Convex Clustering of Generalized Linear Models with Application on Purchase Likelihood Prediction

Xinwei Deng

xdeng@vt.edu

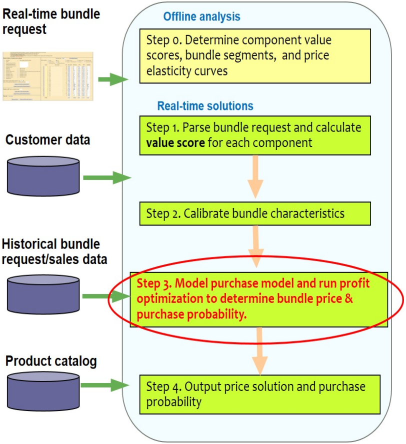
Department of Statistics, Virginia Tech

Outline

- Background & Motivation
- Brief Review of Current Method
- Proposed Approach: Adaptive Convex Clustering
- Simulation Study
- Application on IT Service Data
- Summary

Background & Motivation

- Customers can construct a personalized bundle and send a Request-For-Quote (RFQ) to the seller.
- The seller needs to determine an optimal price for each RFQ.
- Overall Goal: $\max_p (p - c(\boldsymbol{d})) \times q(\boldsymbol{d}, p)$.
  - $\boldsymbol{d}$: bundle features
  - p: quoted price of an order $\boldsymbol{d}$
  - $q(\cdot)$: purchase probability of an order $\boldsymbol{d}$
  - $c(\boldsymbol{d})$: cost to fulfill an order $\boldsymbol{d}$

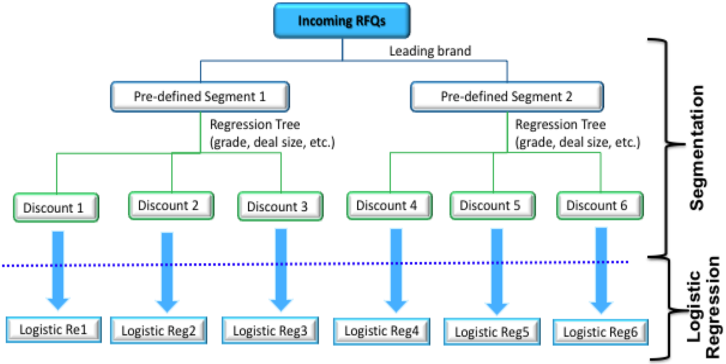


Overall Flow for Real-Time Pricing

Background & Motivation

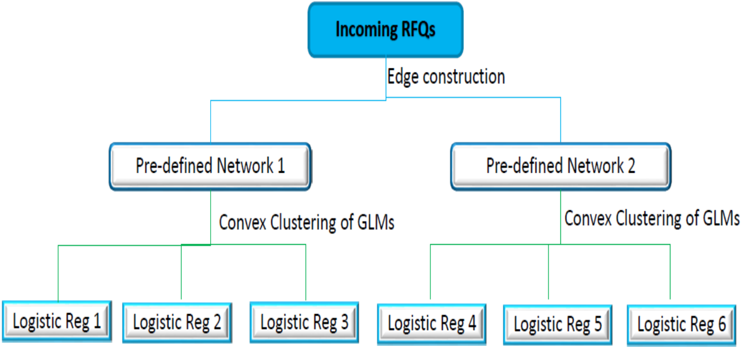
- Main Objective:
  - Predict the purchase probability $q(\boldsymbol{d}, p)$ of a Request-For-Quote (RFQ) for a product configuration ($\boldsymbol{d}$) from a prospective buyer.
  - Model based segmentation for RFQs.

Current Practice (Xue et al., 2015)



- It is a two-step procedure: segmentation first, and then model fitting for each segment.

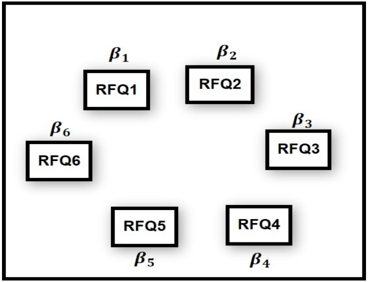
Illustration of The Proposed Method



- It can simultaneously achieve segmentation and model fitting.

Notation

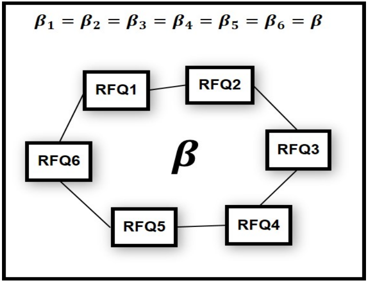
- Denote each RFQ as an observation set.
- In particular, $\text{RFQ}_i = \{\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are bundle features, $\boldsymbol{\beta}_{\boldsymbol{i}} \in \mathbb{R}^{p+1}$ are corresponding coefficients, and the response $y_i \in \{1, -1\}$ indicates whether or not the corresponding client made a purchase.
- Each pair of observations (j, k) can be connected, where $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ will be compared.



