Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Adaptive Convex Clustering of Generalized Linear Models With Application in Purchase Likelihood Prediction

Shuyu Chu, Huijing Jiang, Zhengliang Xue & Xinwei Deng

To cite this article: Shuyu Chu, Huijing Jiang, Zhengliang Xue & Xinwei Deng (2020): Adaptive Convex Clustering of Generalized Linear Models With Application in Purchase Likelihood Prediction, Technometrics, DOI: 10.1080/00401706.2020.1733094

To link to this article: https://doi.org/10.1080/00401706.2020.1733094

TECHNOMETRICS
2020, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/00401706.2020.1733094

Adaptive Convex Clustering of Generalized Linear Models With Application in Purchase Likelihood Prediction

Shuyu Chu[a], Huijing Jiang[a], Zhengliang Xue[a], and Xinwei Deng[b]

[a]IBM T. J. Watson Research Center, Yorktown Heights, NY; [b]Department of Statistics, Virginia Tech, Blacksburg, VA

ABSTRACT
In the pricing of customized products, it is challenging to accurately predict the purchase likelihood of potential clients for each personalized request. The heterogeneity of customers and their responses to the personalized products leads to very different purchase behavior. Thus, it is often not appropriate to use a single model to analyze all the pricing data. There is a great need to construct distinctive models for different data segments. In this work, we propose an adaptive convex clustering method to perform data segmentation and model fitting simultaneously for generalized linear models. The proposed method segments data points using the fused penalty to account for the similarity in model structures. It ensures that the data points sharing the same model structure are grouped into the same segment. Accordingly, we develop an efficient algorithm for parameter estimation and study its consistency properties in estimation and clustering. The performance of our approach is evaluated by both numerical examples and case studies of real business data.

ARTICLE HISTORY
Received August 2018
Accepted February 2020

KEYWORDS
Adaptive weight; Classification; Network lasso; Segmentation

1. Introduction

In the precision marketing, companies offering highly customized products/services encounter the challenge on how to predict the purchase likelihood of potential clients for each personalized request. To optimize the Request-for-Quote (RFQ) pricing decision, it is essential to accurately predict the seller's win probability of any quotes. The win probability means the likelihood that a prospective buyer will complete the purchase after receiving the seller's price offering for each RFQ. In practice, a seller provides a variety of products for which a customer can construct a personalized bundle (combination of products) and submit the RFQ to the seller. With regard to each RFQ, the seller will make a pricing decision to maximize the expected revenue (or profit) based on the win probability estimation.

Offering personalized bundles will give customers tremendous opportunity to configure their customized RFQs. However, it also poses a big challenge to estimate the win probability for almost unlimited number of possible product bundles. Since the majority of the RFQs are configured in a distinctive way, traditional segmentation and regression models are usually not applicable. Xue, Wang, and Ettl (2015) proposed a top-down and bottom-up method to decompose the bundles and aggregate back for creating bundle features. In their method, segmentation is conducted based on the seller's pricing decisions, while the win probability in each segment is estimated in response to the customers' purchase behavior. A potential issue of this approach is the disparity between segmentation and probability modeling when sellers' pricing decision cannot fully reflect customers' purchase behavior. Our article proposes an approach that integrates the segmentation and predictive modeling in a coherent framework that is purely based on customer behaviors.

A variety of techniques on segmentation have been studied in the literature. Clustering is one of the most common unsupervised learning techniques to explore data and group similar objects together. The conventional k-means clustering and hierarchical clustering (Johnson 1967; Hartigan and Wong 1979) group data points mainly based on the similarity of input features or covariates, regardless of the modeling performance on the responses. To cluster data points based on the similarities of their model structures, one can conduct k-means clustering based on model coefficients $\boldsymbol{\beta}$, or model the data via mixture models where latent variables are used to indicate the clustering membership (Muthén 2001; Jung and Wickrama 2008).

While these clustering approaches mainly focus on the similarity of either input features or modeling structures, one growing interest is to identify clusters accounting for both aspects. Region-specific linear models have been proposed in the literature to partition the data based on the performance of linear classifiers in each region of input features (Wang and Saligrama 2012; Jose et al. 2013). One disadvantage of this method arises from the nonconvex objective function and lack of generalization error analysis (Oiwa and Fujimaki 2014). To overcome the challenges, Oiwa and Fujimaki (2014) proposed the partition-wise linear models to achieve partitions by means of convex structured regularizations. In addition to the parametric methods, Qiu (2011) proposed a nonparametric approach assuming the underlying regression

CONTACT Xinwei Deng xdeng@vt.edu Department of Statistics, Virginia Tech, Blacksburg, VA 24061.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.

function has jumps. When the number of jumps is unknown, this method requires to identify all the possible candidate jumps through a series of hypothesis tests, which could be challenging.

Recently, the convex clustering method has received increasing attentions on finding the groups of similar objects via fused regularizer (Hocking et al. 2011; Lindsten, Ohlsson, and Ljung 2011; Chen et al. 2015; Wang et al. 2016; Radchenko and Mukherjee 2017). Convex clustering was introduced by Hocking et al. (2011) and Lindsten, Ohlsson, and Ljung (2011) as a convex relaxation of hierarchical clustering. Later, Hallac, Leskovec, and Boyd (2015) extended it to a more general convex optimization problem. By imposing a penalty on the differences of model coefficient vectors, data points sharing the same coefficient values are clustered into the same segment automatically. Different from many existing clustering algorithms which are greedy by nature (Chi and Lange 2015), the convex clustering model can be solved efficiently with a global optimum because of the convex objective function. Moreover, it conducts model estimation and clustering at the same time without prespecifying the number of clusters. A sequence of clustering results can be obtained by varying the regularization parameter.

However, the fused lasso based regularization in the convex clustering model tends to inappropriately shrink model coefficient estimates, making the estimates' absolute values smaller than the actual. It could result in the biased estimation of model parameters and produce suboptimal solutions in parameter estimation (Meinshausen and Bühlmann 2006; Zou 2006; Chen et al. 2015; Lange and Keys 2015; Wang and Wang 2014). Moreover, under the context of convex clustering for GLMs, the commonly used algorithms such as alternating direction method of multipliers (ADMM) (Hallac, Leskovec, and Boyd 2015) are likely to get stuck on local optimum for nonlinear objective functions.

In this article, we propose an *adaptive convex clustering approach* to conduct segmentation and model fitting in a simultaneous fashion for GLMs. The proposed method aims to address both the shrinkage issue in the model estimation and the local convergence issue in the optimization. It focuses on the convex clustering of both input features and modeling structures under the generalized linear regression scenario, where the optimization function could be nonlinear. The key contributions of this work are summarized as follows.

First, we impose adaptive convex clustering penalties on the differences of model coefficient vectors, encouraging homogeneity within each segment and heterogeneity across different segments. Instead of assigning prespecified weights on the regularization parameter, we use the adaptive weights to alleviate the shrinkage problems of model coefficients. More weights will be assigned to data points that are likely coming from the same cluster to push their coefficient estimates being exactly the same. Meanwhile, less weight will be given to data points from different clusters to achieve less biased and more accurate coefficient estimation. A Bayesian interpretation on how to assign the weights is also discussed.

Second, we develop an iteratively weighted least squares (IWLS) based algorithm for efficient parameter estimation. The key idea is to linearize the nonlinear objective functions and update all parameters simultaneously at each iteration. Compared with the ADMM-based algorithm, the proposed IWLS-based algorithm shows the superior performance of converging to the global optimum in the empirical study.

Finally, we investigate the asymptotic properties of the adaptive convex clustering in the context of GLMs. In the current literature, Tan and Witten (2015) proved the close connection between convex clustering and single linkage hierarchical clustering, and provided a finite sample bound for the prediction error of convex clustering. Radchenko and Mukherjee (2017) studied theoretical properties of a problem closely related to convex clustering. However, the asymptotic properties of adaptive convex clustering for GLMs have not yet been comprehensively studied.

The rest of this article is organized as follows. After a brief literature review in Section 2, we detail the proposed adaptive convex clustering model for GLMs in Section 3, where a Bayesian interpretation and asymptotic properties are presented. The IWLS-based algorithm for parameter estimation is developed in Section 4. Section 5 demonstrates the performance of proposed method via several numerical examples. A real application to the RFQ pricing is analyzed in Section 6. Section 7 concludes our work with further discussion.

2. Review of Convex Clustering Approach

The convex clustering was originally developed as a convex relaxation of the hierarchal clustering (Hocking et al. 2011). Denote the N data observations by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ in a data matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$, where each row denotes a data point, and p is the number of covariates. The convex clustering optimization problem (Hocking et al. 2011; Lindsten, Ohlsson, and Ljung 2011; Radchenko and Mukherjee 2017) is formulated through clustering the rows as follows,

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{c}_i||_2^2 + \lambda \sum_{i<j} w_{ij}||\boldsymbol{c}_i - \boldsymbol{c}_j||_q, \qquad (1)$$

where λ is a positive tuning parameter, and $w_{ij} \geq 0$ is a prespecified weight to incorporate the prior information of clustering. Here $\boldsymbol{c}_i \in \mathbb{R}^p$ is denoted as the cluster center attached to the data point $\boldsymbol{x}_i$. Different norms on the differences $\boldsymbol{c}_i - \boldsymbol{c}_j$ were considered in the literature. For example, $q = \{1, 2, \infty\}$ were analyzed in Hocking et al. (2011).

