tea drink	[6872]cola
[450]personal ice (cream)	[6833]coffee drink	[4289]instant curry	[5946]lactic acid bacteria	[5895]coffee drink
[5879]yogurt	[3279]energy drink	[4224]salad oil	[5887]yogurt	[5506]cup noodle

Table 2: Model Fit and Comparisons

Model	RMSE1(In Sample)	RMSE2(Out of Sample)	Number of Covariates
(a) Topic-Factor Dimension-Reduced	50.41	14.403	39,560(200)
(b) Effective T-F Dimension-Reduced	45.622	8.168	2131(200)
(c) Topic Regression	46.617	19.164	1000
(d) Category Regression	93.472	33.956	239

Table 3: Number of Effective Covariates

Topic:k	1	2	3	4	5	6	7	8	9	10
Number of Variables	190	205	235	284	179	281	286	74	239	158

Table 4(1): Feature Selection for Topics

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
1	[4891]miso	2.318	[3064]rice cracker	0.079	[3064]rice cracker	-0.019	[2092]tomato juice	-0.089	-	-
	[4929]coffee drink	2.184	[3523]cup noodle	0.030	[1639]cup noodle	0.010	[235]yogurt	-0.043	-	-
	[5025]rice cracker	2.141	[7211]barbecue sauce	-0.001	-	-	-	-	-	-
	[5577]suger	2.093	-	-	-	-	-	-	-	-
	[3249]energy drink	1.946	-	-	-	-	-	-	-	-
	[7034]coffee drink	1.942	-	-	-	-	-	-	-	-
	[1927]family ice	-1.885	-	-	-	-	-	-	-	-
	[3287]fruit juice	1.764	-	-	-	-	-	-	-	-
	[494]sauce	1.709	-	-	-	-	-	-	-	-
	[7384]premium ice	1.649	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
2	[5790]biscuit and cookie	3.000	[6012]cheese	-0.014	[2245]snack	0.154	[5550]cup noodle	-0.260	-	-
	[4919]family ice (cream)	1.948	-	-	[2553]chocolat	-0.090	[3128]cup noodle	-0.123	-	-
	[3846]cup noodle	1.764	-	-	-	-	[2112]instant soup	0.108	-	-
	[1345]chocolat	1.764	-	-	-	-	-	-	-	-
	[2727]soy	1.680	-	-	-	-	-	-	-	-
	[1826]dressing	1.674	-	-	-	-	-	-	-	-
	[89]coffee drink	1.532	-	-	-	-	-	-	-	-
	[3587]personal ice	1.462	-	-	-	-	-	-	-	-
	[6860]instant coffee	-1.460	-	-	-	-	-	-	-	-
	[1413]biscuit and cookie	-1.389	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
3	[1287]cooked curry	-2.088	-	-	[3225]salad oil	0.486	[2031]dressing	-0.178	-	-
	[4825]lactic acid bacteria b	1.913	-	-	-	-	[2030]dressing	0.145	-	-
	[1138]mayonnaise	-1.607	-	-	-	-	-	-	-	-
	[494]sauce	1.516	-	-	-	-	-	-	-	-
	[6402]chocolat	1.379	-	-	-	-	-	-	-	-
	[3116]candy	1.352	-	-	-	-	-	-	-	-
	[469]dessert	-1.316	-	-	-	-	-	-	-	-
	[166]personal ice (cream)	1.294	-	-	-	-	-	-	-	-
	[3559]cheese	1.272	-	-	-	-	-	-	-	-
	[883]carbonated flavor	1.128	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
4	[258]water	3.963	[3182]cup noodle	0.165	-	-	[869]carbonated flavor	-0.049	-	-
	[5304]biscuit and cookie	-3.253	-	-	-	-	-	-	-	-
	[4309]instant curry	3.042	-	-	-	-	-	-	-	-
	[3731]carbonated flavor	3.008	-	-	-	-	-	-	-	-
	[4792]family ice (cream)	2.941	-	-	-	-	-	-	-	-
	[5184]chocolat	-2.626	-	-	-	-	-	-	-	-
	[2334]snack	-2.54	-	-	-	-	-	-	-	-
	[4094]fruit juice	2.431	-	-	-	-	-	-	-	-
	[6122]chocolat	2.328	-	-	-	-	-	-	-	-
	[6832]cola	2.314	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
5	[7897]instant noodle	2.174	[5046]dessert	0.036	[4710]biscuit and cookie	0.451	[5255]chocolat	-0.328	[2227]rice cracker	-0.108
	[6815]black tea drink	-1.998	[4983]health food	-0.015	[3419]snack	0.276	[5916]family ice	-0.047	-	-
	[3020]dressing	1.781	[2297]snack	0.000	166]Principle of Chinese foo	-0.187	[3015]mayonnaise	-0.020	-	-
	[4297]cooked curry	1.584	-	-	-	-	[6451]100% fruit juice	-0.006	-	-
	[6735]miso	1.538	-	-	-	-	-	-	-	-
	[6862]	1.528	-	-	-	-	-	-	-	-
	[7717]energy drink	1.528	-	-	-	-	-	-	-	-
	[2486]candy	-1.443	-	-	-	-	-	-	-	-
	[505]candy	-1.413	-	-	-	-	-	-	-	-
	[6211]candy	1.410	-	-	-	-	-	-	-	-

