ARTICLE TEMPLATE

Sparse Assortment Personalization in High Dimensions

Jingyu Shao[a] , Ruipeng Dong[a] and Zemin Zheng[a]

[a]International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui, 230026, China

ARTICLE HISTORY
Compiled September 20, 2021

ABSTRACT
The data-driven conditional multinomial logit choice model with customer features has a good performance in assortment personalization problem when a low-rank structure of parameter matrix is considered. Despite the recent theoretical and algorithmic advances, parameter estimation in the choice model is still a challenging task especially when the predictors are more than the observations. Motivated by this concern, we suggest a penalized likelihood approach based on the feature matrix, to recover the sparse structure from populations and products toward the assortment. Our proposed method considers low-rank and sparsity structure simutaneously, which can further reduce model complexity and improve estimation and prediction accuracy. New algorithm sparse factoral gradient descent (SFGD) is proposed to estimate the parameter matrix, which enjoys high interpretability and efficient performance in computing. As a first-order method, SFGD works well in high dimensional scenario due to the absence of Hessian matrix. Simulation studies show that SFGD algorithm outperforms the state-of-art methods in estimation, sparsity recovery as well as the average regret.

KEYWORDS
Assortment personalization; Sparsity; Penalized likelihood; Factoral gradient descent; Low rank matrix approximation

1. Introduction

As an important part of revenue management, assortment planning has a wide range of applications in the fields of retailing, advertising settings and e-commerce. The personalization techniques are used to optimize the selection of products or services for certain customers. A key factor to optimize the assortment successfully is the ability to understand and predict the demand or preferences of customers. In the scenario of many online application, the customer-specific data will be available for companies. The feature information of customer data is of great significance to model the relationship between features and purchase decisions. In [1,2], transactional data are used to estimate customer preferences. And they rely on the discrete context of each customer's type, that is, certain types of customers are discovered before the estimation. And a practical algorithm for personalization under inventory constraint was provided in [3]. The full features data of customers are considered in [4], where the covariate information was learned from the data.

CONTACT Zemin Zheng. Email: zhengzm@ustc.edu.cn

To better understand customers preference and demand in practice, logit models are commonly used by far. It is the advantage of interpretability and simplicity that make the logit model a popular choice. Such framework is widely used in targeted advertising [5], pricing [6] and assortment personalization [1,3,4]. The data-driven logit model framework, as a special kind of generalized linear model, use the information of customers features to estimate the coefficients, based on which the assortment optimization will be finished. In the big data application, it's challenging to learn and infer the dependence structures since the responses and predictors in such GLM framework may be related through a few latent pathways or a subset of predictors. Futhermore, with the exponential growth in data volume, the curse of dimensionality and massive amounts of data make the estimation and prediction harder to process. To successfully recover the sparse structure of predictor associated with response, regularization methods such as lasso [7], group lasso [8] and group lasso for logistic regression [9].

In the multi-responses scenario, the data-driven multinomial logit model tackles the associations between the predictors and responses via a sparse and low-rank representation of the coefficient matrix. The sparse reduced-rank regression has been extensively researched in the literature, such methods maintain the iterpretability of the estimated matrix by eliminating irrelevant features and the low-rank structure helps to reduce the number of free parameters of model [10–13]. Sparse reduced-rank regression has amounts of application such as social network community discovery [14], subspace clustering [15], motion segmentation [16]. To the best of our konwledge, in the assortment personalization problem, the simultaneous sparse and low-rank structure in the coefficient matrix of the multinomial logit model are rarely considered in the literatures. To meet this in our multinomial logit model, we choose the framework of penalized likelihood.

To derive a sparse reduced-rank approximation of parameter matrix, it's popular to choose L_1 and nuclear norm regularizers. There are several methods for solving the penalized likelihood problem since the convex relaxations to sparsity and low-rankness of a matrix. The resulting preoblem is convex and then can be solved by alternating direction method of multipliers (ADMM) [17], see [14]. Other method such as sequential co-sparse unit-rank method [18], sparse eigenvalue decomposition [19]. All the above sparse reduced-rank approaches enjoy the desirable theoretical properties. However, they can not be used directly in the penalized likelihood framework. For the GLM problem, the factored gradient descent method [20] is commonly used in the problem that can be posed as matrix factorization, see [21] for the precise convergence rate guarantees for general convex function. Such first-order method works in an alternative way and it does not require SVD of parameter matrix at each step, which makes high efficiency computation a possibility in solving the penalized likelihood problem.

The main contributions of this article are threefold. First of all, we give the framework of the assortment personalization problem, which maximizes the expected revenue over feasible assortment. We introduce the customer features that are related to the utility model, and present our data-driven conditionally multinomial logit choice model. For the sparsity of parameter matrix, we use a group lasso type penalty to derive the rowwise sparsity, which is in same the features selection for customers. Through this we make the estimation in high-dimensional features scenario available. Second, To solve the penalized maximum likelihood problem, we propose a first-order sparse factored gradient descent (SFGD) approach, in which both the sparsity and low-rank structures are considered. Due to the low-rank of parameter matrix, we are able to use the SVD to reduce the amount of parameters. We illustrate the details of the thresholding rule in SFGD and how it proceeds in the alternative updating of two matrix derived from the decomposition. Moreover we show the local convergence of SFGD and also present the structure-aware dynamic assortment personalization procedure based on the SFGD method. Third, our simulation which contains the high-dimensional settings show that SFGD can consistently estimate the parameter matrix and accurately recover the support of features. The average regret of different structure settings are compared with the growth of time horizon, and the SFGD method with sparse reduced-rank structure considered outperforms the sparsity structure-ignorant methods.

2. Model specification

In this section we present our modeling framework for data-driven assortment personalization problems, where customer features is considered. Throughout this paper, we use bold letter to denote matrix and vector. In this paper, $\boldsymbol{z}_i$ is the column vector of ith row of $\boldsymbol{Z}$, and z_{ij} is the j element of vector $\boldsymbol{z}_i$ without special instructions. For any matrix $\boldsymbol{Z} = (z_{ij})$, denote by $||\boldsymbol{Z}||_F = \sqrt{\sum_{i,j} z_{ij}^2}$, $||\boldsymbol{Z}||_{2,1} = \sum_i ||\boldsymbol{z}_i||_2$ and $||\boldsymbol{Z}||_\infty = max_{ij}|z_{ij}|$ the Frobenius norm, rows $l_{2,1}$-norm and entrywise l_∞-norm. And $\sigma_1(\boldsymbol{Z})$ means the largest sigular value of $\boldsymbol{Z}$.

In the assortment personalization problem, the retailer records the observed transctional data in the past, in which contain the customers features, items (products) chosen by customers and the assortment arrangement provided by retailer. For a time horizon T, the decision maker had observed the customers data in the past time $t = 1, ..., T$. In time t, the decision maker got the customer data $\boldsymbol{x}_t$ with p features that include individual information, the assortment $S_t \subset \{1, ..., q\}$ and the items $j_t \in \{1, ..., q\}$ which was chosen by the t customer.