Hallac, Leskovec, and Boyd (2015) extended the objective function of the convex clustering to a large variety of convex functions. They proposed to cluster the data points based on the similarity of their corresponding model coefficients by generalizing group lasso to a network setting. Both the number of clusters and the model coefficients are unknown parameters and estimated simultaneously during the model estimation procedure. Denote the complete set of observations by $(\boldsymbol{u}_i, y_i), i = 1, \dots, N$, where y_i is the variable in response to $\boldsymbol{u}_i$. Furthermore, the vector $\boldsymbol{u}_i$ is an union of model predictors $\boldsymbol{x}_i$ and the clustering features $\boldsymbol{v}_i$. The model predictors $\boldsymbol{x}_i$ contain p covariates used in the regression model, while the clustering features $\boldsymbol{v}_i$ could either be chosen from $\boldsymbol{x}_i$ or totally different from it.

Let $\mathcal{E}$ be the set of pairwise indices between two data points, whose model coefficients are regularized. Thus, $E = |\mathcal{E}|$ is the total number of pairs to be regularized in the model. Note that each observation i can have another observation j such that $(i,j) \in \mathcal{E}$, and $j \neq i$. Then the convex optimization problem is defined as follows,

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) + \lambda \sum_{(j,k)\in\mathcal{E}} w_{jk} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2, \qquad (2)$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N \in \mathbb{R}^{p+1}$ are unknown coefficient vectors at N observations. The $f_i(\cdot)$ is the loss function defined for each observation i.

The penalty term on model coefficients $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_k\|$, for any $(j,k) \in \mathcal{E}$, aims to cluster data that share similar model coefficients into the same segment. The segments are automatically formed according to the coefficient values related to each observation. Clearly, the overall regularization parameter λ controls the amount of penalization. Here a prespecified weight w_{jk} between data points j and k is used to incorporate the prior knowledge on how likely the data points j and k belong to the same segment based on the distance of clustering features. For example, we can specify the weight as $w_{jk} \propto \frac{1}{\|\boldsymbol{v}_j - \boldsymbol{v}_k\|}$. By using this weight, the data with similar clustering features tend to be grouped into the same segment.

3. The Proposed Adaptive Convex Clustering

3.1. Shrinkage Problem in Convex Clustering

In the convex clustering model in (2), a global shrinkage parameter λ controls the overall penalization, and the w_{jk} is prespecified by the clustering features. Thus, the overall penalty weight λw_{jk} is unchanged throughout the model optimization, which may cause inappropriate shrinkage for the coefficient estimates, especially when the paired observations $\boldsymbol{u}_j, \boldsymbol{u}_k$ $((j,k) \in \mathcal{E})$ actually come from different segments.

Let us take a simple example to elaborate this point. Suppose that a dataset, as shown in Figure 1, contains two true clusters in the space spanned by the clustering features, $\boldsymbol{v}_j = \left(u_{1j}, u_{2j}\right)'$. Each cluster has its own set of coefficients (β_0, β_1) corresponding to the logistic regression with a single covariant $x_j = u_{3j}$. The binary responses are simulated separately in each cluster. The true coefficients of two clusters and the corresponding estimates from the convex clustering in (2) are listed in Table 1. The number of clusters in the convex clustering can be determined by tuning the regularization parameter λ. More details can be found in Section 5. To demonstrate the shrinkage problem in the convex clustering, we refit the logistic regression for each obtained cluster from the convex clustering model. The GLM-refitted estimates of coefficients in Table 1 thus serve as the benchmark with no shrinkage issues. From Table 1, it is clearly seen that the absolute values of convex clustering estimates are always smaller than those in the GLM-refit. It indicates that the coefficients are highly shrunk during the estimation procedure in the convex clustering method.

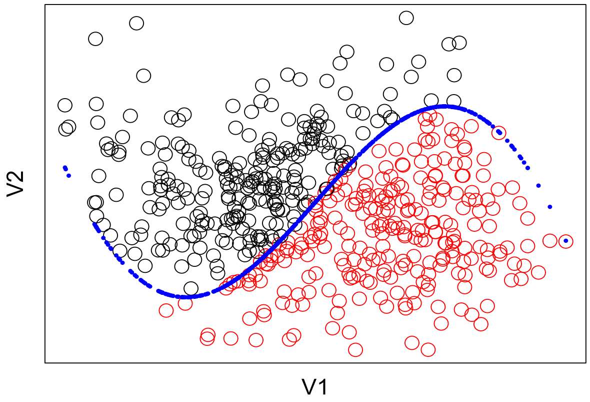


Figure 1. Simulated data with two clusters in the space spanned by clustering feature of V_1, V_2.

Table 1. Estimated model coefficients of the convex clustering method.

Segment		True	GLM-refit	Convex clustering
1	β_{01}	−1	−0.646	−0.601
	β_{11}	2.5	1.901	1.414
2	β_{02}	1.5	1.365	0.719
	β_{12}	−3.5	−3.504	−1.798

Note that the prespecified weight can be calculated as $w_{jk} \propto \frac{1}{\|\boldsymbol{v}_j - \boldsymbol{v}_k\|}$. Two observations $\boldsymbol{u}_j, \boldsymbol{u}_k$ are paired (i.e., $(j,k) \in \mathcal{E}$) based on the similarity measures between $\boldsymbol{v}_j$ and $\boldsymbol{v}_k$, regardless of whether they truly belong to the same segment. Thus, “wrong” pairs are inevitably created, especially for data points near the boundary of clusters (the blue curve shown in Figure 1). For two observations $\boldsymbol{u}_j, \boldsymbol{u}_k$ with similar measures, if the underlying truth is that they are in the same segment, then λw_{jk} will correctly push $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ toward zero. On the other hand, if $\boldsymbol{u}_j, \boldsymbol{u}_k$ belong to different segments, the value of $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ is supposed to be nonzero. However, they are still pushed toward zero incorrectly to minimize the overall penalty term. As a result, the absolute values of $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ are seriously shrunk. This inappropriate shrinkage would result in the biased coefficient estimates for paired observations from different segments and induce suboptimal estimation risk (Zou 2006). Meinshausen and Bühlmann (2006) also showed the disagreement of optimal prediction and the accurate estimation of the true model in the context of penalized regressions.

The model coefficients play an essential role in the estimation of purchase likelihood prediction since both data segmentation and model prediction are highly influenced by the estimation accuracy of model coefficients. To alleviate the shrinkage problem, we develop an adaptive convex clustering model in Section 3.2, where an adaptive shrinkage parameter $\lambda_{jk}^{(t)}$ is introduced to penalize coefficient differences $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ at iterative t.

3.2. The Adaptive Convex Clustering

As discussed in Section 3.1, the shrinkage problem in the convex clustering is mainly caused by the prespecified penalization of coefficient differences $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$. Inspired by the adaptive lasso in Zou (2006), a simple and effective remedy is to assign adaptive weights to different coefficient pairs. The proposed adaptive convex clustering model can be considered as a generalization of the conventional convex clustering. Specifically, we update the regularization parameter $\lambda_{jk}^{(t+1)}$ in each iteration according to $||\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_k^{(t)}||_2$. If $||\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_k^{(t)}||_2$ is large, $(\boldsymbol{u}_j, \boldsymbol{u}_k)$ are more likely from the same cluster and we penalize them less in next iteration $(t+1)$. Thus at iteration $(t+1)$, the estimation problem in (2) becomes,

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \sum_{i=1}^{N} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) + \sum_{(j,k)\in\mathcal{E}} \lambda_{jk}^{(t+1)} w_{jk} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2, \tag{3}$$

where $\lambda_{jk}^{(t+1)} \propto \frac{1}{||\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_k^{(t)}||_2}$ and $\boldsymbol{\beta}_j^{(t)}$ is the updated estimate of $\boldsymbol{\beta}_j$ at iteration $(t+1)$.

Obviously, the formulation in (3) is equivalent to using one global λ with adaptive penalty weight,

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \sum_{i=1}^{N} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) + \lambda \sum_{(j,k)\in\mathcal{E}} \gamma_{jk}^{(t+1)} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2, \tag{4}$$

where $\gamma_{jk}^{(t+1)} \propto \frac{w_{jk}}{||\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_k^{(t)}||_2}$ is the adaptive weight at iteration $(t+1)$.

When the loss function f_i is the negative log-likelihood function, we can justify the adaptive weights in our proposed adaptive convex clustering model from the Bayesian point of view. Note that the construction of adaptive weights is equivalent to assigning an appropriate prior to coefficients $\boldsymbol{\beta}$. Let us denote the prior density function as $p(\boldsymbol{\beta}|\boldsymbol{\theta})$. Then the maximizing-a-posterior (MAP) estimate of $\boldsymbol{\beta}$ is obtained by solving

$$\hat{\boldsymbol{\beta}}_{\mathrm{MAP}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \log p(\boldsymbol{\beta}|\boldsymbol{\theta}),$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the response variable, and $\boldsymbol{\theta} \in \Theta$ contains all unknown hierarchical parameters. The second item $\log p(\boldsymbol{\beta}|\boldsymbol{\theta})$ can be thought of as a penalty term added on the log-likelihood of the data $f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\theta})$ (Lee et al. 2010).