Historical Transaction Data

Notation

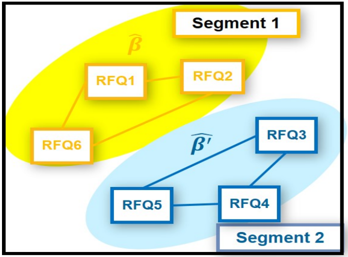
- Denote each RFQ as an observation set.
- In particular, $\text{RFQ}_i = \{\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are bundle features, $\boldsymbol{\beta}_{\mathbf{i}} \in \mathbb{R}^{p+1}$ are corresponding coefficients, and the response $y_i \in \{1, -1\}$ indicates whether or not the corresponding client made a purchase.
- Each pair of observations (j, k) can be connected, where $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ will be compared.



One Global Logistic Regression

Notation

- Denote each RFQ as an observation set.
- In particular, $\text{RFQ}_i = \{\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are bundle features, $\boldsymbol{\beta}_{\boldsymbol{i}} \in \mathbb{R}^{p+1}$ are corresponding coefficients, and the response $y_i \in \{1, -1\}$ indicates whether or not the corresponding client made a purchase.
- Each pair of observations (j, k) can be connected, where $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ will be compared.



Self-segmented Modeling via Convex Clustering

Convex Clustering

- How to achieve clustering and modeling fitting simultaneously?
- Original idea of convex clustering: given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of a data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, it is to

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

- *Idea of convex clustering for GLM*: given data $\{\boldsymbol{x}_i, y_i\}, i = 1, \ldots, N$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{1, -1\}$, we can consider

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

  - f_i is the negative log-likelihood at node i under a network setting.
  - $\lambda w_{jk} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ denotes the penalty for edge (j, k).
  - $w_{jk} \geqslant 0$ is pre-defined weight for edge (j, k), and is set as $w_{jk} = 1$ for simplicity.

Example: Convex Clustering for Logistic Regression

- Consider the logistic regression where the response $y_i \in \{-1, 1\}$. Then $Pr(y_i = 1|\boldsymbol{x}_i)$ is

$$p_i(\boldsymbol{x}_i, \boldsymbol{\beta}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}.$$

- The negative log-likelihood function f_i can be written as,

$$f_i = \log(1 + \exp(-y_i \boldsymbol{x}_i^T \boldsymbol{\beta}_i)).$$

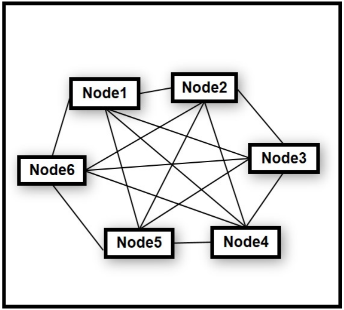
- The first part of objective in the convex clustering for GLM becomes,

$$\sum_{i=1}^{N} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) = \sum_{i=1}^{N} \log(1 + \exp(-y_i \boldsymbol{x}_i^T \boldsymbol{\beta}_i)).$$

Convex Clustering under Network Representation

- Denote the set of observation ids as $\mathcal{V}$ and the set of pairwise ids as $\mathcal{E}$.

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



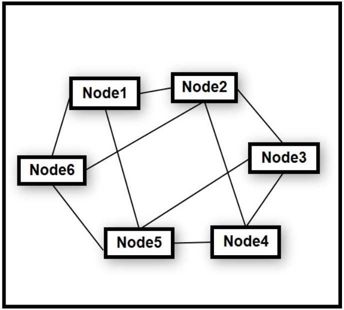
Network Representation

- Initial construction of penalty term: Initial coefficient pairs to be regularized are based on some similarity measures or prior knowledge.

Convex Clustering under Network Representation

- Denote the set of observation ids as $\mathcal{V}$ and the set of pairwise ids as $\mathcal{E}$.

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



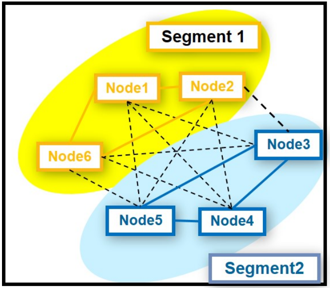
Network Representation

- Initial construction of penalty term: Initial coefficient pairs to be regularized are based on some similarity measures or prior knowledge.

Convex Clustering under Network Representation

- Denote the set of observation ids as $\mathcal{V}$ and the set of pairwise ids as $\mathcal{E}$.