Table 4(2): Feature Selection for Topics

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
6	[4124]coffee drink	2.021	[7430]dessert	-0.290	-	-	[3783]carbonated flavo	0.066	-	-
	[7196]soup	-1.678	[1560]rice cracker	0.174	-	-	-	-	-	-
	[3019]dressing	-1.673	[3184]cup noodle	-0.086	-	-	-	-	-	-
	[6540]100% fruit juice	-1.613	[7148]family ice (cream)	0.000	-	-	-	-	-	-
	[713]yogurt	1.580	-	-	-	-	-	-	-	-
	[5884]vegetable juice	1.517	-	-	-	-	-	-	-	-
	[2493]candy	1.445	-	-	-	-	-	-	-	-
	[7158]family ice (cream)	1.364	-	-	-	-	-	-	-	-
	[6143]chocolat	1.327	-	-	-	-	-	-	-	-
	[1125]Principle of Chinese	1.265	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
7	[1345]chocolat	2.085	[3631]dessert	0.013	[1634]cup noodle	1.035	[3994]dessert	-0.518	-	-
	[2064]regular coffee	1.754	[4686]chocolat	0.003	[1878]regular coffee	-0.532	[6777]suger	-0.117	-	-
	[7166]yogurt	1.668		-	[3196]cup noodle	0.333	[6874]vegetable juice	-0.036	-	-
	[1264]instant curry	1.667	-	-	[1142]Chinese food	0.272	[6781]bread crumb	-0.007	-	-
	[6269]soup	1.661	-	-	[3345]coffee drink	0.171	-	-	-	-
	[4097]rice cracker	-1.464	-	-	[1273]cooked curry	-0.105	-	-	-	-
	[6518]Janpanese tea and barley tea drink	1.429	-	-	-	-	-	-	-	-
	[6103]chocolat	-1.422	-	-	-	-	-	-	-	-
	[6428]yogurt	1.393	-	-	-	-	-	-	-	-
	[6122]chocolat	1.362	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
8	[3589]milk	-0.489	-	-	-	-	-	-	-	-
	[6620]coffee drink	0.424	-	-	-	-	-	-	-	-
	[6772]suger	0.232	-	-	-	-	-	-	-	-
	[3493]cup noodle	-0.231	-	-	-	-	-	-	-	-
	[2493]candy	0.225	-	-	-	-	-	-	-	-
	[5579]suger	-0.207	-	-	-	-	-	-	-	-
	[3101]chocolat	0.205	-	-	-	-	-	-	-	-
	[359]personal ice (cream)	0.205	-	-	-	-	-	-	-	-
	[89]coffee drink	0.170	-	-	-	-	-	-	-	-
	[4786]milk	0.164	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
9	[1086]Principle of Chinese	1.415	-	-	[6891]rice cracker	-0.965	[6316]energy drink	0.470	-	-
	[4263]Principle of Chinese	1.291	-	-	[6053]chocolat	-0.646	[6317]energy drink	-0.345	-	-
	[1149]salad oil	1.193	-	-	[2598]chocolat	-0.532	[2167]vegetable juice	-0.313	-	-
	[1230]instant curry	1.185	-	-	[1680]cup noodle	-0.442	[2090]tomato juice	0.143	-	-
	[6734]instant soup	-1.072	-	-	[2534]chocolat	-0.334	[2095]tomato juice	0.088	-	-
	[4839]cheese	1.058	-	-	-	-	[2156]vegetable juice	-0.054	-	-
	[5988]dressing	0.981	-	-	-	-	[2959]dressing	0.010	-	-
	[274]candy	0.978	-	-	-	-	-	-	-	-
	[743]vegetable juice	0.967	-	-	-	-	-	-	-	-
	[1578]health food	-0.947	-	-	-	-	-	-	-	-