Denote $\boldsymbol{\beta}_e = \boldsymbol{\beta}_j - \boldsymbol{\beta}_k$, $e = 1, \ldots, E$, as the difference of coefficients for any $(j,k) \in \mathcal{E}$. According to Lee et al. (2010) and Jiang, Lozano, and Liu (2012), the hierarchical priors given to the coefficient difference on each pair $\boldsymbol{\beta}_e = (\beta_{e,0}, \ldots, \beta_{e,p})'$ are listed below,

$$\begin{aligned} \beta_{e,i}|\sigma_e^2 &\sim N(0, \sigma_e^2), \quad i = 0, 1, \ldots, p, \\ \sigma_e^2|\tau_e &\sim G\left(\frac{p+1}{2}, 2\tau_e^2\right), \\ \tau_e|a_e, b_e &\sim \mathrm{IG}(a_e, b_e), \end{aligned} \tag{5}$$

where $G(a,b)$ denotes the Gamma distribution and $\mathrm{IG}(a,b)$ represents the inverse Gamma distribution. The joint distribution of each coefficient difference vector $\boldsymbol{\beta}_e$ is

$$p(\boldsymbol{\beta}_e|\tau_e) = \frac{(2\tau_e)^{-p}\pi^{-(p-1)/2}}{\Gamma((p+1)/2)} \exp\left(-\frac{||\boldsymbol{\beta}_e||_2}{\tau_e}\right).$$

Besides, we have

$$\tau_e|\boldsymbol{\beta}_e^{(t)}, a_e, b_e \sim \mathrm{IG}(a_e + p, b_e + ||\boldsymbol{\beta}_e^{(t)}||_2).$$

Thus, the corresponding iterative procedure to solve for $\boldsymbol{\beta}$ is,

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} & -\log f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) \\ & + \sum_{e\in\mathcal{E}} ||\boldsymbol{\beta}_e||_2 \int \frac{1}{\tau_e} p\left(\tau_e|\boldsymbol{\beta}_e^{(t)}, a_e, b_e\right) d\tau_e, \end{aligned}$$

and it can be reformulated as

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} -\log f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{e\in\mathcal{E}} \gamma_e^{(t+1)} ||\boldsymbol{\beta}_e||_2, \tag{6}$$

where $\gamma_e^{(t+1)} = \frac{a_e + p}{||\boldsymbol{\beta}_e^{(t)}||_2 + b_e}$. This adaptive weight formula is consistent with the proposed adaptive weight structure in (4).

In addition, the prior distribution of all regression coefficients $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_N^T\right)^T$, which is a vector of length $N(p+1)$, can be defined as follows,

$$\pi(\boldsymbol{\beta}|a_1, \ldots, a_E, b_1, \ldots, b_E) \propto \prod_{(j,k)\in\mathcal{E}} p(\boldsymbol{\beta}_j, \boldsymbol{\beta}_k|a_e, b_e),$$

where a_e, b_e for $e = 1, \ldots, E$ are the hierarchical parameters at each edge e, and the distributions of edges are assumed to be conditionally independent (Liben-Nowell and Kleinberg 2007).

The hierarchical representation for the prior distribution of regression coefficient vector can be expressed as

$$\begin{aligned} &\pi(\boldsymbol{\beta}|a_1, \ldots, a_E, b_1, \ldots, b_E) \\ &\propto \int\int \prod_{e=1}^{E} (\sigma_e^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}\right) \\ &\prod_{e=1}^{E} \pi(\sigma_e^2|\tau_e) \prod_{e=1}^{E} \pi(\tau_e) \prod_{e=1}^{E} d\sigma_e^2 \prod_{e=1}^{E} d\tau_e, \end{aligned}$$

where $\Sigma_{\boldsymbol{\beta}}^{-1}$ is the $N(p+1) \times N(p+1)$ symmetric precision matrix with the following structure,

$$\Sigma_{\boldsymbol{\beta}}^{-1} = \begin{bmatrix} \sum_{j\in\mathcal{N}(1)} \frac{1}{\sigma_{(1,j)}^2} & -\frac{1}{\sigma_{(1,2)}^2} & 0 & \cdots & 0 \\ -\frac{1}{\sigma_{(2,1)}^2} & \sum_{j\in\mathcal{N}(2)} \frac{1}{\sigma_{(2,j)}^2} & 0 & \cdots & -\frac{1}{\sigma_{(2,N)}^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{1}{\sigma_{(N,2)}^2} & -\frac{1}{\sigma_{(N,3)}^2} & \cdots & \sum_{j\in\mathcal{N}(N)} \frac{1}{\sigma_{(N,j)}^2} \end{bmatrix} \\ \otimes \mathbf{1}_{p+1},$$

where $\mathcal{N}(i)$ denotes the neighbors of observation i, the subscript (i,j) denotes pair $(i,j) \in \mathcal{E}$, and $\sigma_{(i,j)}^2 = \sigma_{(j,i)}^2$. The symbol $\otimes$ represents Kronecker product. The off-diagonal element $-\frac{1}{\sigma_{(i,j)}^2}$ is nonzero if and only if $(i,j) \in \mathcal{E}$, that is, observations i and j are paired.

3.3. Consistency Properties of Clustering and Parameter Estimation

In this section, we study the asymptotic properties of the adaptive convex clustering in the context of GLMs. More precisely, we assume that the number of covariates p is fixed and the number of observation N grows to infinity. Some notation and assumptions need to be introduced before stating the results.

First, we assume that the model has a clustering representation with $1 \leqslant K \ll N$ number of actual clusters in the whole dataset $\left\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, N\right\}$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the covariate vector and y_i is the corresponding response variable. Denote $\boldsymbol{y} = [y_1, \ldots, y_n]^T$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$, $\boldsymbol{z}_i = \left(1, \boldsymbol{x}_i^T\right)^T$ and $\boldsymbol{\beta}_i \in \mathbb{R}^{p+1}$ the corresponding coefficient vector at observation i. Without loss of generality, let $\mathcal{A} = \{(j,k) : \boldsymbol{\beta}_j^* \neq \boldsymbol{\beta}_k^*\}$ be the set of unique pairs of coefficient vectors, and its cardinality is $|\mathcal{A}| = K(K-1)/2$. Then $\boldsymbol{\beta}_{\mathcal{A}} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T\right)^T$ is the collection of all unique coefficients (sample-wise) where $\boldsymbol{\beta}_k \in \mathbb{R}^{p+1}, k = 1, \ldots, K$ are the unique coefficient vectors for each cluster. Here $\boldsymbol{\beta}_{\mathcal{A}}^* = \left(\boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_K^{*T}\right)^T$ denote the true model parameters and n_k as the number of observations in each cluster. The corresponding design matrix $\boldsymbol{Z}^* = \operatorname{diag}\left(\boldsymbol{Z}_1^*, \ldots, \boldsymbol{Z}_K^*\right)$ is a N by $K(p+1)$ block-diagonal matrix, whose blocks are of size $n_k \times (p+1)$ and defined as $\boldsymbol{Z}_k^* = [\mathbf{1} \quad \boldsymbol{X}_k^*]$. Here $\boldsymbol{X}_k^*$ stands for the $n_k \times p$ sub-matrix of $\boldsymbol{X}^*$ with rows corresponding to the observations falling in cluster k. Thus, the vector of all response variables is arranged into $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T\right)^T$, where $\boldsymbol{y}_k := \{y_i, c_i = k\}$ with c_i as the cluster indicator.

In GLMs, assume that each response y_i comes from a distribution in the exponential family with canonical parameter ξ_i and dispersion parameter ϕ. The corresponding probability density function can be written as

$$f\left(y_i\right) = \exp\left\{\frac{y_i \xi_i - b\left(\xi_i\right)}{a_i\left(\phi\right)} + c\left(y_i, \phi\right)\right\}, \tag{7}$$

where $a_i(\phi), b(\xi_i)$, and $c\left(y_i, \phi\right)$ are known functions. In this framework, it can be shown that,

$$\mathrm{E}\left(Y_i\right) = \mu_i = b'\left(\xi_i\right),$$

$$\operatorname{var}\left(Y_i\right) = \sigma_i^2 = b''\left(\xi_i\right) a\left(\phi\right) = \phi b''\left(\xi_i\right)/q_i,$$

where $b'\left(\xi_i\right)$ and $b''\left(\xi_i\right)$ are the first and second derivatives of $b\left(\xi_i\right)$.

Next, we assume for $s_i = k$, the expected value of Y_i, μ_i is a linear function of the independent variables, $\boldsymbol{x}_i$, such that

$$\mathrm{E}\left(Y_i\right) = \mu_i = g^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_k^*\right),$$

or,

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_k^*,$$

for $s_i = k$, where g is the link function, and the quantity η_i is called the linear predictor which is same as the canonical parameter ξ_i.

Without loss of generality, the adaptive convex clustering estimates $\hat{\boldsymbol{\beta}}^{*(N)}$ are given by

$$\hat{\boldsymbol{\beta}}^{*(N)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(-y_i(\boldsymbol{x}_i^T \boldsymbol{\beta}_i) + b\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_i\right)\right) \tag{8}$$

$$+ \lambda_N \sum_{(j,k) \in \mathcal{E}} \gamma_{(j,k)} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2.$$

Accordingly, we write the optimal Fisher information matrix,

$$\mathbf{I}\left(\boldsymbol{\beta}_{\mathcal{A}}^*\right) = \begin{bmatrix} \mathbf{I}(\boldsymbol{\beta}_1^*) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}(\boldsymbol{\beta}_K^*) \end{bmatrix},$$

where $\mathbf{I}(\boldsymbol{\beta}_k^*)$ is the Fisher information matrix related to the kth true clustered model.