2.1. Data-driven Conditionally Multinomial Logit Choice Model

In the data-driven assortment problem, we assume that the customer data matrix $\boldsymbol{X}$ of size $T \times p$ obtained directly from the past, also known as feature vectors. We assume that customers choose among the products according to some conditional probability $\mathbb{P}_{\boldsymbol{\Theta}}(j|S)$, when the assortment was shown to the customer. Here $\boldsymbol{\Theta}$ is the parameter matrix which play an important role in conditional multinomial logit choice model, and the choice of $\boldsymbol{\Theta}$ will be presented later. For each item $j \in \{1, ..., q\}$, let r_j be its associated revenue. Here $r_0 = 0$ in the revenue for no-purchase option. Then the decision maker maximizes the expected revenue

$$f(S) = \sum_{j \in S} r_j \mathbb{P}_{\boldsymbol{\Theta}}(j|S) \tag{1}$$

over feasible assortment $S \subset \{1, ..., q\}$. The assortment personalization problem aims to find the assortment that maximize the expected revenue

$$\hat{S} = \underset{S \subseteq \{1,...,q\}}{\text{argmax}}\ f(S).$$

In order to have a clear view of $\boldsymbol{\Theta}$, we first introduce the utility of items. A popular way to model the customer choice probability is to utilize the random utility model

[22]. We assume that a customer with feature vector $\boldsymbol{x} \in \mathbb{R}^p$ has utility

$$U_j^x = V_j^x + \epsilon_j \tag{2}$$

for each product j, where V_j^x can be interpreted as the mean utility of product j for this customer and ϵ_j is a standard Gumbel random variable with mean zero. When a decision maker offers the assortment $S \subset \{1, ..., q\}$ to a customer with features $\boldsymbol{x}$, the customer will choose the product in S with the highest U_j^x. The utility V_0^x of no-purchase option it to be zero. Here we assume that the mean utility is given by a linear model $V_j^x = \left\langle \boldsymbol{x}, \boldsymbol{\theta}_j^* \right\rangle$, where $\boldsymbol{\theta}_j^* \in \mathbb{R}^p$ for $1 \leq j \leq q$. And hence we get the mean utility matrix on all items $\boldsymbol{V}^x = \boldsymbol{X}\boldsymbol{\Theta}^*$, where the underlying parameter matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_q) \in \mathbb{R}^{p \times q}$.

The data-driven conditionally multinomial logit choice model is over time $t = 1, ..., T$ and items $j = 1, ..., q$. Here we introduce two random variables: customer I and item (choice) J. By a well-known result from discrete choice theory [23], given the assortment $S \subset \{1, ..., q\}$, we derive a personalized case of choice probability

$$\mathbb{P}_{\boldsymbol{\Theta}^*}(J = j|S) = \frac{e^{V_j^x}}{1 + \sum_{j' \in S} e^{V_{j'}^x}} = \frac{\exp\{\left\langle \boldsymbol{x}, \boldsymbol{\theta}_j^* \right\rangle\}}{1 + \sum_{j' \in S} \exp\{\left\langle \boldsymbol{x}, \boldsymbol{\theta}_{j'}^* \right\rangle\}}. \tag{3}$$

We choose the linear model of $\boldsymbol{x}$ and $\boldsymbol{\theta}_j^*$ to represent V_j^x, then the choice J has the conditional distribution

$$\mathbb{P}_{\boldsymbol{\Theta}^*}(J = j|I = \boldsymbol{x}_t; S) = \frac{1}{1 + \sum_{j' \in S} \exp(\boldsymbol{x_t}^T \boldsymbol{\theta}_{j'}^*)} \times \begin{cases} 1 & j = 0 \\ 0 & j \neq 0, j \notin S \\ \exp(\boldsymbol{x_t}^T \boldsymbol{\theta}_j^*) & j \neq 0, j \in S \end{cases} \tag{4}$$

where $J = 0$ means that no product was purchased in assortment S. The no-purchase option is common in the choice model. In our data-driven framework, the decision maker is able to observe customer features $\boldsymbol{x}_t \in \mathbb{X}, t = 1, ..., T$, where $\mathbb{X} \subset \mathbb{R}^p$ is a space of possible contexts. We also assume that $\boldsymbol{x}_t$ is scaled to satisfy $||\boldsymbol{x}_t||_\infty \leq 1$ for $t = 1, ..., T$.

2.2. Penalized Maximum Likelihood Approach

We suppose that we have T observations $(\boldsymbol{x_t}, j_t, S_t)$ for $t = 1, ..., T$ where S_t comes from the set of subsets of $\{1, ..., q\}$ of size K, and j_t are *iid* according to model (4). Based on the specific form of $\mathbb{P}_{\boldsymbol{\Theta}^*}(J = j|I = \boldsymbol{x}_t; S)$ in (4), we now define the loss function constructed from negative log-likelihood:

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{\Theta}) = \frac{1}{T} \sum_{t=1}^{T} \log((1 + \sum_{j \in S_t} e^{\boldsymbol{x}_t^T \boldsymbol{\theta}_j})(I_{(j_t=0)} + I_{(j_t \neq 0)} e^{\boldsymbol{x}_t^T \boldsymbol{\theta}_{j_t}})^{-1}) \tag{5}$$

Similar to classic methods, we often assume that the underlying parameter matrix $\boldsymbol{\Theta}^*$ has a certain special structure such as low-rank structure of $\boldsymbol{\Theta}^*$ and the sparsity of $\boldsymbol{\Theta}^*$ [12]. In the customer choice model, it's reasonable to assume that for a customer

$\boldsymbol{x}$, only few features have significant impact on the utility of choosing different items. Since the sparsity depend on the items, we introduce the rowwise sparsity of $\boldsymbol{\Theta}^*$.

In order to recover the sparse structure of the parameter matrix, the regularization method could be helpful for varible selection as well as sparsity recovery. In the sparse reduced-rank leaning, we tend to recover the sparsity and low-rank structure simultaneously. Choose from a large amount of features which is also a procedure of variable selection of our generalized multi-response regression problem. When large numbers of predictor variables (i.e. features) are available, some of them may not helpful for estimation as well as prediction. Therefore, it is of great significance to perform feature selection through shrinkage method. In our problem, setting the entire row of $\boldsymbol{\Theta}$ as zero corresponds to excluding a feature of customer data. Inspired by the regularization method in regression, we choose a grouped lasso type [8,12] penalty to aviod overfitting and improve interpretability. Then we have the form of our problem as

$$\begin{aligned}&\text{minimize } \mathcal{Q}(\boldsymbol{X};\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta}) + \lambda||\boldsymbol{\Theta}||_{2,1}\\ &\text{s.t. } \text{rank}(\boldsymbol{\Theta}) \leq \tilde{r}\end{aligned} \tag{6}$$

If $\boldsymbol{\Theta}^*$ has rank r, we may factor $\boldsymbol{\Theta}^*$ to find vectors $\boldsymbol{u}_i = (u_{i1}, ..., u_{ir})^T$ and $\boldsymbol{v}_j = (v_{j1}, ..., v_{jr})^T$ for $i = 1, ..., p$, $j = 1, ..., q$, such that θ^*_{ij} is approximately equal to $\sum_{l=1}^{r} u_{il}v_{jl}$. Denote that $\boldsymbol{U} = (u_{il})_{p\times r}$, $\boldsymbol{V} = (v_{jl})_{q\times r}$, then the right factors $\boldsymbol{V}$ can be thought of latent item weights, and the left factors $\boldsymbol{U}$ as latent features [1]. We now derive an appealing sparse SVD representation of $\boldsymbol{X\Theta}$ as

$$\begin{aligned}&\frac{1}{\sqrt{T}}\boldsymbol{X\Theta} = \left(\frac{1}{\sqrt{T}}\boldsymbol{XU}_0\right)\boldsymbol{D}_0\boldsymbol{V}_0^T = \frac{1}{\sqrt{T}}\boldsymbol{XUV}^T\\ &\text{s.t. } \left(\frac{1}{\sqrt{T}}\boldsymbol{XU}_0\right)^T\left(\frac{1}{\sqrt{T}}\boldsymbol{XU}_0\right) = \boldsymbol{V}_0^T\boldsymbol{V}_0 = \boldsymbol{I}_r\end{aligned} \tag{7}$$

where $\boldsymbol{U} = \boldsymbol{U}_0\boldsymbol{D}_0 \in \mathbb{R}^{p\times r}$, $\boldsymbol{V} = \boldsymbol{V}_0 \in \mathbb{R}^{q\times r}$, $\boldsymbol{D}_0 = \text{diag}\{d_1^0, ..., d_r^0\}$, and $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix. This encourages us to use $\boldsymbol{U}$ and $\boldsymbol{V}$ to factorize the matrix $\boldsymbol{\Theta}$, that is, $\boldsymbol{\Theta} = \boldsymbol{UV}^T$. Now we derive the factored form of objective problem