Topic	Sales	Coef	Price	Coef	Prom1	Coef	Prom2	Coef	Prom3	Coef
10	[4297]cooked curry	2.070	[476]dessert	-0.379	[4643]chocolat	0.121	[3895]cup noodle	0.278	-	-
	[494]sauce	1.979	[271]chocolat	0.332	[4407]barley tea	-0.012	[3394]dessert	0.098	-	-
	[3287]fruit juice	1.578	[2659]chewing gum	0.210	-	-	[5483]soup	-0.073	-	-
	[7176]yogurt	-1.340	[5007]personal ice (cream	-0.168	-	-	[3393]dessert	0.012	-	-
	[3068]candy	1.283	[4407]barley tea	0.089	-	-	-	-	-	-
	[3847]cup noodle	1.181	[6584]100% fruit juice	0.078	-	-	-	-	-	-
	[3010]dressing	-1.159	[1672]cup noodle	0.073	-	-	-	-	-	-
	[6936]candy	1.014	[1621]snack	-0.069	-	-	-	-	-	-
	[7260]dessert	-0.960	[5021]yogurt	-0.055	-	-	-	-	-	-
	[6831]carbonated flavor	0.936	[6586]vegetable juice	0.053	-	-	-	-	-	-

Figure 1: Overview of Model

$$Y = FX + \varepsilon$$

non-zero data

Zero data

$N \times P_y$; P_y > N; Many Zero

$N \times P_x$; P_x > N; Many Zero

T_{1y}

T_{ky}

Original Data

T_{1x}

T_{kx}

Topic Model

...

...