Theorem 1 shows that the adaptive convex clustering estimates $\hat{\boldsymbol{\beta}}^{*(N)}$ enjoy the asymptotic properties if λ_N is chosen appropriately under some mild regularity conditions. The detailed proof can be found in the supplementary materials.

Theorem 1. Let $\mathcal{A}_N^* = \{(j,k) : \hat{\boldsymbol{\beta}}_j^{*(N)} \neq \hat{\boldsymbol{\beta}}_k^{*(N)}\}$. Suppose that $\frac{\lambda_N}{\sqrt{N}} \rightarrow 0$ and $\lambda_N \rightarrow \infty$; then under some mild regularity conditions, the adaptive convex clustering estimator $\hat{\boldsymbol{\beta}}^{*(N)}$ must satisfy the following properties:

1. Consistency in clustering: $\lim_{n \to \infty} P\left(\mathcal{A}_N^* = \mathcal{A}\right) = 1$;
2. Asymptotic normality: $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(N)} - \boldsymbol{\beta}_{\mathcal{A}}^*) \rightarrow_d N(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta}_{\mathcal{A}}^*)^{-1})$, as $n \to \infty$,

where n is the sample size of the smallest cluster.

From Theorem 1, it is seen that the proposed adaptive convex clustering can be asymptotically consistent in both clustering and parameter estimation.

4. The Adaptive IWLS-Based Algorithm

Several algorithms have been developed in the literature to solve the convex clustering problems (Zhu et al. 2014; Chen et al. 2015; Lange and Keys 2015). However, the general convex optimization methods may not work very well for the problem where p, N, and E are potentially large. Hallac, Leskovec, and Boyd (2015) developed an algorithm based on the ADMM (Boyd et al. 2011; Parikh and Boyd 2014; Chen et al. 2015) to solve the convex clustering problem efficiently. Note that the ADMM-based algorithm searches the optimal solutions by going through every pair of unknown parameters in the regularization term. Such an algorithm may get stuck in local optimum especially when the objective function is nonlinear. We refer to Boyd et al. (2011) for the detailed description about the ADMM method.

To overcome this challenge, we develop an IWLS based algorithm for the adaptive convex clustering of GLMs. The IWLS-based algorithm is commonly used for GLMs with promising performance. Here we adopt the idea of IWLS for parameter estimation in the adaptive convex clustering of GLMs. The key idea of the IWLS is a Newton–Raphson approach, which uses a second-order Taylor expansion for the log-likelihood function of the GLMs. It means that in each iteration, the log-likelihood function is approximated by a quadratic function. Specifically, let us denote $\hat{\boldsymbol{\beta}}^{(t)}$ the estimate at iteration t, and we can calculate $\hat{\eta}_i^{(t)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^{(t)}$, and $\hat{\mu}_i^{(t)} = g^{-1}\left(\hat{\eta}_i^{(t)}\right)$. At iteration $(t+1)$, the log-likelihood function at $(\boldsymbol{x}_i, y_i)$, denoted as $f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i)$, is approximated by

$$f_i^{(t+1)} = \hat{w}_i^{(t)} (\hat{z}_i^{(t)} - \boldsymbol{x}_i^T \boldsymbol{\beta}_i)^2, \tag{9}$$

where

$$\hat{z}_i^{(t)} = \hat{\eta}_i^{(t)} + \left(y_i - \hat{\mu}_i^{(t)}\right) \frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\Big|_{\hat{\mu}_i^{(t)}} \quad \text{and}$$

$$\hat{w}_i^{(t)} = q_i \Big/ \left[b''\left(\theta_i^{(t)}\right) \left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\Big|_{\hat{\mu}_i^{(t)}}\right)^2 \right].$$

Details of IWLS based approximation for logistic regression can be found in the supplemental materials. Then the approximated objective function of (4) becomes

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{N} \hat{w}_i^{(t)} (\hat{z}_i^{(t)} - \boldsymbol{x}_i^T \boldsymbol{\beta}_i)^2 + \lambda \sum_{(j,k)\in\mathcal{E}} \gamma_{jk}^{(t+1)} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2, \tag{10}$$

which can be solved efficiently by using the CVXPY (Diamond and Boyd 2016), a conventional convex optimization tool. In case of large datasets, the CVXPY can be paralleled easily. The iterative procedure is repeated until changes in $\hat{\boldsymbol{\beta}}^{(t+1)}$ are sufficiently small.

The proposed IWLS-based algorithm iteratively updates the unknown parameters $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N$ simultaneously by taking advantage of the second-order derivative (i.e., Hessian matrix). While the ADMM-based algorithm updates the parameters pair-by-pair in a similar fashion as the coordinate descent methods. Thus, it is expected that the proposed IWLS-based method will converge to the optimum more efficiently than the ADMM-based algorithm. But it would be difficult to surely guarantee the convergence of global optimum in practice because the IWLS-based method may encounter the singularity (or near singularity) of the Hessian matrix. More detailed numerical comparisons between the proposed IWLS-based algorithm and the ADMM-based algorithm will be shown in Section 5. Note that the objective function in the proposed adaptive convex clustering will change during the optimization iteration t because of the updating of the adaptive weights $\gamma_{jk}^{(t)}$. Particularly, the weights for zero-coefficient differences get inflated (to infinity), whereas the weights for nonzero-coefficient differences converge to a constant value. Thus, the objective function remains convex, making the estimation of coefficients converge to the estimates in each estimated segment.

5. Simulation Study

To assess the performance of the proposed method, we conducted a set of simulation studies. In particular, Section 5.1 aims to verify the shrinkage problem of the conventional convex clustering. We first compare the performance of the conventional convex clustering using ADMM and IWLS algorithms under the scenario where the true clusters are totally separated from each other. Next, we will illustrate how the shrinkage problem arises in the conventional convex clustering when the clusters are not well separated.

In Section 5.2, we focus on the comparison of the proposed approach with several existing methods under a more complex example where four clusters are adjacent to each other. Without loss of generality, we take logistic regression model as an example of GLMs in the numerical study.

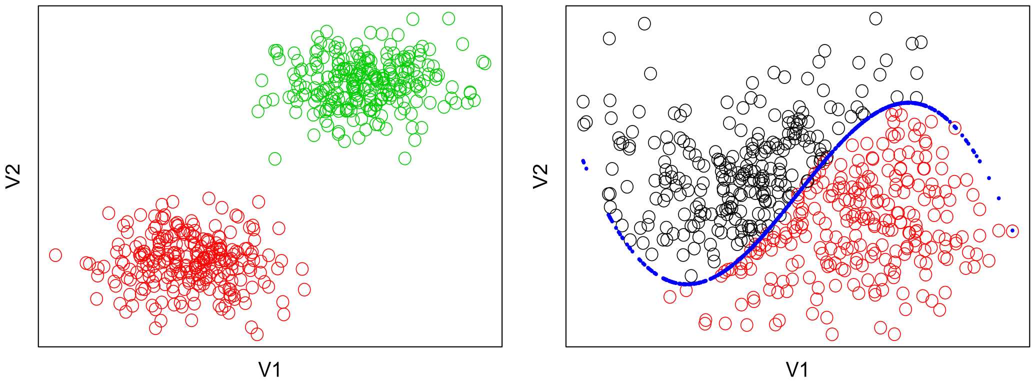
The simulated data contain N observations with binary responses at each observation $y_i \in \{-1, 1\}$. Specifically, three continuous features, U_1, U_2, U_3 are simulated for each observation, where U_1 and U_2 are used as clustering features, $\boldsymbol{v}_j = (u_{1j}, u_{2j})^T$, to calculate the prespecified weight w_{jk}. The set of observation pairs $\mathcal{E}$ is formed by connecting five nearest neighbors to each observation (Hallac, Leskovec, and Boyd 2015). The Euclidean distance calculated by $\boldsymbol{v}_j$ is used to define the nearest neighbors. Thus, each observation has at least five connected observations. For simplicity, only $X_1 = U_3$ is included in the logistic regression model in the simulation study.

For each simulation, data are randomly split into training (80%) and testing (20%) datasets. Then cross-validation is conducted for the training data to find the optimal λ value which maximizes the prediction accuracy based on AUC, the area under the receiver operating characteristic (ROC) curve. The cutoff point in the logistic regression is chosen in the way that the sum of sensitivity and specificity is maximized. In addition, each simulation setting was repeated for 50 times and the corresponding mean and standard deviation of resulting statistics are reported.

5.1. Evaluation on Shrinkage Problem

In this section, two scenarios as listed below are generated with the number of true clusters $K = 2$: (D1) the clusters are fully separated and (D2) the clusters are adjacent.

(D1) Separated clusters with $K = 2$: $N = 500$ observations of (U_1, U_2) are simulated as in Figure 2(D1). The two clusters



(D1) Separated, $K = 2$

(D2) Adjacent, $K = 2$

Figure 2. Illustration of simulated data under scenarios (D1) and (D2).

are clearly separated from each other. Particularly, 250 data points in one cluster are generated from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and 250 data points in another cluster are generated from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Here $\boldsymbol{\mu}_1 = (1.2, 3.3)^T$, $\boldsymbol{\mu}_2 = (1.5, 4.2)^T$, and $\boldsymbol{\Sigma} = \text{diag}(0.01, 0.02)$.