$$\begin{aligned}&\text{minimize } \mathcal{Q}(\boldsymbol{X};\boldsymbol{UV}^T) = \mathcal{L}(\boldsymbol{X};\boldsymbol{UV}^T) + \lambda||\boldsymbol{U}||_{2,1}\\ &\text{s.t. } \boldsymbol{U} \in \mathbb{R}^{p\times r}, \boldsymbol{V} \in \mathbb{R}^{q\times r}, r \leq \tilde{r}, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I},\end{aligned} \tag{8}$$

where the constraint $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ is from (7) for the purposes of identifiability, and tuning parameters λ which is choosen by a information criterion will be discussed later. By the constraint $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and the definition of row l_1 norm, we have $||\boldsymbol{\Theta}||_{2,1} = ||\boldsymbol{U}||_{2,1} = \sum_{i=1}^{p} ||\boldsymbol{u}_i||_2$, where $\boldsymbol{u}_i$ is the ith row vector of $\boldsymbol{U}$. Therefore, we can shrinkage $||\boldsymbol{\theta}_i||_2$ into zero by setting ith row of $\boldsymbol{U}$ as zeros, and then derive the rowwise sparsity on the $\boldsymbol{U}$ and $\boldsymbol{\Theta}$ accordingly. In our generalized multi-response regression problem, all the items have probabilities to be choosen and that motivate us to introduce rowwise sparsity instead of columnwise sparsity.

We define our estimator $\hat{\boldsymbol{\Theta}}$ for $\boldsymbol{\Theta}^*$ as the solution of the maximum likelihood problem with low rank assumption rank$(\boldsymbol{\Theta}^*) \ll \min\{p, q\}$. Since problem (8) is convex, we can apply a variety of convex methods on it. In the next part, we have a nice try using a first-order algorithm on the non-convex and factored form.

3. Algorithm

With the convexity of $\mathcal{Q}(\boldsymbol{X};\boldsymbol{\Theta})$, many fast optimization approaches such as alternating direction method of multipliers, accelerated projected gradient descent and factored gradient descent, could perform well. The commonly used method for estimating the parameter matrix with low-rank structure is the factored gradient descent method [20]. In this section, we will introduce a data-driven Sparse Factored Gradient Descent (SFGD) algorithm to appromix $\boldsymbol{\Theta}^*$ with low-rank and sparse structure. SFGD is an interactive method, in which the rowwise sparsity of $\boldsymbol{U}$ is considered and the updates of $\boldsymbol{U}$ and $\boldsymbol{V}$ are overlapped. In the update of $\boldsymbol{U}$, we use a subgradient approach which cooperates with gradient descent method.

3.1. Sparse Factored Gradient Descent Method

In the scenario of high dimensions, the computation of Hessian matrix could be difficult or even infeasible. In this part, we introduce a first-order algorithm for computing $\hat{\boldsymbol{\Theta}}$, which works on the factored form of low-rank as well as sparse constraint likelihood optimization problem (8). We first consider the problem without regularization

$$\begin{aligned} &\text{minimize } \mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) \\ &\text{s.t. } \boldsymbol{U}\in\mathbb{R}^{p\times r}, \boldsymbol{V}\in\mathbb{R}^{q\times r}, r\le\tilde{r}. \end{aligned} \tag{9}$$

It's clear that the algorithm reduces computational work since this model has only $r\times(p+q)$ optimization parameters rather than $p\times q$. Moreover, our SFGD algrithm works in an alternative way, that is, we optimize the factors $\boldsymbol{U}$ and $\boldsymbol{V}$ of the parameter matrix $\boldsymbol{\Theta}=\boldsymbol{U}\boldsymbol{V}^T$ rather than producing SVD at each step.

From the convexity of $\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$, it is feasible to use the gradient descent method. Inspired by factored form of our problem, we introduce the factored gradient descent procedure, which is a data-driven non-convex method and a fundamental part of SFGD. The SFGD algorithm first solves the unconstrained problem (9) by the alternate updating rule

$$\begin{aligned} \boldsymbol{U}' &= \boldsymbol{U}-\eta\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) \\ \boldsymbol{V}' &= \boldsymbol{V}-\eta\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T), \end{aligned} \tag{10}$$

which is closely connected to the alternating convex search (ACS) method as in [18,24]. The main difference our SFGD is that $\boldsymbol{U}$ and $\boldsymbol{V}$ are overlapping to each other with rank $r\ge 1$ rather than the unit-rank problem. We start the line search with a step size of $\eta=1$, after which the adaptive step size is repeatedly decreased by a shinkage factor β_{dec} until the objective decreases.

It is easy to compute the gradients of the objective of (9). According to the chain rule of differentiable function, we have

$$\begin{aligned} \nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \boldsymbol{X}^T\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} \\ \nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T)^T\boldsymbol{X}\boldsymbol{U}, \end{aligned}$$

here we do not need explicitly form $\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T)$ to compute the gradients, we have the following form of gradients

$$
\begin{aligned}
\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}\boldsymbol{x}_t\boldsymbol{v}_j^T}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}} - \boldsymbol{x}_t\boldsymbol{v}_{j_t}^T\right)\\
\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}\boldsymbol{e}_j\boldsymbol{x}_t^T}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}} - \boldsymbol{e}_{j_t}\boldsymbol{x}_t^T\right)\boldsymbol{U},
\end{aligned}
$$

which clarifies the gradient descent direction, see appendix for more details.

Now we introduce the rowwise sparsity of $\boldsymbol{\Theta}$. To solve this problem, in the (m)th step, we use subgradient method to screen the rows of $\boldsymbol{U}^{(m)}$, which aims to find sparsity when $||\boldsymbol{u}_i^{(m)}||_2 = 0$. Denote by x_{t_j} the jth element of $\boldsymbol{x}_t$. For any $i = 1, ..., p$, we use the subgradient method on (8) with respect to $\boldsymbol{u}_i^{(m)}$, and let the subgradient be zero, which leads to

$$
\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t\boldsymbol{U}^{(m)}\boldsymbol{v}_j}x_{t_j}\boldsymbol{v}_j}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t\boldsymbol{U}^{(m)}\boldsymbol{v}_j}} - x_{t_j}\boldsymbol{v}_{j_t}\right) + \lambda\frac{\boldsymbol{u}_i^{(m)}}{||\boldsymbol{u}_i^{(m)}||_2} = 0. \tag{11}
$$

Let $\boldsymbol{s}_i = \frac{\boldsymbol{u}_i^{(m)}}{||\boldsymbol{u}_i^{(m)}||_2}$ if $||\boldsymbol{u}_i^{(m)}||_2 \neq 0$. And $\boldsymbol{s}_i$ is a r vector satisfying $||\boldsymbol{s}_i||_2 < 1$ if $||\boldsymbol{u}_i^{(m)}||_2 = 0$, then we have

$$
\boldsymbol{s}_j = -\frac{1}{\lambda T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t\boldsymbol{U}^{(m)}\boldsymbol{v}_j}x_{t_j}\boldsymbol{v}_j}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t\boldsymbol{U}^{(m)}\boldsymbol{v}_j}} - x_{t_j}\boldsymbol{v}_{j_t}\right). \tag{12}
$$

We present the details of SFGD as follows

1) For the $(m+1)$th repeat in gradient descent, we screen the rowwise sparsity before finally updating $\boldsymbol{U}^{(m)}$.
2) For $i = 1, ..., p$, denote x_{t_j} for the jth element of $\boldsymbol{x}_t$, then compute s_j. And update the jth row of $\boldsymbol{U}^{(m)}$ by a threshold rule

$$
\boldsymbol{u}_i^{(m+1)} = \frac{1}{||\boldsymbol{s}_i||_2 - 1}\big(||\boldsymbol{s}_i||_2 - 1\big)_+\boldsymbol{u}_i^{(m)},
$$

   where $(z)_+ = \max\{0, z\}$ for all $z \in \mathbb{R}$. And without loss of generality, if $||\boldsymbol{s}_i||_2 - 1 = 0$, then we let $\boldsymbol{u}_i^{(m+1)} = \boldsymbol{u}_i^{(m)}$.
3) After screening for all rows of $\boldsymbol{U}^{(m)}$, update $\boldsymbol{U}^{(m)}$ derive $\boldsymbol{U}^{(m+1)}$ and then enter the next iteration or stop.