$N \times T_{1y}$; T_{1y}> N; Less Zero

a_1

b_1

a_k

b_k

Bayes Factor

$y_1 = V_1 a_1 + \eta_{y1}$

$x_1 = U_1 b_1 + \eta_{x1}$

$y_k = V_k a_k + \eta_{yk}$

$x_k = U_k b_{1k} + \eta_{xk}$

$$a_1 = H_1 b_1 + \eta_{h1}$$

$$a_k = H_k b_k + \eta_{hk}$$

$$\widehat{Y} = \sum_{k=1}^{10} \widehat{y_k}$$

$E(F)$

$P_y \times P_x$ coefficients matrix

26
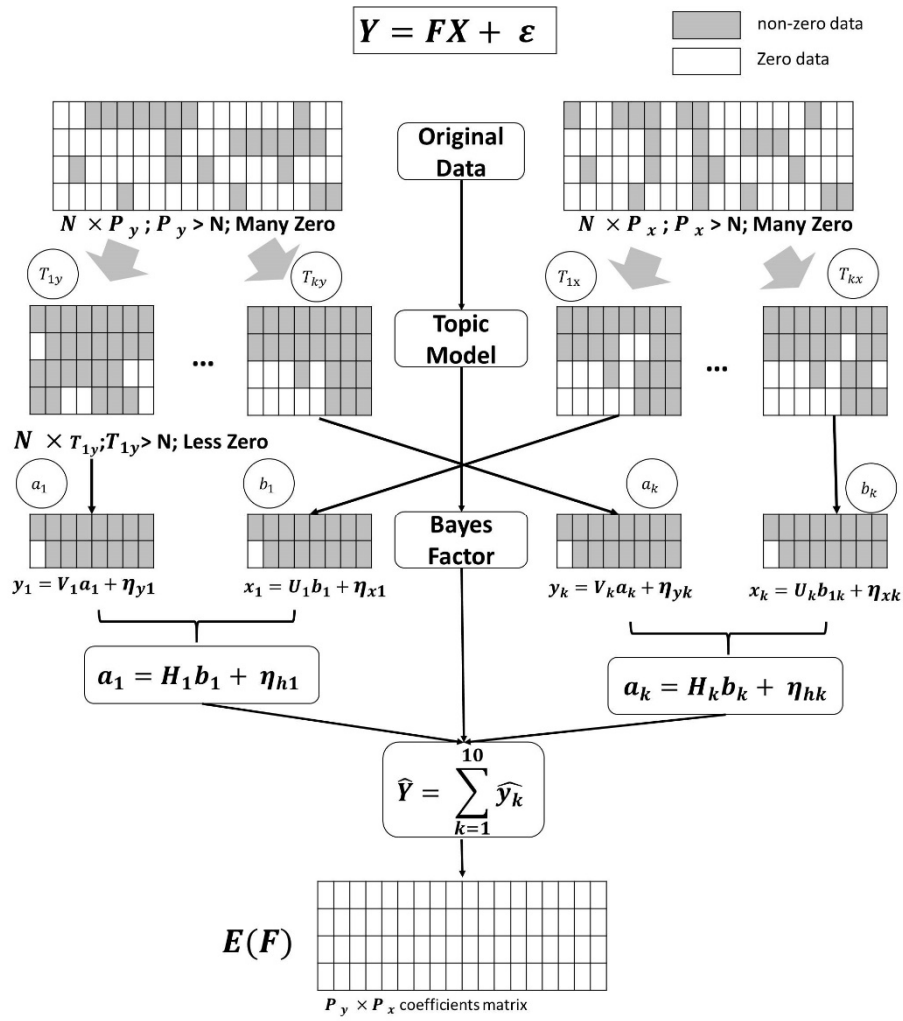
Figure 1: Sales of a Milk Product (JAN: 4902705065161)
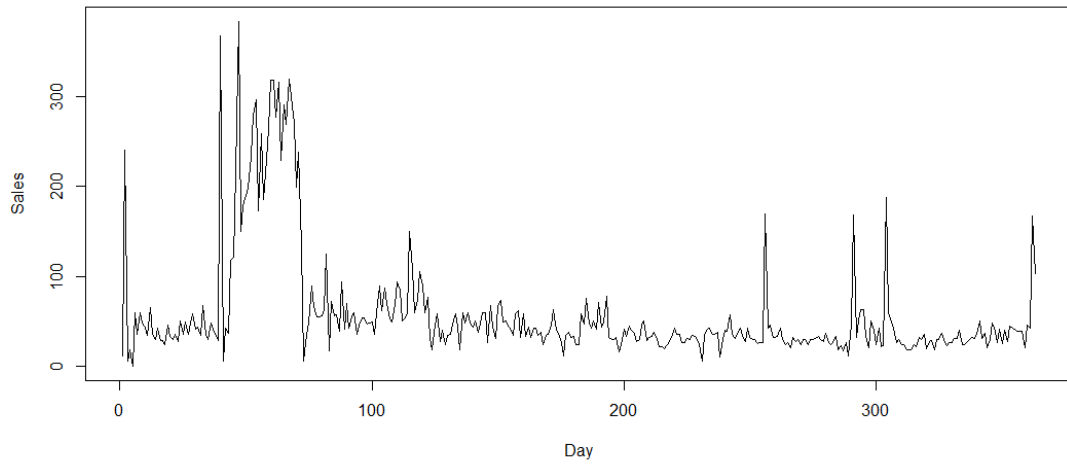
Figure 2: Averaged Topic Distribution for a Product
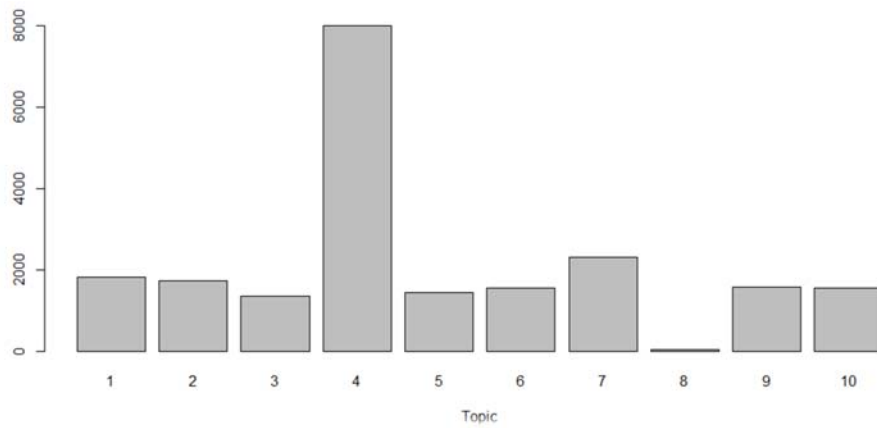
Figure 3: Topic Sales Decomposition
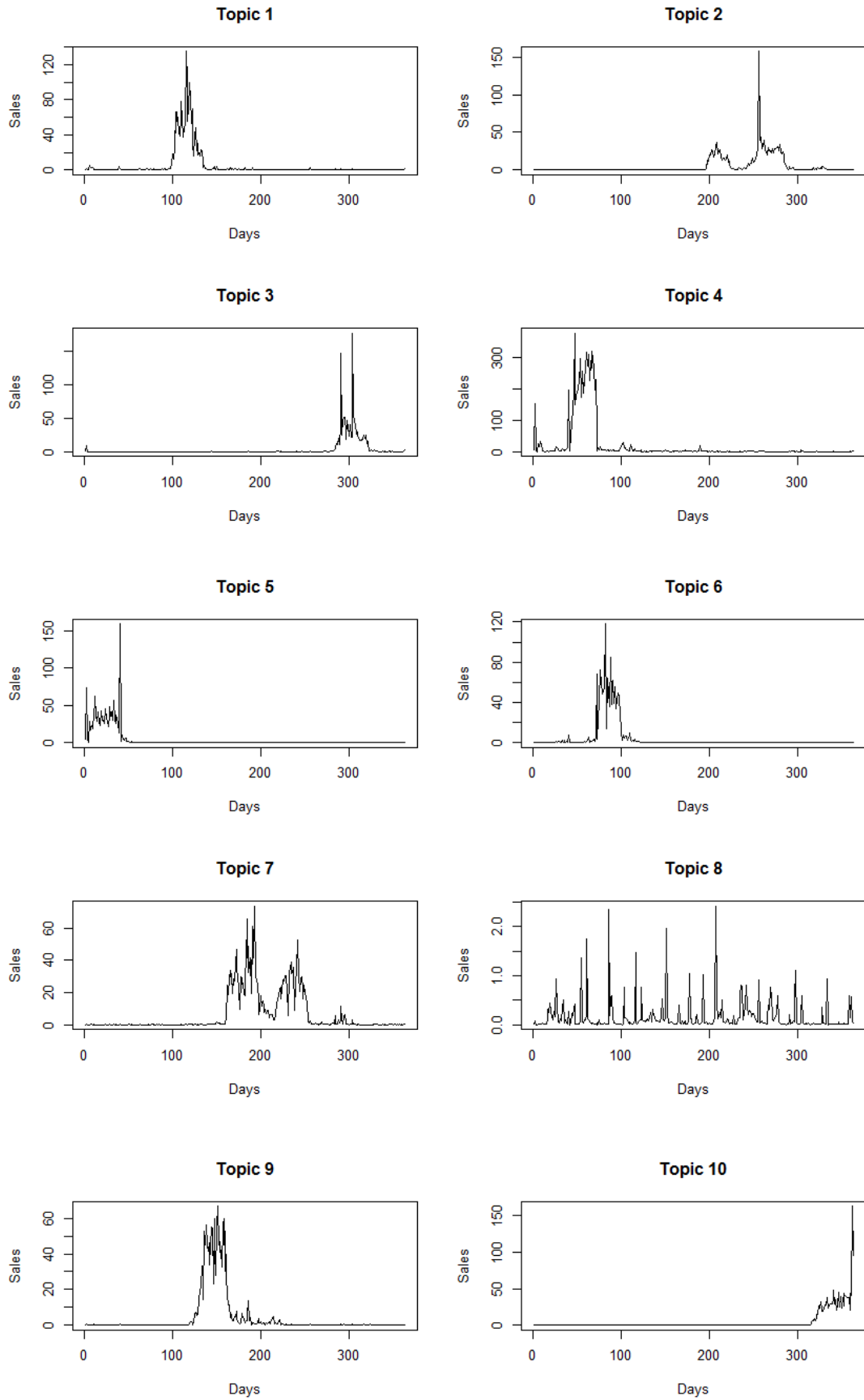
Figure 4: Model Fit and Forecasting