(D2) Adjacent clusters with $K = 2$: $N = 500$ observations of $\boldsymbol{v}_i = (u_{1i}, u_{2i})^T$ are first simulated from two multivariate normal distributions: 250 data points from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and 250 data points from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = (4.2, 5.9)^T$, $\boldsymbol{\Sigma}_1$ is a correlation matrix with off-diagonal value to be 0.05, and $\boldsymbol{\mu}_2 = (3.5, 5.6)^T$ and $\boldsymbol{\Sigma}_2$ a correlation matrix with off-diagonal value to be 0.1. Secondly, the data are split into two adjacent clusters by a nonlinear function, $f_1(u) = \sin(u^*) - 1.5u^3$, where u^* is the normalized u. The corresponding blue curve is shown in Figure 2(D2).

The shrinkage problem is not very severe in (D1) where the clusters are clearly separated. The setting of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is designed to make the two clusters well separated with the variance in $\boldsymbol{\Sigma}$ being relatively small. In (D2), we set the values of $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$ such that data from the two clusters are not well separated at the boundary. Then the segmentation and estimation are relatively difficult, especially for the data near the boundary between two clusters. The reason is that, the data around the boundary might be considered as "close" neighbors in terms of $\boldsymbol{v}_i$ values even though they actually belong to different clusters. Thus, the pairs are wrongly connected and the fixed penalty weights push their corresponding coefficients toward zero. This leads to the shrinkage problem as discussed in Section 3.

Note that in each scenario, the values of $X_1 = U_3$ are simulated from $\mathcal{N}(1.0, 0.1)$ and normalized. A unique set of true coefficients, $\boldsymbol{\beta_k} = [\beta_{k0}, \beta_{k1}]^T, k = 1, \ldots, K$, is assigned to each cluster obtained previously. The true values can be found in Tables 2 and 3. Finally, the binary responses $\boldsymbol{y}$ are simulated based on the predictor variable X_1 as well as the corresponding true coefficients $\boldsymbol{\beta}_k$.

The models compared in this section are the conventional convex clustering model in (1) with two different algorithms ADMM and IWLS, respectively. We denoted these two methods as CC-ADMM and CC-IWLS.

Table 2 shows the mean estimation results under scenario (D1). The corresponding standard errors are reported in parentheses as well. The CC-ADMM(GLM) and CC-IWLS(GLM) are the results obtained from refitting the logistic regression models under each cluster estimated from the CC-ADMM and CC-IWLS, respectively. The computation time listed is the average time per iteration per λ value.

Note that the simulated data in (D1) are clearly separated by (X_1, X_2) as shown in Figure 2(D1). By connecting the nearest five neighbors to each observation, all observation pairs belong to the same true cluster. Thus, there are no shrinkage problems under this scenario, and both CC-ADMM and CC-IWLS are able to find the true clusters. Therefore, the two GLM-refitted results are exactly the same in Table 2. However, the CC-ADMM does not find the optimal estimate of $\boldsymbol{\beta}_k$, while CC-IWLS gives almost the same coefficient estimates as those obtained from the CC-IWLS(GLM). This indicates that the ADMM algorithm failed to find the global optimum. In terms of computation time, the CC-IWLS takes about two thirds of the time compared with the CC-ADMM. To summarize, the CC-IWLS results in more accurate estimation and is computationally more efficient.

Table 3 shows the estimation results under scenario (D2). Since there are two adjacent clusters, the segmentation and estimation are relatively difficult compared with (D1), especially for the data near the boundary. Based on the results of GLM refitting, it is seen that the CC-IWLS produced more accurate clustering results than the CC-ADMM. Due to the inaccuracy in cluster assignment, the estimates of CC-ADMM(GLM) are far

Table 2. Coefficient estimation for separated clusters with $K = 2$ (D1).

Segment		True	CC-ADMM(GLM)	CC-ADMM	CC-IWLS(GLM)	CC-IWLS
1	β_{01}	−1	−1.020 (0.249)	−0.745 (0.185)	−1.020 (0.249)	−1.047 (0.220)
	β_{11}	2.5	2.530 (0.353)	1.883 (0.208)	2.530 (0.353)	2.534 (0.387)
2	β_{02}	1.5	1.564 (0.364)	0.869 (0.375)	1.564 (0.364)	1.52 (0.285)
	β_{12}	−3.5	−3.611 (0.621)	−2.024 (0.709)	−3.611 (0.621)	−3.51 (0.586)
Computation time (min)			–	2.22 (0.97)	–	1.51 (0.214)

Table 3. Coefficient estimation (based on 50 iterations) for adjacent clusters with $K = 2$ (D2).

Segment		True	CC-ADMM(GLM)	CC-ADMM	CC-IWLS(GLM)	CC-IWLS
1	β_{01}	−1	−2.572 (17.54)	−0.398 (0.878)	−0.997 (0.437)	−0.410 (0.236)
	β_{11}	2.5	0.310 (41.12)	1.151 (1.742)	2.606 (0.572)	1.182 (0.329)
2	β_{02}	1.5	11.761 (28.60)	0.942 (0.524)	1.569 (0.580)	0.634 (0.230)
	β_{12}	−3.5	−29.358 (78.24)	−1.922 (0.947)	−3.739 (1.378)	−1.269 (0.477)

away from the truth, and their corresponding standard errors are inflated. Besides, the coefficient estimates of CC-ADMM have relatively larger standard errors as well. It indicates that the ADMM algorithm may not converge to the global optimum. In addition, comparing CC-ADMM and CC-IWLS with their corresponding GLM refits, both algorithms consistently give smaller absolute coefficient values.

The results shown in both (D1) and (D2) scenarios indicate that the CC-ADMM encounters both shrinkage and convergence issues for the logistic objective function.

5.2. Methods Comparison

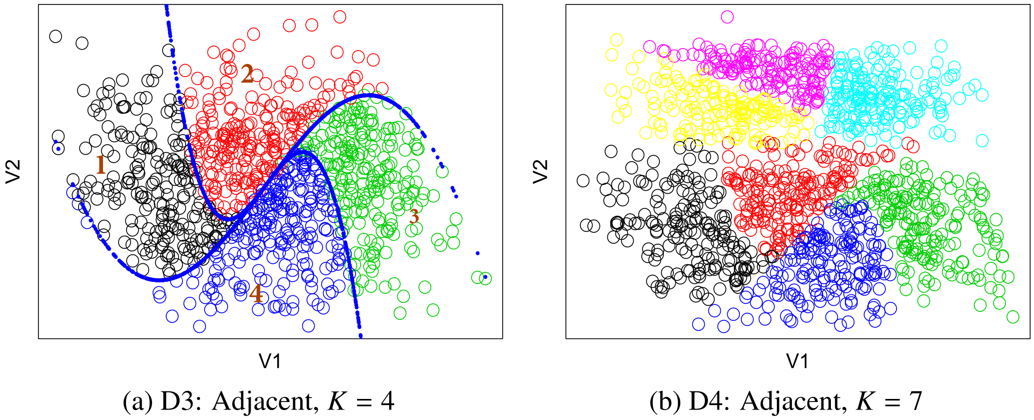
In this section, we first simulate a more complex scenario with $K = 4$ true clusters (D3) as follows.

(D3) Adjacent clusters with $K = 4$: $N = 900$ observations of (U_1, U_2) are simulated shown in Figure 3. Particularly, $U_1 \sim \mathcal{N}(0, 0.4)$, and $U_2 \sim \mathcal{N}(0, 0.5)$. The data are split into four adjacent clusters by two nonlinear functions, $h_1(u) = \sin(u^*) - 1.5(u^*)^3$ and $h_2(u) = \sin(u^*) - 1.5(u^*)^3$, where u^* is the normalization of u. Besides, X_3 is simulated the same way as in (D1) and (D2).

In this study, we compare our proposed approach, denoted as ACC-IWLS, with the existing methods, CC-ADMM, the global model, and the k-means method listed below. We also include the *Optimal Model* as a benchmark, which corresponds to the best result achievable under the assumption that the true segmentation is known.

Global model: Fit one logistic regression model.
k-means method: Cluster data by k-means according to U_1, U_2 and then fit logistic regression under each cluster.
ACC-IWLS: Fit adaptive convex clustering model using the IWLS-based algorithm.
Optimal model: Fit logistic regression model for each true segment.

Note that both the CC-ADMM method and our proposed methods can end up with more than four clusters. To evaluate the accuracy of estimated coefficients for the individual four clusters, we marked them as four segments "1," "2," "3," and "4" as shown in Figure 3(a). In Table 4, we compare the coefficients of estimated clusters that have majority of data points falling into these four segments with the true coefficient values. The results in Table 4 indicate that the coefficient estimation from the proposed ACC-IWLS method is more accurate than that of the CC-ADMM. Besides, comparing both algorithms with their GLM refits, it is clear that the CC-ADMM shows much more serious shrinkage problem than the ACC-IWLS.



(a) D3: Adjacent, $K = 4$ (b) D4: Adjacent, $K = 7$

Figure 3. Illustration of simulated data under scenario (D3) and (D4).

Table 4. Coefficients estimation results for adjacent clusters with $K = 4$ (D3).