Our SFGD method could be initialized by the technique from [21], which only need the gradients of $\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta})$. By the SVD of $-\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{0})$, it entails $-\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{0}) = \tilde{\boldsymbol{U}}\text{diag}(\tilde{\sigma}_1, ..., \tilde{\sigma}_{\min\{p,q\}})\tilde{\boldsymbol{V}}^T$. Denote by $\tilde{\boldsymbol{U}}_{\tilde{r}}$ and $\tilde{\boldsymbol{V}}_{\tilde{r}}$ the first $\tilde{r}$ columns of $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$. Let $\boldsymbol{E}_1$ be the $p \times q$ matrix that has value one in the $(1,1)$ element with others zeros. Then we initialize

$$
\begin{aligned}
\boldsymbol{U}^0 &= \omega^{-1/2}\text{diag}(\sqrt{\tilde{\sigma}_1}, ..., \sqrt{\tilde{\sigma}_{\tilde{r}}})\tilde{\boldsymbol{U}}_{\tilde{r}}\\
\boldsymbol{V}^0 &= \omega^{-1/2}\text{diag}(\sqrt{\tilde{\sigma}_1}, ..., \sqrt{\tilde{\sigma}_{\tilde{r}}})\tilde{\boldsymbol{V}}_{\tilde{r}},
\end{aligned}
$$

where $\omega = ||\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{0}) - (\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{E_1}) + \lambda\boldsymbol{E_1})||_F$. On the other hand, the termination of our algorithm is met when the decrease in the objective function value is smaller than the tolerance τ.

The selection of tunning parameter λ is based on the information criterion which will be disscussed later. Considering the overshooting problem in the line search process, the step size shrinkage factor β works to adjust η to make sure that the local optimal will not be missed. The details of SFGD is presented in Algorithm 3.1.

Algorithm 3.1 Sparse Factored Gradient Descent (SFGD)

Input: Feature-item-assortment data $\{(\boldsymbol{x}_t, j_t, S_t)\}_{t=1}^{T}$; dimensions of $\boldsymbol{\Theta}$: p, q; rank $\tilde{r}$; regularizing coefficient λ; step size shrinkage factor β; and tolerance τ. $\boldsymbol{U} \leftarrow \boldsymbol{U}^0$, $\boldsymbol{V} \leftarrow \boldsymbol{V}^0$, $f^{'} \leftarrow \infty$.

```
repeat
    $\eta \leftarrow 1, f \leftarrow f^{'}, \Delta\boldsymbol{U} \leftarrow -\lambda\boldsymbol{U}, \Delta\boldsymbol{V} \leftarrow -\lambda\boldsymbol{V}$
    for $t = 1, ..., T$ do
        for $j \in S_t$ do
            $w_j \leftarrow e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}, W \leftarrow W + w_j$
        end for
        $\Delta\boldsymbol{U} \leftarrow \Delta\boldsymbol{U} - \frac{1}{N}\big(\boldsymbol{x}_t\boldsymbol{v}_{j_t}^T - \frac{1}{W}\sum_{j\in S_t} w_j\boldsymbol{x}_t\boldsymbol{v}_j^T\big)$
        $\Delta\boldsymbol{V} \leftarrow \Delta\boldsymbol{V} - \frac{1}{N}\big(\boldsymbol{e}_{j_t}\boldsymbol{x}_t^T - \frac{1}{W}\sum_{j\in S_t} w_j\boldsymbol{e}_j\boldsymbol{x}_t^T\big)\boldsymbol{U}$
    end for
    repeat
        $\boldsymbol{U}^{'} \leftarrow \boldsymbol{U} + \eta\Delta\boldsymbol{U}, \boldsymbol{V}^{'} \leftarrow \boldsymbol{V} + \eta\Delta\boldsymbol{V}$
        $f^{'} \leftarrow \mathcal{L}(\boldsymbol{X};\boldsymbol{U}^{'}\boldsymbol{V}^{'T}) + \lambda||\boldsymbol{U}^{'}\boldsymbol{V}^{'T}||_{2,1}$
        $\eta \leftarrow \beta\eta$
    until $f^{'} \leq f$
    for $i = 1, ..., p$ do
        $\boldsymbol{u}_i^{'} = \frac{1}{||\boldsymbol{s}_i||_2 - 1}\big(||\boldsymbol{s}_i||_2 - 1\big)_+\boldsymbol{u}_i^{'}$ with $\boldsymbol{s}_i$ defined in (12)
    end for
    $\boldsymbol{U} \leftarrow \boldsymbol{U}^{'}, \boldsymbol{V} \leftarrow \boldsymbol{V}^{'}$
until $\frac{f - f^{'}}{f^{'}} \leq \tau$
```

Output: $\hat{\boldsymbol{\Theta}} = \boldsymbol{U}\boldsymbol{V}^T$

3.2. Local convergence of SFGD

To illustrate the performance of convergence, we first introduce the definition of M-smooth. Let $g : \mathbb{R}^{p\times q} \rightarrow \mathbb{R}$ be a convex differentiable function. Then g is M-smooth if there exists $M > 0$ such that $||\nabla g(\boldsymbol{\Theta}_1) - \nabla g(\boldsymbol{\Theta}_2)||_F \leq M||\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2||_F$, $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{p\times q}$. And this further implies the following bound:

$$g(\boldsymbol{\Theta}_2) \leq g(\boldsymbol{\Theta}_1) + \langle \nabla g(\boldsymbol{\Theta}_1), \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1 \rangle + \frac{M}{2}||\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1||_F^2.$$

From the construction above, we find it's easy for $\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta})$ and thus for $\mathcal{Q}(\boldsymbol{X};\boldsymbol{\Theta})$ to meet the M-smooth condition at each interation of SFGD. Now we give the convergence performance of SFGD algorithm as follows

Theorem 3.1. *Let $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^T$ and $\boldsymbol{\Theta}' = \boldsymbol{U}'\boldsymbol{V}'^T$ denote the input and output of an interation of SFGD. Then there exist a constant $M > 0$ and $k \in \mathbb{Z}^+ \cup \{0, +\infty\}$, if the step size $\eta \leq \frac{\beta^k}{3M(\sigma_1(\boldsymbol{U}^T\boldsymbol{U})+1)}$, where $0 \leq \beta < 1$ is the shrinkage factor, then*

$$\mathcal{Q}(\boldsymbol{X}; \boldsymbol{\Theta}') \leq \mathcal{Q}(\boldsymbol{X}; \boldsymbol{\Theta}) - \frac{5}{6}\eta\beta^k ||[\nabla\mathcal{L}(\boldsymbol{X}; \boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2. \tag{13}$$

And there exists a local optimum $\tilde{\boldsymbol{\Theta}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^T$ such that $\mathcal{Q}(\boldsymbol{X}; \boldsymbol{\Theta})$ converges to $\mathcal{Q}(\boldsymbol{X}; \tilde{\boldsymbol{\Theta}})$.

3.3. The Structure-Aware Dynamic Assortment Personalization Problem

In the senario of dynamic assortment personalization problem, we learn from the past until the time horizon T firstly by affording the random assortments S_t and recording the observations $(\boldsymbol{x}_t, j_t, S_t)$, which is the procedure of exploration. In the next step, we implement the SFGD algorithm to estimate $\boldsymbol{\Theta}$ with both low-rank and sparse structure. With the growing of T, the in sample prediction of the assortment then can be derived accordingly by maximizing the expected revenue (1). For the given p, q and r, there exists a critical value as function $C(T)$ that depends on T, such as $C(T) = Cr(p+q)\log(T)$ in [1]. When T meets $C(T)$, the problem turns into the exploitation, which is the out of sample prediction on the assortment. Denote by $\mathcal{A}$ the collection of observations, and $C(T)$ is slowly varying with respect to T. Then we present the details of our dynamic assortment personalization problem in Algorithm 3.3.