(a) Topic-Hierarchical Factor Regression
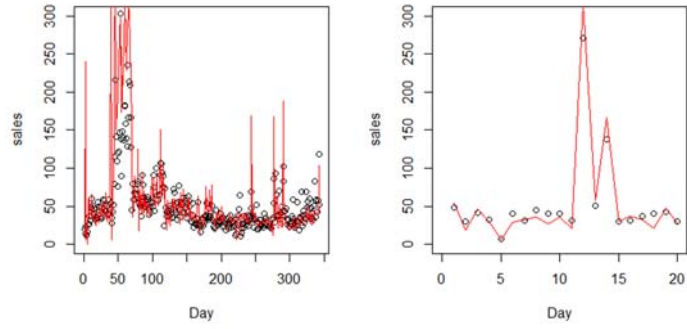
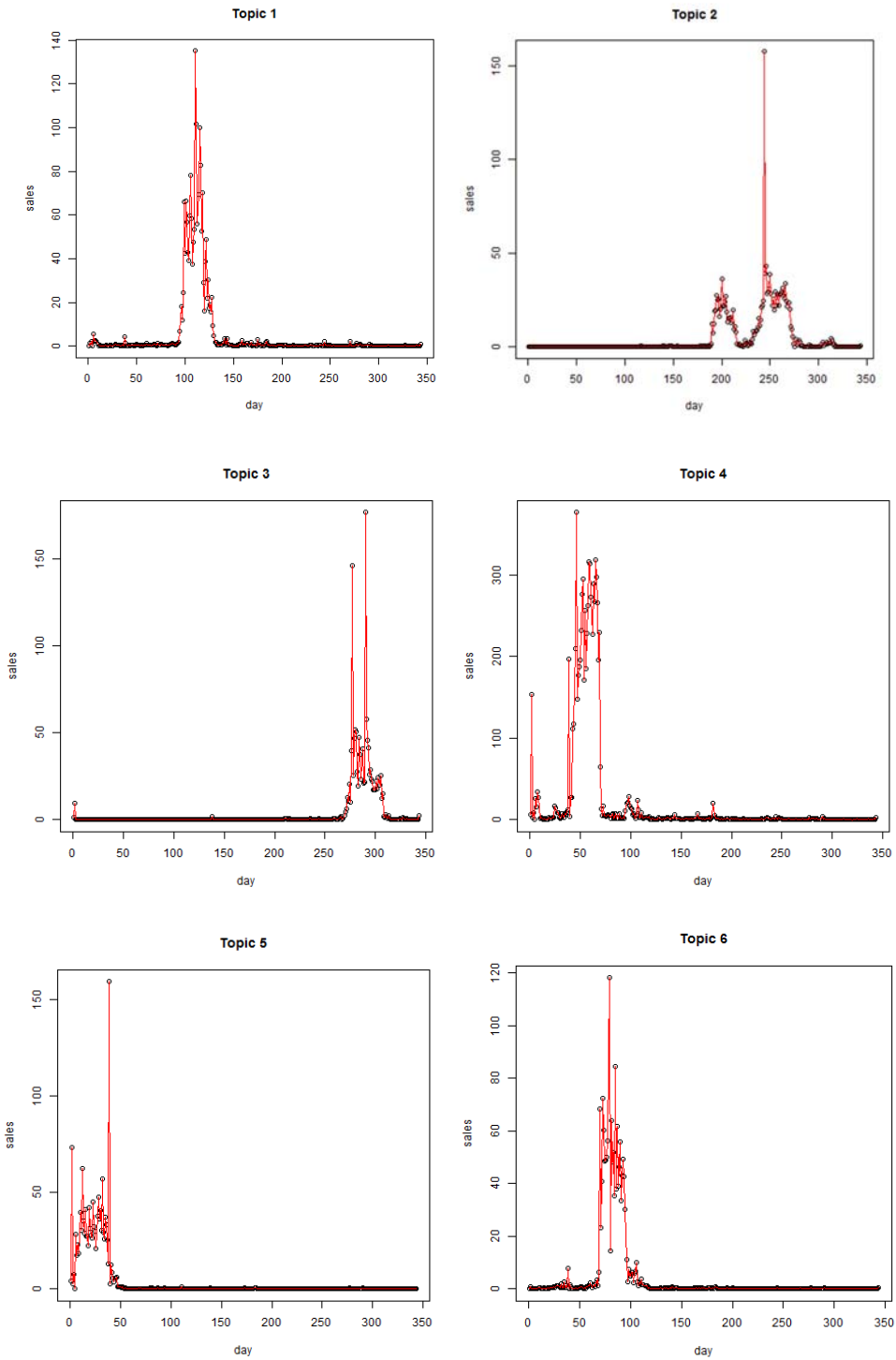### (b) Efficient Topic-Hierarchical Factor Regression
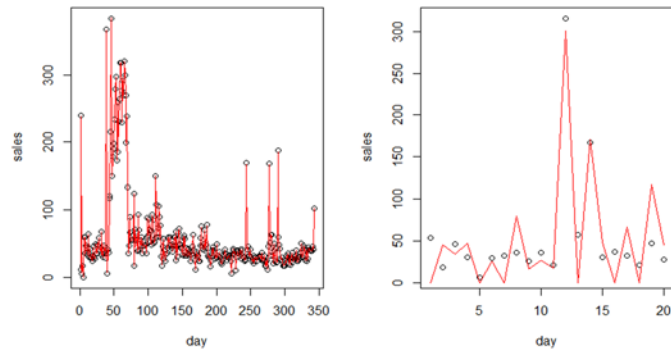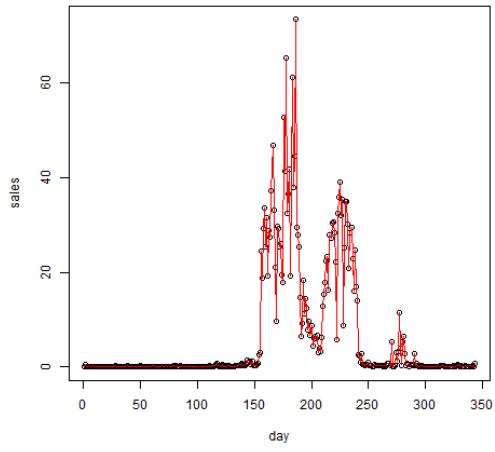