Segment		True	CC-ADMM(GLM)	CC-ADMM	ACC-IWLS(GLM)	ACC-IWLS
1	β_{01}	−1	−1.176 (0.749)	−0.533 (0.269)	−1.117 (0.654)	−0.998 (0.530)
	β_{11}	2.5	2.966 (1.060)	1.218 (0.324)	2.628 (0.970)	2.272 (0.863)
2	β_{02}	1.5	1.395 (0.639)	0.509 (0.212)	1.419 (0.975)	1.220 (0.574)
	β_{12}	−3.5	−3.617 (1.386)	−1.258 (0.252)	−3.359 (1.817)	−2.452 (0.974)
3	β_{03}	0.5	0.593 (0.433)	0.406 (0.196)	0.666 (0.406)	0.688 (0.231)
	β_{13}	1.5	1.951 (0.769)	0.884 (0.292)	1.761 (1.193)	1.308 (0.548)
4	β_{04}	−0.5	−0.615 (0.550)	−0.271 (0.176)	−0.554 (0.541)	−0.777 (0.361)
	β_{14}	−1.5	−2.023 (1.172)	−1.081 (0.266)	−1.652 (0.719)	−1.170 (0.436)

Furthermore, we evaluate and the performance of these methods via several criteria listed in Table 5. The F_1 score is defined as

$$F_1 = \frac{2PR}{P+R},$$

where $P = tp/(tp + fp)$ is the precision, and $R = tp/(tp + fn)$ is the recall or sensitivity. The larger F_1 score implies higher clustering accuracy. Here tp is the true positive, fp is the false positive, tn is the true negative, and fn is the false negative, respectively. The Frobenius norm measures the estimation accuracy of coefficients, and is defined as

$$\text{Fnorm} = \sqrt{\sum_{i=1}^{n}\sum_{j=0}^{p}\left(\hat{\beta}_{ij} - \beta_{ij}\right)^2},$$

where n is the number of testing observations. In addition, the AUC is the area under the ROC curve, which can be used to evaluate how well the model is distinguishing between classes.

The result shows that proposed ACC-IWLS generally outperforms other methods with the lowest classification error and Fnorm, and the highest AUC value. The global model gives the worst performance. The classification errors of both CC-ADMM and ACC-IWLS are no better than the global model, which might be due to the shrinkage problem in the data. Besides, the ACC-IWLS performs better than the CC-ADMM across all criteria except recall. A possible reason why the CC-ADMM has a high recall score is that the CC-ADMM tends to predict positives for almost all the observations, which leads to a very low precision. In conclusion, the proposed adaptive convex clustering model with IWLS based algorithm (ACC-IWLS) produces a desirable result on data segmentation as well as model prediction.

To further evaluate the performance of the proposed method in the case of a larger number of clusters, we consider a simulation study with the number of clusters $K = 7$.

(D4) Adjacent clusters with $K = 7$: $N = 1250$ observations of (U_1, U_2) are simulated in a similar way as the setting in (D3). The data are split into seven adjacent clusters as shown in Figure 3. Besides, X_3 is simulated in the same way as in (D3).

The performance of different methods are reported in Table 6. All results are based on 50 iterations except the CC-ADMM method, which is based on 15 iterations. It is worth remarking that the computation of the CC-ADMM method in this case is very slow with more than 20 hr per run on a standard Mackbook Pro laptop with 4GB RAM. From the results in Table 6, it is seen that the proposed ACC-IWLS outperforms global and k-means significantly, especially for classification error, Fnorm, and AUC. Both CC-ADMM and ACC-IWLS have comparable performance in terms of Fnorm and AUC. However, the proposed ACC-IWLS gives better classification error, F_1 and recall values, which can be explained by the fact that our proposed method can address the coefficient shrinkage problem. In general, our proposed method can be applied to the data with relatively larger number of clusters and preserves its merits over other methods.

6. IT Service Pricing Data Application

In this section, we apply the methodology to the historical pricing data from a major IT service provider. Specifically, we evaluate two datasets corresponding to the software brands of analytics and security, for simplicity, denoted as Brand1 and Brand2, respectively, throughout the rest of this article. The adaptive convex clustering model will be applied to the data of each brand independently. The goal is to cluster and fit logistic

Table 5. Performance comparison for adjacent clusters with $K = 4$ (D3): The classification error and Fnorm are smaller the better, while AUC, F_1, precision, and recall are larger the better.

Models	Classification error	Fnorm	AUC	F_1	Precision	Recall
CC-ADMM	0.41 (0.041)	57.46 (5.139)	0.75 (0.042)	0.59 (0.059)	0.59 (0.055)	0.62 (0.112)
Global	0.46 (0.038)	70.14 (0.898)	0.54 (0.044)	0.59 (0.120)	0.54 (0.082)	0.70 (0.203)
k-means	0.43 (0.035)	67.08 (1.403)	0.61 (0.040)	0.55 (0.104)	0.59 (0.067)	0.56 (1.969)
ACC-IWLS (proposed)	0.31 (0.04)	53.46 (12.02)	0.75 (0.046)	0.68 (0.060)	0.70 (0.048)	0.69 (0.125)
Optimal	0.22 (0.036)	12.12 (5.509)	0.87 (0.029)	0.78 (0.045)	0.79 (0.048)	0.78 (0.082)

Table 6. Performance comparison results for adjacent clusters with $K = 7$ (D4).

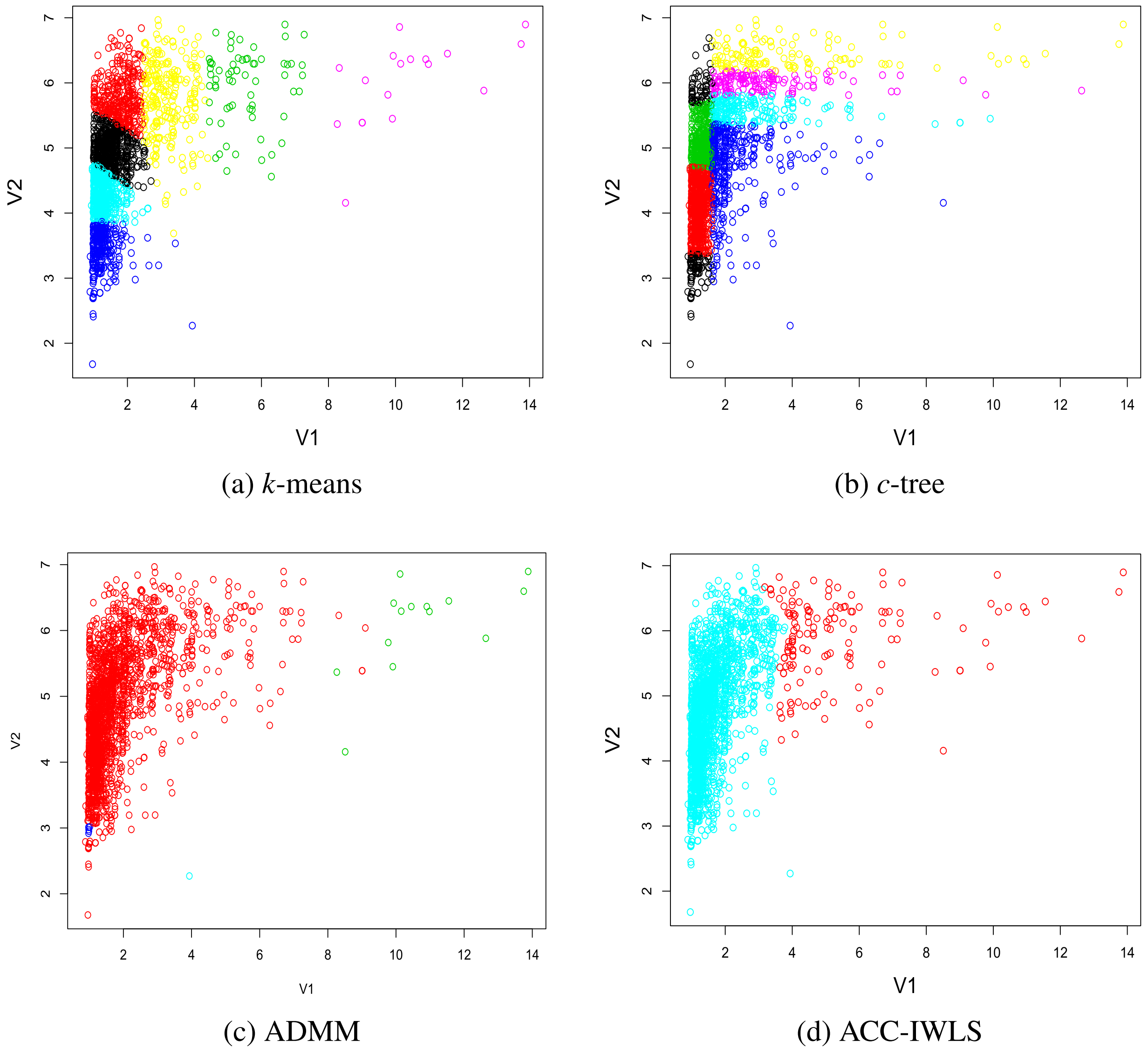
Models	Classification error	Fnorm	AUC	F_1	Precision	Recall
CC-ADMM	0.35 (0.040)	63.30 (9.53)	0.81 (0.041)	0.60 (0.064)	0.63 (0.082)	0.59 (0.106)
Global	0.45 (0.040)	97.56 (0.773)	0.50 (0.046)	0.25 (0.130)	0.61 (0.154)	0.19 (0.152)
k-means	0.35 (0.043)	77.58 (6.106)	0.72 (0.050)	0.64 (0.055)	0.61 (0.095)	0.67 (0.067)
ACC-IWLS (proposed)	0.32 (0.05)	62.10 (15.01)	0.81 (0.065)	0.66 (0.056)	0.64 (0.070)	0.69 (0.088)
Optimal	0.19 (0.019)	18.58 (5.066)	0.90 (0.018)	0.79 (0.022)	0.79 (0.042)	0.80 (0.042)

regression models to predict the purchase likelihood for the quotes in each brand, where the RFQs sharing similar purchase behavior are expected to be clustered together.