After the exploration step, it yields the structure-aware estimate $\hat{\boldsymbol{\Theta}}$, and based on $\hat{\boldsymbol{\Theta}}$, we derive the conditional distribution by (4) with respect to the new data $\boldsymbol{x}_t$. Now we can see that our structure-aware dynamic assortment personalization approach serve for every upcoming individual at time t rather than several types of customers as in [1].

4. Simulation studies

In this part, we implement the simulation to show the advantages of our approach. We use the GIC [25] for high dimensional penalized likelihood settings to select the tuning parameter λ by minimizing

$$\mathrm{GIC}_{a_T}(\lambda) = \frac{1}{T}\{\mathcal{L}(\boldsymbol{X}; \hat{\boldsymbol{\Theta}}_\lambda) + a_T|\alpha_\lambda|\},$$

where $\alpha_\lambda \subset \{1, ..., p\}$ is the rowwise support of the estimate $\hat{\boldsymbol{\Theta}}$. Denote by α_0 the true rowwise support of $\hat{\boldsymbol{\Theta}}$, then there exist a λ_0 such that $\alpha_{\lambda_0} = \alpha_0$. And the a_T is a positive sequence depending only on T. We choose a modified BIC type of a_T such that $a_T = C_T\log(T)$ with a diverging C_T sequence [26]. In this article, we use this strategy by letting $C_T = c\log(\log(T + p + q))$ where c is a positive constant. In the following analysis, we select λ by minimize $\mathrm{GIC}_{a_T}(\lambda)$.

Algorithm 3.3 Structure-Aware Dynamic Assortment Personalization

```
Input: $C(T), \lambda$
$\mathcal{A} \leftarrow \emptyset$
for $T = 1, 2, ...$ do
    $t \leftarrow T$
    if $T \leq C(T)$ then
        Exploration:
        choose $S_t$ uniformly at random form $\{1, ..., q\}$ of size $K$,
        observe $(\boldsymbol{x}_t, j_t, S_t)$ and $\mathcal{A} \leftarrow \mathcal{A} \cup (\boldsymbol{x}_t, j_t, S_t)$,
        $$\mathcal{L}(\boldsymbol{X}; \boldsymbol{\Theta}) = \frac{1}{T} \sum_{(\boldsymbol{x}_t, j_t, S_t) \in \mathcal{A}} \log \frac{1 + \sum_{j \in S_t} e^{\boldsymbol{x}_t^T \boldsymbol{\theta}_j}}{I_{(j_t = 0)} + I_{(j_t \neq 0)} e^{\boldsymbol{x}_t^T \boldsymbol{\theta}_{j_t}}},$$
        $\hat{\boldsymbol{\Theta}} \in \left\{ \boldsymbol{\Theta} : \text{argmax } Q(\boldsymbol{X}; \boldsymbol{\Theta}), \text{s.t. rank}(\boldsymbol{\Theta}) \leq \tilde{r} \right\}$
    else
        Exploitation:
        $S_t \in \left\{ S : \underset{S \subseteq \{1, ..., q\}}{\text{argmax}} \sum_{j \in S} r_j \mathbb{P}_{\hat{\boldsymbol{\Theta}}}(j|S) \right\}$
    end if
end for
Output: $S_1, ..., S_T$
```

4.1. Estimation accuracy

We first generate the true $\boldsymbol{\Theta}^*$ as follow. First we generate $p \times q$ matrix $\boldsymbol{\Theta}_0$ from elemental standard normal, take SVD of $\boldsymbol{\Theta}_0$ as $\boldsymbol{\Theta}_0 = \boldsymbol{U}\text{diag}(\sigma_1, \sigma_2, ...)\boldsymbol{V}^T$, reserve the first r sigular values and derive $\boldsymbol{\Theta}_1 = \boldsymbol{U}\text{diag}(\sigma_1, ..., \sigma_r, 0, ..., 0)\boldsymbol{V}^T$. then $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_1/\text{sd}(\text{vec}(\boldsymbol{\Theta}_1))$. Finally, we derive the rowwise sparsity by randomly choose $S \subseteq \{1, ..., p\}$ and $|S| = s$ as the corresponding nonsparse rows of $\boldsymbol{\Theta}_2$ with other rows zeros, which yeilds $\boldsymbol{\Theta}^*$. Let customer data $\boldsymbol{X}$ be drawn from normal distribution $N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$, assortments $S_t, t = 1, ..., T$ be uniformly drawn from $\{1, ..., q\}$ with subset size $K = 10$, then derive j_t according to the conditional distribution as demonstrated in (4). We choose $r = 2, \tilde{r} = 2r, s = 10, \tau = 10^{-10}, \eta = 0.05$ and $c = 5$.

In the methods comparison, we consider the ordinary factored gradient descent (OFGD) method, which solve problem (9) by alternative updating rule (10). The OFGD only recover the low-rank structure of $\boldsymbol{\Theta}$. Moreover, we also introduce the MLE method with structure free of $\boldsymbol{\Theta}$, that is both sparse and low-rank structures of $\boldsymbol{\Theta}$ are ignored and in which the rank of $\boldsymbol{\Theta}$ is choosen as $\min\{p, q\}$.

The error of estimation is measured by root mean squared error (RMSE)

$$\text{RMSE}(\boldsymbol{\Theta}) = \frac{1}{\sqrt{pq}} ||\boldsymbol{\Theta} - \boldsymbol{\Theta}^*||_F.$$

To evaluate the error in utility as declared in (2), we also introduce $Er(\boldsymbol{X}\boldsymbol{\Theta})$ defined by

$$Er(\boldsymbol{X\Theta}) = \frac{1}{\sqrt{Tq}}||\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)||_F.$$

Moreover we choose two indicators: false positive rate $FPR = \frac{FP}{FP+TN}$ and false negtive rate $FNR = \frac{FN}{FN+TP}$ to evaluate the results of sparsity recovery, in which for $\boldsymbol{\theta}_i$, if $\boldsymbol{\theta}_i^* = \mathbf{0}$ but $\hat{\boldsymbol{\theta}}_i \neq \mathbf{0}$, then i goes into the counter of FP, and FN, TN, TP would be calculated by analogy.

In Table 1 we compare the performance of OFGD, SFGD and MLE in 100 replications and we also report the results in high-dimensional setting with $p > T$ in Table 1. As reported in the Table 1, undering the setting $c = 5$ in the tunning of λ. The structure-aware SFGD method outperforms all other methods in the error of both estimation and utility. Moreover, as we expected, OFGD method that only consider the low-rank structure performs better than the structure-ignorant MLE method. And SFGD enjoys the ability in sparse recovery. In the high dimensional setting when $p > T$, SFGD still maintain the performance of estimation accuracy as well as sparse recovery.

Table 1. Results in methods OFGD, SFGD, and MLE with different p, q, T settings, 100 replications (standard deviations are shown in parentheses).