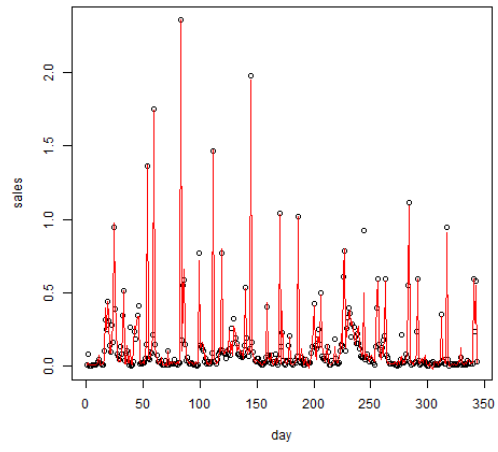### (c) Topic Regression
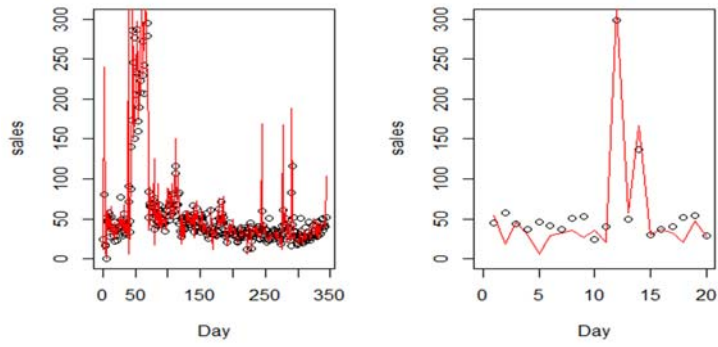
(d) Category Regression
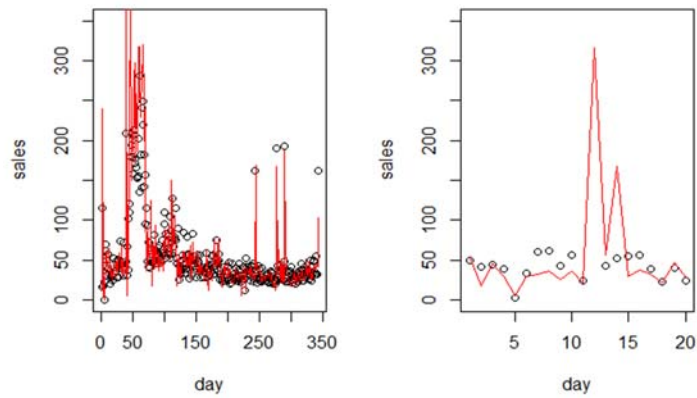
Figure5: Estimated Sales in Topics

Topic 7

Topic 8

Topic 9

Topic 10