Brand1 contains $N_1 = 2682$ observations, whereas Brand2 contains $N_2 = 2642$ observations. Each observation i corresponds to a unique RFQ. Three features considered here are grade (U_1 = entitled price/value score), deal-size (U_2 = log(total entitled price)), and normalized price (U_3 = quote price/value score) (Xue, Wang, and Ettl 2015). The three features are included in the logistic regression model to predict the purchase likelihood. The U_1 and U_2 are chosen as the clustering features according to domain expert's knowledge. To form the set of pairs $\mathcal{E}$, each RFQ is connected to its five nearest neighbors in terms of the Euclidean distance of $\boldsymbol{v}_i = (u_{1i}, u_{2i})'$. The binary response ($Y_i$) at each node is coded as 1/−1 indicating whether or not the corresponding client made a purchase.

For each brand, the RFQs are randomly split into the 80% training and 20% testing datasets. The 5-fold cross-validation is conducted on each training dataset to choose the optimal value of λ as well as the cutoff point in the logistic regression. Specifically, λ is chosen to maximize the median AUC values in 5-fold cross-validation, while the cutoff point is chosen to maximize the sum of sensitivity and specificity. The optimal values will be directly used in the testing data. We compare the proposed approach ACC-IWLS with the global model, the k-means method, and the c-tree method implemented in Xue, Wang, and Ettl (2015).

The comparison of four approaches would provide important insights into the modeling performance. The global model assumes the purchase behavior of all RFQs can be captured in a uniform pattern, in which all RFQs are fit into a single logistic regression model without any segmentation. In contrast, both the k-means method and c-tree method have separate steps of segmentation and regression. Specifically, the k-means method uses an unsupervised learning approach to cluster RFQs based on quote features. However, the c-tree method first conducts tree regression classification considering the discount off the entitled price as a measure of the sellers' pricing strategy in response to RFQ features. After the segmentation, logistic regression models are fit for each segment. To make a fair comparison, the number of clusters used in the k-means method is the same as that obtained from the c-tree method. For the proposed ACC-IWLS,



(a) k-means

(b) c-tree

(c) ADMM

(d) ACC-IWLS

Figure 4. Comparison of clustering results for Brand1.

it integrates the clustering and regression in a single step to learn the buyer's response to the pricing decision and quote features.

The performance of all four models are compared in Figures 4 and 5, Tables 7 and 8 for Brand1 and Brand2, respectively.

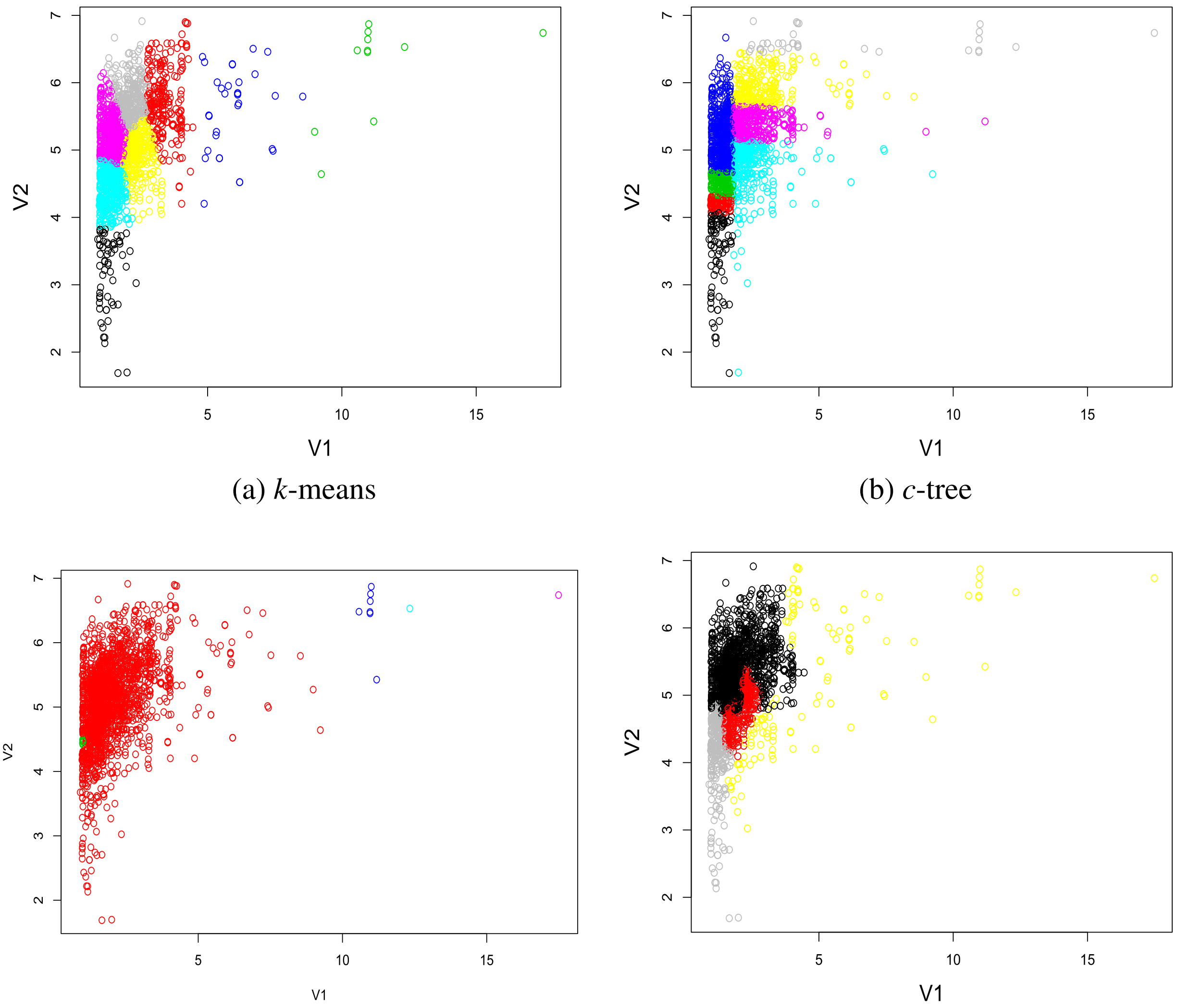
In the case of Brand1, the proposed adaptive convex clustering only results in two segments. The prediction performance in Table 7 indicates that the proposed method is comparable to the global method and performs relatively better than the *k*-means, the *c*-tree, and the CC-ADMM method with a higher value of AUC. It seems that segmentation will not obviously improve the prediction accuracy of Brand1 compared with the global model. One possible explanation is that the data in Brand1 may not have segmentation patterns just based on the current features since there is a very limited number of variables available to describe the bundling features in the complex configuration. We also find the *c*-tree model is unable to outperform the global model, which means a supervised learning based on the seller's pricing strategy does not necessarily improve the fit of logistics regression. Moreover, the *k*-mean method makes the performance even worse compared to the global method, indicating that the unsupervised learning may result in poor segmentation and deteriorate the prediction of purchase behavior. We also

Table 7. Modeling performance comparison for Brand1: The classification error is the smaller the better, while the F_1, AUC, precision, and recall are the larger the better.

	Classification error	F_1	AUC	Precision	Recall
Global	0.40	0.53	0.653	0.47	0.62
k-means	0.44	0.54	0.640	0.44	0.71
c-tree	0.41	0.56	0.640	0.47	0.69
CC-ADMM	0.42	0.54	0.635	046	0.65
ACC-IWLS (proposed)	0.40	0.55	0.659	0.47	0.67

Table 8. Modeling performance comparison for Brand2: The classification error is the smaller the better, while the F_1, AUC, precision, and recall are the larger the better.

	Classification error	*F*-score	AUC	Precision	Recall
Global	0.40	0.41	0.614	0.32	0.57
k-means	0.42	0.43	0.622	0.32	0.63
c-tree	0.39	0.37	0.606	0.31	0.47
CC-ADMM	0.75	0.40	0.541	0.25	1.00
ACC-IWLS (proposed)	0.32	0.39	0.659	0.37	0.41



(a) *k*-means

(b) *c*-tree

(c) ADMM

(d) ACC-IWLS

Figure 5. Comparison of clustering results for Brand2.

notice that the CC-ADMM does not provide better performance than the proposed method.

It is worth pointing out that currently U_1 and U_2 are used as clustering features. The Normalized Price U_3 contains the information of quote pricing decision, and it cannot be used as a clustering feature since the embedded quote price is unknown (Xue, Wang, and Ettl 2015). Please see the supplemental materials for detailed explanations. To further evaluate the proposed ACC model performance for data in Brand1, we have run one more experiment (named as Exp-A) using the Normalized Price U_3 as a clustering feature. That is, the clustering features are changed to U_1 and U_3, and all others remain the same setting. The corresponding performance for this Exp-A case is reported in the supplemental materials. Compared with other models, it is seen that the proposed method for the Exp-A case provides better performance in terms of classification error, AUC, and precision. It indicates that if more informative features are available for use, our proposed model can gain some advantage for analyzing data in Brand1.