	RMSE	Er($X\Theta$)	FPR%	FNR%	CPU time
		$p = 50, q = 25;$	$r = 2, T = 400$		
OFGD	2.1656(0.2698)	1.6229(0.3595)	100(0)	0(0)	13.63(0.22)
SFGD	0.6438(0.0288)	0.2938(0.0181)	3.42(1.24)	0(0)	15.70(1.17)
MLE	6.1391(0.4325)	4.7971(0.3621)	100(0)	0(0)	43.82(3.35)
		$p = 100, q = 100;$	$r = 2, T = 200$		
OFGD	1.9767(0.1276)	2.8220(0.4002)	100(0)	0(0)	12.18(0.22)
SFGD	0.4858(0.0736)	0.5721(0.0542)	1.98(0.46)	0(0)	14.31(1.45)
MLE	10.4705(0.5249)	13.8533(0.6313)	100(0)	0(0)	76.82(5.25)
		$p = 300, q = 100;$	$r = 2, T = 200$		
OFGD	1.8704(0.0902)	4.6054(0.3512)	100(0)	0(0)	20.35(4.13)
SFGD	0.3866(0.0873)	0.8175(0.0463)	2.17(0.31)	1.35(0.47)	44.78(7.48)
MLE	13.2192(0.6139)	16.8519(0.8791)	100(0)	0(0)	133.37(9.21)

4.2. Regret for low-rank and sparse structure

Next, we consider the dynamic assortment personalization problem. We compare the average regret [27] of 3 mothods. One alternative is the structure-ignorant algorithm in which we fit a single MNL model by MLE to whole population without the low-rank and sparse structure on $\boldsymbol{\Theta}$. We first illustrate the definition of average regret as follows

Definition 4.1. Given an instance $(p, q, \boldsymbol{\Theta}^*)$, the average regret of algorithm π at time T is

$$\text{AveRegret}(T; \pi) = \mathbb{E}^{\pi_{\Theta^*}}\left[\frac{1}{T}\sum_{t=1}^{T} r_t\right] - \mathbb{E}^{\pi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t\right].$$

In the simulation of average regret, the feature matrix $\boldsymbol{X}$ and the underlying $\boldsymbol{\Theta}^*$ are generated as the previous simulation on estimation accuracy. Now we construct the ture revenue of each product as follows:

1. K out of q items have revenue parameters $r_i = 1$.
2. For the other $(q - K)$ items, both their revenue are uniformly distribution on $[0.05, 0.1]$.

To make a comparsion, we also introduce an average regret $O(r\max(p, q)\frac{\log(T)}{T})$ of [1] as a baseline that consider customer types rather than features data.

In Figure 1, we report all the result for different settings of p, q and T after 100 replications. From Figure 1, we can see that SFGD has lower average regret than that of both OFGD and MLE, which means both low-rank and sparse structures have the power to reduce the regret level. And with the growing of types p and items q, considering both low-rank and sparse structure is more nearly to the baseline.

We can find that the SFGD method will stabilize at a mean regret level much lower than the OFGD and MLE methods. Before reaching the minimum regret level, with growing of the time horizon T, the average regret will decrease for SFGD, while for OFGD and MLE the regret will not decrease with larger T. Futhermore, in all the settings, the OFGD method that only use the low-rank structure enjoys a better performance than the structure-ignorant MLE. Therefore, our results confirm the necessity of the sparse recovery as well as the effectiveness of our proposed algorithm for dynamic assortment personalization problem.

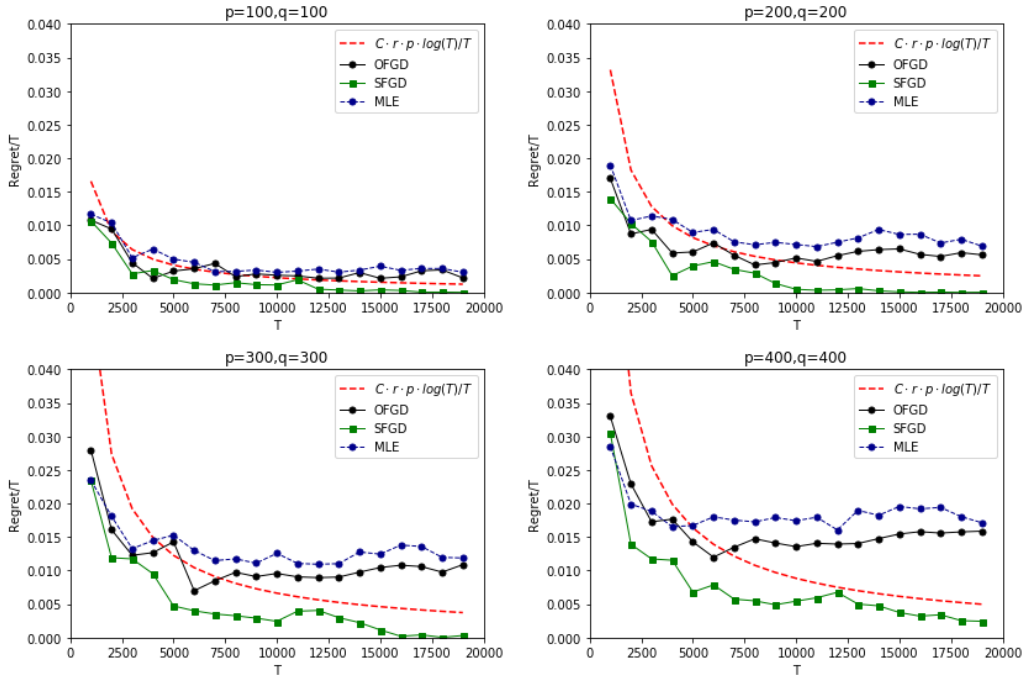

Figure 1. Comparison of average regret between our proposed methods OFGD, SFGD and MLE. Time horizon T ranges from 1000 to 20000. The constant of baseline is chosen as $C = 0.12$.

The Structure-aware method that containing low-rank and sparsity structures is of great significance when handling the large scale customer features data espicially in the high dimensional senario $p \geq T$. We also see the effective varibles selection capability of the grouped lasso type shrinkage method on $\mathbf{\Theta}$, which is computed iteratively through our proposed SFGD method. Futhermore, with the growth of horizon T, it's SFGD that always enjoys the lowest average regret among different algorithms.

5. Discussion

In this paper, we focus on the assortment personalization problem using the data-driven conditionally multinomial logit choice model where the sparse and low-rank settings of parameter matrix are considered. Then we present SFGD method for our penalized maximum likelihood problem (that is, a negtive likelihood loss function plus certain penalties), leading to computational efficiency. Moreover, we prove that SFGD enjoy the local convergence property, and the simulations show that SFGD achieve nice estimation accuracy and features selection ability with massive and high-dimensional data.

For future research, one interesting direction is the non-asymptotic analysis of the multinomial logit penalized likelihood in high-dimensional setting, which describes the estimation accuracy statistically. Another research direction is to extend SFGD method to the co-sparse framework which consider the both rowwise and columnwise sparsity of parameter matrix.

Acknowledgements

The authors sincerely thank the editor, the associate editor, and the referees for their valuable comments that helped improve the article substantially.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Natural Science Foundation of China (Grants 72071187, 11671374, 71731010, and 71921001) and Fundamental Research Funds for the Central Universities (Grants WK3470000017 and WK2040000027).

References

[1] Kallus N, Udell M. Dynamic assortment personalization in high dimensions. Oper Res. 2020;68(4):1020-1037.

[2] Bernstein F, Kök AG, Xie L. Dynamic assortment customization with limited inventories. Manufacturing & Service Oper Management. 2015;17(4):538-553.

[3] Golrezaei N, Nazerzadeh H, Rusmevichientong P. Real-time optimization of personalized assortments. Management Sci. 2014;60(6):1532-1551.

[4] Chen X, Owen Z, Pixton C, et al. A statistical learning approach to personalization in revenue management. Management Sci. 2021;0(0):1-19.

[5] Luo X, Andrews M, Fang Z, et al. Mobile targeting. Management Sci. 2014;60(7):1738-1756.

[6] Xue Z, Wang Z, Ettl M. Pricing personalized bundles: A new approach and an empirical study. Manufacturing & Service Oper Management. 2016;18(1):51-68.

[7] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267-288.

[8] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B. 2006;68(1):49-67.

[9] Meier L, Van De Geer S, Bhlmann P. The group lasso for logistic regression. J R Stat Soc Ser B. 2008;70(1):53-71.

[10] Negahban S, Wainwright M J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann Statist. 2011;39(2):1069-1097.

[11] Chen K, Chan KS, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. J R Stat Soc Ser B. 2012;74(2):203-221.

[12] Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am Statist Ass. 2012;107(500):1533-1545.

[13] Chen K, Dong H, Chan K S. Reduced rank regression via adaptive nuclear norm penalization. Biometrika. 2013;100(4):901-920.