For Brand2, the proposed method shows significant improvement in comparison with the global model and the *k*-mean method especially in terms of the classification error and AUC, while the CC-ADMM method has the worst performance. In this case, our proposed model performance based segmentation shows its advantage in simultaneously detecting the segments and fitting the regression model. On the other hand, the *c*-tree method only shows a minor improvement in terms of classification error, which means the seller's price differentiation does not fully reflect the heterogeneity in customer's purchase behavior. It is worth pointing out that the CC-ADMM method classifies all labels to be 1's in the test data, resulting in poor classification error and very high recall value. As shown in Figure 5(c), the CC-ADMM method does not provide a meaningful segmentation result.

As a brief summary, we clearly observe the disconnection between segmentation and regression in the two-step methods: the *k*-mean method and the *c*-tree method. In particular, an unsupervised learning method, such as the *k*-means method, is very likely to give a blinded segmentation and thereby deteriorate the modeling fitting. Also, a supervised learning method based on seller's pricing strategy does not show any remarkable improvement. It seems the seller is unable to fully capture the heterogeneity in customers' purchase behavior during the pricing of personalized offerings. On the contrary, a one-step model, the proposed ACC-IWLS, performs the best and is capable of learning hidden patterns of customer behavior in a business setting as complex as personalized configurations.

7. Discussion

This work considers the convex clustering problem where the different subsets of the data have heterogeneous model structures. The proposed adaptive convex clustering model with IWLS-based algorithm (ACC-IWLS) addresses the potential shrinkage problem of model coefficients estimation in the conventional convex clustering. It provides accurate results of both data segmentation and model estimation with an efficient algorithm. In particular, the weights in the penalty term of the proposed method are iteratively reweighted based on the similarity of their corresponding model coefficients to alleviate the shrinkage problem. The construction of adaptive convex clustering also has a meaningful interpretation from the Bayesian perspective. An IWLS-based algorithm is developed to facilitate the computation of parameter estimation. We would like to remark that due to the data availability and restrictions in business operation, the proposed method could have some limitations to include more informative features for improving the model performance.

One potential direction for future research is to extend this work to a "soft" clustering setting. That is, instead of hard clustering as presented in this article, we can fit a mixture of GLMs to heterogeneous data so that a probability or likelihood of each data point to be in each clusters is assigned. The traditional mixture models cluster data purely based on the similarity of model structures while our "soft" clustering problem aims to perform clustering according to both model structures and input features. One way to extend the mixture model to account for the similarity of the input features is to impose a prior distribution on the cluster membership. The prior distribution assumes data are more likely to be assigned to the same cluster if their input features are more similar. Compared with the current "hard" clustering, the "soft" clustering approach will provide a confidence of the clustering assignment. However, it can be more computationally expensive since it often requires the Markov chain Monte Carlo (MCMC) for computation.

Note that customer attributes are also important characteristics that can potentially impact the data segmentation as well as model fitting. Unfortunately, the available historical data only contain products' attributes. The proposed approach can be further improved if one can collect more customer or other relevant information such as seller's E-mail contents. In addition, the set of pairwise indices $\mathcal{E}$ in the penalty term is constructed mainly based on the similarity measures between each pair of observations. Such a construction would preclude grouping two observations that are far away from each other in terms of their similarity measures into the same segment. It is worth pointing out that identify overlapping clusters is still challenging, although the proposed methodology works quite well for separate and adjacent clusters. It might be useful to incorporate proper prior knowledge such as informative clustering features in finding the correct overlapping clusters.

There are other potential research directions to further enhance the proposed methodology. Note that the computation cost of the proposed method becomes expensive when the underlying number of clusters gets large. One possible solution is to investigate how the proposed method and algorithm can be adapted for parallel computing, especially for the adaptive IWLS-based algorithm. In the real business application, it is desirable to obtain a relatively balanced cluster structure among RFQs. Thus, one direction is to include the balance condition among clusters in the penalty term to avoid dominating clusters or extremely small clusters. Another open research topic is the stability of cross-validation for data under GLMs. It will be interesting to investigate how the model performance will vary under different cross-validation scenarios, and how to stabilize it if large variations exist. Finally, in practice, the field manager would like to highlight the business impact of the optimal price, and to pick up the model that has the most revenue improvement. The revenue improvement is related to an optimization problem where the objective function involves the estimated win probability in a complicated manner under various scenarios (Xue, Wang, and Ettl 2015). It will be interesting to investigate how to appropriately connect the proposed method with the decision optimization in a unified framework.

Supplementary Materials

The online supplementary materials for this article include technical proofs, algorithms, and the Python code used in the numerical study.

Acknowledgments

The authors would like to sincerely thank the editor, associate editor, and two referees for their insightful and constructive comments for helping improve the article.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [5]

Chen, G. K., Chi, E. C., Ranola, J. M. O., and Lange, K. (2015), "Convex Clustering: An Attractive Alternative to Hierarchical Clustering," *PLoS Computational Biology*, 11, e1004228. [2,5]

Chi, E. C., and Lange, K. (2015), "Splitting Methods for Convex Clustering," *Journal of Computational and Graphical Statistics*, 24, 994–1013. [2]

Diamond, S., and Boyd, S. (2016), "CVXPY: A Python-Embedded Modeling Language for Convex Optimization," *Journal of Machine Learning Research*, 17, 1–5. [6]

Hallac, D., Leskovec, J., and Boyd, S. (2015), "Network Lasso: Clustering and Optimization in Large Graphs," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 387–396. [2,5,6]

Hartigan, J. A., and Wong, M. A. (1979), "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society*, Series C, 28, 100–108. [1]

Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. (2011), "Clusterpath an Algorithm for Clustering Using Convex Fusion Penalties," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, ACM, New York, pp. 745–752. [2]

Jiang, H., Lozano, A. C., and Liu, F. (2012), "A Bayesian Markov-Switching Model for Sparse Dynamic Network Estimation," in *Proceedings of 2012 SIAM International Conference on Data Mining*, SIAM, pp. 506–515. [4]

Johnson, S. C. (1967), "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241–254. [1]

Jose, C., Goyal, P., Aggrwal, P., and Varma, M. (2013), "Local Deep Kernel Learning for Efficient Non-Linear SVM Prediction," in *International Conference on Machine Learning*, pp. 486–494. [1]

Jung, T., and Wickrama, K. (2008), "An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling," *Social and Personality Psychology Compass*, 2, 302–317. [1]

Lange, K., and Keys, K. L. (2015), "The Proximal Distance Algorithm," arXiv no. 1507.07598. [2,5]

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2010), "A Hierarchical Bayesian Framework for Constructing Sparsity-Inducing Priors," Tech. Rep., University of Oxford, UK. [4]

Liben-Nowell, D., and Kleinberg, J. (2007), "The Link-Prediction Problem for Social Networks," *Journal of the Association for Information Science and Technology*, 58, 1019–1031. [4]

Lindsten, F., Ohlsson, H., and Ljung, L. (2011), "Clustering Using Sum-of-Norms Regularization: With Application to Particle Filter Output Computation," in *Statistical Signal Processing Workshop (SSP)*, IEEE, pp. 201–204. [2]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [2,3]

Muthén, B. O. (2001), "Latent Variable Mixture Modeling," in *New Developments and Techniques in Structural Equation Modeling*, eds. G. A. Marcoulides & R. E. Schumacker, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 21–54. [1]

Oiwa, H., and Fujimaki, R. (2014), "Partition-Wise Linear Models," in *Advances in Neural Information Processing Systems*, pp. 3527–3535. [1]

Parikh, N., and Boyd, S. P. (2014), "Proximal Algorithms," *Foundations and Trends in Optimization*, 1, 127–239. [5]

Qiu, P. (2011), "Jump Regression Analysis," in *International Encyclopedia of Statistical Science*, eds. M. Lovric, Berlin, Heidelberg: Springer, pp. 702–704. [1]

Radchenko, P., and Mukherjee, G. (2017), "Convex Clustering via ℓ_1 Fusion Penalization," *Journal of the Royal Statistical Society*, Series B, 79, 1527–1546. [2]

Tan, K. M., and Witten, D. (2015), "Statistical Properties of Convex Clustering," *Electronic Journal of Statistics*, 9, 2324. [2]

Wang, J., and Saligrama, V. (2012), "Local Supervised Learning Through Space Partitioning," in *Advances in Neural Information Processing Systems*, eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, pp. 91–99. [1]

Wang, M., and Wang, X. (2014), "Adaptive Lasso Estimators for Ultrahigh Dimensional Generalized Linear Models," *Statistics & Probability Letters*, 89, 41–50. [2]

Wang, Q., Gong, P., Chang, S., Huang, T. S., and Zhou, J. (2016), "Robust Convex Clustering Analysis," in *IEEE 16th International Conference on Data Mining (ICDM)*, IEEE, pp. 1263–1268. [2]

Xue, Z., Wang, Z., and Ettl, M. (2015), "Pricing Personalized Bundles: A New Approach and an Empirical Study," *Manufacturing & Service Operations Management*, 18, 51–68. [1,10,12,13]

Zhu, C., Xu, H., Leng, C., and Yan, S. (2014), "Convex Optimization Procedure for Clustering: Theoretical Revisit," in *Advances in Neural Information Processing Systems*, pp. 1619–1627. [5]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [2,3]