[14] Zhou K, Zha H, Song L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In: Artif Intel and Statist; PMLR; 2013. p. 641-649.

[15] Wang Y X, Xu H, Leng C. Provable Subspace Clustering: When LRR meets SSC. In: NIPS; 2013. p. 5.

[16] Feng J, Lin Z, Xu H, et al. Robust subspace segmentation with block-diagonal prior. In: Proc IEEE Conf Comput vision and Pattern Recognit; IEEE; 2014. p. 3818-3825.

[17] Boyd S, Parikh N, Chu E. Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc; Vol. 3, 2011.

[18] Mishra A, Dey DK, Chen K. Sequential co-sparse factor regression. J Comput Graph Statist. 2017;26(4):814-825.

[19] Zheng Z, Bahadori MT, Liu Y, et al. Scalable Interpretable Multi-Response Regression via SEED. J Mach Learn Res. 2019;20(107):1-34.

[20] Jain P, Netrapalli P, Sanghavi S. Low-rank matrix completion using alternating minimization. In: Proc 45th Ann ACM Symp on Theory of Comput; ACM; 2013. p. 665-674.

[21] Bhojanapalli S, Kyrillidis A, Sanghavi S. Dropping convexity for faster semi-definite optimization. In: Conf on Learning Theory; PMLR; 2016. p. 530-582.

[22] Golrezaei N, Nazerzadeh H, Rusmevichientong P. Real-time optimization of personalized assortments. Management Sci. 2014;60(6):1532-1551.

[23] Train KE. Discrete choice methods with simulation. Cambridge university press; 2nd ed, 2009.

[24] Gorski J, Pfeuffer F, Klamroth K. Biconvex sets and optimization with biconvex functions: a survey and extensions. Mathemat Methods of Oper Res. 2007;66(3):373-407.

[25] Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. J R Stat Soc Ser B. 2013;75(3):531-552.

[26] Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B. 2009;71(3):671-683.

[27] Chen X, Krishnamurthy A, Wang Y. Robust dynamic assortment optimization in the presence of outlier customers. arXiv preprint arXiv:1910.04183. 2019.

[28] Mirsky L. A trace inequality of John von Neumann. Monatshefte fr mathematik. 1975;79(4):303-306.

Appendix A. Details on Sparse Factored Gradient Descent

We first let $\boldsymbol{e}_l$ be the lth unit vector with lth element 1 and others zeros, and $\boldsymbol{e}_t \in \mathbb{R}^T, \boldsymbol{e}_j \in \mathbb{R}^q$. Then we rewrite the loss

$$\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta}) = \frac{1}{T}\sum_{t=1}^{T}\left(log\left(1+\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{e}_j}\right) - \boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{e}_{j_t}\right) \tag{A1}$$

and this leads to the gradient and Hessian with respect to $\boldsymbol{\Theta}$ as

$$\begin{aligned}
\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta}) &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}^T\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{e}_j}\boldsymbol{e}_t\boldsymbol{e}_j^T}{1+\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{e}_j}} - \boldsymbol{e}_t\boldsymbol{e}_{j_t}^T\right). \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j}\boldsymbol{x}_t\boldsymbol{e}_j^T}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j}} - \boldsymbol{x}_t\boldsymbol{e}_{j_t}^T\right). \\
\nabla^2\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta}) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j}(\boldsymbol{x}_t\boldsymbol{e}_j^T)^{\otimes 2}}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j}} - \frac{(\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j}\boldsymbol{x}_t\boldsymbol{e}_j^T)^{\otimes 2}}{(1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{\Theta}\boldsymbol{e}_j})^2}\right) \qquad \text{(A2)}
\end{aligned}$$

where $\boldsymbol{Z}^{\otimes 2} = \boldsymbol{Z} \otimes \boldsymbol{Z}$ is the symmetric linear operator on matrices, and $\nabla\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta}), \nabla^2\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta})$ have size $p \times q$ and $pq \times pq$. Since we have the chain rule as

$$\begin{aligned}
\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \nabla_{\boldsymbol{U}\boldsymbol{V}^T}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} \\
\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \nabla_{\boldsymbol{U}\boldsymbol{V}^T}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T)^T\boldsymbol{U},
\end{aligned}$$

then use the result above, we obtain

$$\begin{aligned}
\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}^T\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{e}_j}\boldsymbol{e}_t\boldsymbol{e}_j^T}{1+\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{e}_j}} - \boldsymbol{e}_t\boldsymbol{e}_{j_t}^T\right)\boldsymbol{V} \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}\boldsymbol{x}_t\boldsymbol{v}_j^T}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}} - \boldsymbol{x}_t\boldsymbol{v}_{j_t}^T\right) \qquad \text{(A3)} \\
\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{X};\boldsymbol{U}\boldsymbol{V}^T) &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{e}_j}\boldsymbol{e}_j\boldsymbol{e}_t^T}{1+\sum_{j\in S_t} e^{\boldsymbol{e}_t^T\boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{e}_j}} - \boldsymbol{e}_{j_t}\boldsymbol{e}_t^T\right)\boldsymbol{X}\boldsymbol{U} \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}\boldsymbol{e}_j\boldsymbol{x}_t^T}{1+\sum_{j\in S_t} e^{\boldsymbol{x}_t^T\boldsymbol{U}\boldsymbol{v}_j}} - \boldsymbol{e}_{j_t}\boldsymbol{x}_t^T\right)\boldsymbol{U}, \qquad \text{(A4)}
\end{aligned}$$

which clarifies the gradient descent direction in Algorithm 1.

Appendix B. Proof of Theorem 3.1

Proof. For simplicity,when $\boldsymbol{X}$ is fixed, we use $\mathcal{Q}(\boldsymbol{\Theta})$ and $\mathcal{L}(\boldsymbol{\Theta})$ instead of $\mathcal{Q}(\boldsymbol{X};\boldsymbol{\Theta})$ and $\mathcal{L}(\boldsymbol{X};\boldsymbol{\Theta})$. From the smoothness of $\mathcal{Q}$ and $\mathcal{L}$, there exists a common constant M such that $\mathcal{Q}(\boldsymbol{\Theta})$ and $\mathcal{L}(\boldsymbol{\Theta})$ are both M-smooth. According to the overlapping of $\boldsymbol{U}$

and $\boldsymbol{V}$, one of $\boldsymbol{U}$ and $\boldsymbol{V}$ updates with the other one fixed. We now devide the problem into the $\boldsymbol{U}$-step and the $\boldsymbol{V}$-step, and we start from the $\boldsymbol{V}$-step.

The $\boldsymbol{V}$-step. Let $\boldsymbol{\Theta}^V = \boldsymbol{U}\boldsymbol{V}'^T$. In the $\boldsymbol{V}$-step, by the chain rule, the updating of $\boldsymbol{V}$ satisfies

$$\boldsymbol{V}' = \boldsymbol{V} - \eta\nabla_{\boldsymbol{V}}\mathcal{Q}(\boldsymbol{\Theta}) = \boldsymbol{V} - \eta[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U},$$

Then $\boldsymbol{\Theta}^V = \boldsymbol{U}\boldsymbol{V}'^T = \boldsymbol{U}[\boldsymbol{V} - \eta\boldsymbol{U}\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{\Theta})]^T = \boldsymbol{\Theta} - \eta\boldsymbol{U}\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{\Theta})$. From the M-smooth of $\mathcal{L}(\boldsymbol{\Theta})$ we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}^V) &\le \mathcal{L}(\boldsymbol{\Theta}) + \big\langle\nabla\mathcal{L}(\boldsymbol{\Theta}), \boldsymbol{\Theta}' - \boldsymbol{\Theta}\big\rangle + \frac{M}{2}||\boldsymbol{\Theta}' - \boldsymbol{\Theta}||_F^2 \\ &= \mathcal{L}(\boldsymbol{\Theta}) + \mathrm{Tr}(\nabla\mathcal{L}(\boldsymbol{\Theta})(\boldsymbol{\Theta}^V - \boldsymbol{\Theta})^T) + \frac{M}{2}\mathrm{Tr}((\boldsymbol{\Theta}' - \boldsymbol{\Theta})^T(\boldsymbol{\Theta}' - \boldsymbol{\Theta})).\end{aligned}$$

By the property of trace, it entails

$$\begin{aligned}\mathrm{Tr}(\nabla\mathcal{L}(\boldsymbol{\Theta})(\boldsymbol{\Theta}^V - \boldsymbol{\Theta})^T) &= -\eta\mathrm{Tr}(\nabla\mathcal{L}(\boldsymbol{\Theta})[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}\boldsymbol{U}^T) \\ &= -\eta\mathrm{Tr}([\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{\Theta})) \\ &= -\eta||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2.\end{aligned}$$

And by the Von Neumann's trace inequality [28] we have

$$\begin{aligned}\mathrm{Tr}((\boldsymbol{\Theta}' - \boldsymbol{\Theta})^T(\boldsymbol{\Theta}' - \boldsymbol{\Theta})) &= \eta^2\mathrm{Tr}([\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{\Theta})) \\ &\le \eta^2||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2 \cdot \sigma_1(\boldsymbol{U}^T\boldsymbol{U}).\end{aligned}$$

Since the step size satisfies $\eta \le \frac{\beta^k}{3M(\sigma_1(\boldsymbol{U}^T\boldsymbol{U})+1)} < \frac{\beta^k}{3M\sigma_1(\boldsymbol{U}^T\boldsymbol{U})}$, then we derive

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}^V) &\le \mathcal{L}(\boldsymbol{\Theta}) - \eta||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2 + \frac{M\eta^2}{2}||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2\sigma_1(\boldsymbol{U}^T\boldsymbol{U}) \\ &\le \mathcal{L}(\boldsymbol{\Theta}) - \eta||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2 + \frac{\eta}{6}\beta^k||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2 \\ &\le \mathcal{L}(\boldsymbol{\Theta}) - \frac{5}{6}\eta\beta^k||[\nabla\mathcal{L}(\boldsymbol{\Theta})]^T\boldsymbol{U}||_F^2. \qquad \text{(B1)}\end{aligned}$$

The $\boldsymbol{U}$-step. Since the fixed $\boldsymbol{V}$ satisfies that $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$, we then have a useful form of $\mathcal{Q}(\boldsymbol{\Theta})$ as

$$\mathcal{Q}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{\Theta}) + \lambda||\boldsymbol{\Theta}\boldsymbol{V}||_{2,1}.$$

Let $\boldsymbol{\Theta}^U = \boldsymbol{U}'\boldsymbol{V}^T$, according to the updating rule, we have

$$\boldsymbol{U}' = \boldsymbol{U} - \eta\nabla_{\boldsymbol{U}}\mathcal{Q}(\boldsymbol{\Theta}) = \boldsymbol{U} - \eta\nabla\mathcal{Q}(\boldsymbol{\Theta})\boldsymbol{V}\boldsymbol{V}^T.$$

Here we don't need the explicit expression of $\nabla\mathcal{Q}(\boldsymbol{\Theta})$, and from the M-smooth of function $\mathcal{Q}(\mathbf{\Theta})$

$$\begin{aligned}
\mathcal{Q}(\mathbf{\Theta}^U) &= \mathcal{Q}(\mathbf{\Theta}) + \left\langle \nabla\mathcal{Q}(\mathbf{\Theta}), \mathbf{\Theta}^U - \mathbf{\Theta} \right\rangle + \frac{M}{2}||\mathbf{\Theta}^U - \mathbf{\Theta}||_F^2 \\
&= \mathcal{Q}(\mathbf{\Theta}) - \eta \mathrm{Tr}(\nabla\mathcal{Q}(\mathbf{\Theta})\boldsymbol{V}\boldsymbol{V}^T[\nabla\mathcal{Q}(\mathbf{\Theta})]^T) \\
&+ \frac{M\eta^2}{2}\mathrm{Tr}(\boldsymbol{V}\boldsymbol{V}^T[\nabla\mathcal{Q}(\mathbf{\Theta})]^T\nabla\mathcal{Q}(\mathbf{\Theta})\boldsymbol{V}\boldsymbol{V}^T) \\
&\leq \mathcal{Q}(\mathbf{\Theta}) - \eta||\nabla\mathcal{Q}(\mathbf{\Theta})\boldsymbol{V}||_F^2 + \frac{M\eta^2}{2}||\nabla\mathcal{Q}(\mathbf{\Theta})\boldsymbol{V}||_F^2\sigma_1(\boldsymbol{V}^T\boldsymbol{V}) \\
&\leq \mathcal{Q}(\mathbf{\Theta}) - \frac{5}{6}\eta\beta^k||\nabla\mathcal{Q}(\mathbf{\Theta})\boldsymbol{V}||_F^2, \qquad \text{(B2)}
\end{aligned}$$

where the first inequality is Von Neumann's and the second inequality is from the fact $\sigma_1(\boldsymbol{V}^T\boldsymbol{V}) = 1$ and $\eta \leq \frac{\beta^k}{3M(\sigma_1(\boldsymbol{U}^T\boldsymbol{U})+1)} < \frac{\beta^k}{3M}$.

In $\mathbf{\Theta'} = \boldsymbol{U'}\boldsymbol{V}'^T$, from the results above, we can treat $\boldsymbol{V'}$ as the fixed part firstly, by (B2) it yields

$$\mathcal{Q}(\mathbf{\Theta'}) \leq \mathcal{Q}(\mathbf{\Theta}^V) - \frac{5}{6}\eta\beta^k||\nabla\mathcal{Q}(\mathbf{\Theta}^V)\boldsymbol{V}||_F^2.$$

In the next place, $\boldsymbol{U}$ is fixed and it enters the $\boldsymbol{V}$-step. And $\nabla\mathcal{Q}(\mathbf{\Theta}) = \nabla\mathcal{L}(\mathbf{\Theta})$ holds in this step, thus by (B1) we have $\mathcal{Q}(\mathbf{\Theta}^V) \leq \mathcal{Q}(\mathbf{\Theta}) - \frac{5}{6}\eta\beta^k||[\nabla\mathcal{L}(\mathbf{\Theta})]^T\boldsymbol{U}||_F^2$, then

$$\begin{aligned}
\mathcal{Q}(\mathbf{\Theta'}) &\leq \mathcal{Q}(\mathbf{\Theta}) - \frac{5}{6}\eta\beta^k||[\nabla\mathcal{L}(\mathbf{\Theta})]^T\boldsymbol{U}||_F^2 - \frac{5}{6}\eta\beta^k||\nabla\mathcal{Q}(\mathbf{\Theta}^V)\boldsymbol{V}||_F^2 \\
&\leq \mathcal{Q}(\mathbf{\Theta}) - \frac{5}{6}\eta\beta^k||[\nabla\mathcal{L}(\mathbf{\Theta})]^T\boldsymbol{U}||_F^2.
\end{aligned}$$

Since $\mathcal{L}(\mathbf{\Theta})$ is constructed by negtive likelihood, it entails

$$\mathcal{Q}(\mathbf{\Theta}) \geq \mathcal{L}(\mathbf{\Theta}) = -\frac{1}{T}\sum_{t=1}^{T}\log(\mathbb{P}_{\mathbf{\Theta}}(J = j_t | I = \boldsymbol{x}_t; S_t)) \geq -1,$$

where $\mathbb{P}_{\mathbf{\Theta}}(J = j_t|I = \boldsymbol{x}_t; S_t)$ is the conditional probability defined in (4). The $\mathcal{Q}(\mathbf{\Theta})$ has a lower bound and the value of $\mathcal{Q}(\mathbf{\Theta})$ descents iteratively through SFGD, which shows there exist a local optimum $\tilde{\mathbf{\Theta}}$ such that $\mathcal{Q}(\mathbf{\Theta})$ converages to $\mathcal{Q}(\tilde{\mathbf{\Theta}})$ by simple analysis. □