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

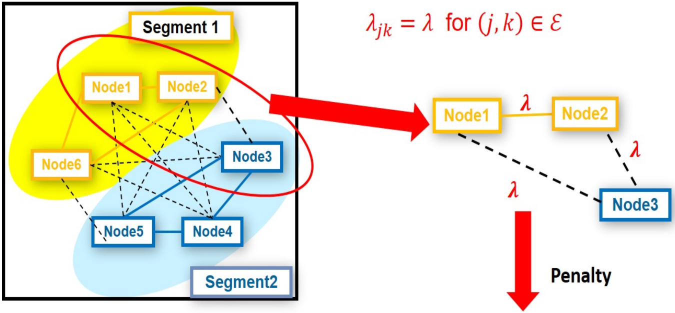


Network Representation

- Initial construction of penalty term: Initial coefficient pairs to be regularized are based on some similarity measures or prior knowledge.

Shrinkage Problem in Convex Clustering

Minimize $\sum_{i \in \mathcal{V}} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i,) + \sum_{(j,k) \in \mathcal{E}} \lambda_{jk} w_{jk} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_k\|_2$.

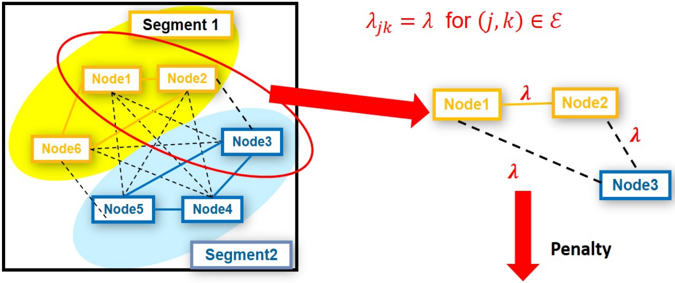


$\lambda w_{12} \|\beta_1 - \beta_2\|_2 + \lambda w_{23} \|\beta_2 - \beta_3\|_2 + \lambda w_{13} \|\beta_1 - \beta_3\|_2$

Shrinkage Problem in Convex Clustering

Minimize $\sum_{i \in \mathcal{V}} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i,) + \sum_{(j,k) \in \mathcal{E}} \lambda_{jk} w_{jk} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_k\|_2.$

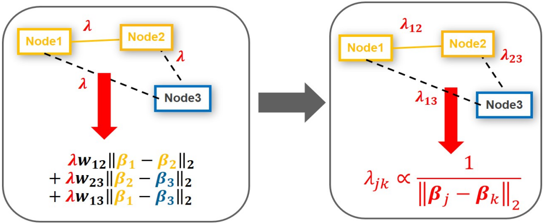
$\lambda_{jk} = \lambda$ for $(j,k) \in \mathcal{E}$



$$\underbrace{\lambda w_{12}\|\beta_1 - \beta_2\|_2}_{\text{Appropriate shrinkage of } \beta_1-\beta_2} + \lambda w_{23}\underbrace{\|\beta_2 - \beta_3\|_2 + \lambda w_{13}\|\beta_1 - \beta_3\|_2}_{\text{Inappropriate shrinkage of } \beta_2-\beta_3, \text{ and } \beta_1-\beta_3}$$

The Proposed Method: Adaptive Convex Clustering

- Alleviate inappropriate shrinkage of $\boldsymbol{\beta}_j - \boldsymbol{\beta}_k$, when j, k belong to different segments.
  - Large λ_{jk} when $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ is small
  - Small λ_{jk} when $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ is large



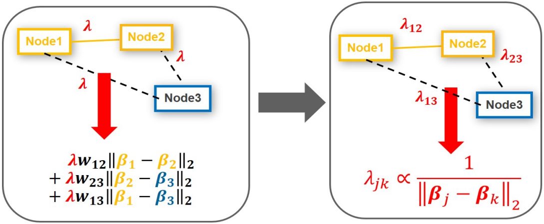
- It is equivalent to using one global λ with adaptive penalty weight,

$$\text{minimize} \quad \sum_{i \in \mathcal{V}} f_i(\boldsymbol{\beta}_i; y_i, \mathbf{x}_i) + \lambda \sum_{(j,k) \in \mathcal{E}} \tilde{w}_{jk} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2,$$

where, $\tilde{w}_{jk} \propto \frac{w_{jk}}{||\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_k||_2}$.

The Proposed Method: Adaptive Convex Clustering

- Alleviate inappropriate shrinkage of $\boldsymbol{\beta}_j - \boldsymbol{\beta}_k$, when j, k belong to different segments.
  - Large λ_{jk} when $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ is small
  - Small λ_{jk} when $||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2$ is large



- It is equivalent to using one global λ with adaptive penalty weight,

$$\text{minimize} \quad \sum_{i \in \mathcal{V}} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) + \lambda \sum_{(j,k) \in \mathcal{E}} \tilde{w}_{jk} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_k||_2,$$

where, $\tilde{w}_{jk} \propto \frac{w_{jk}}{||\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_k||_2}$.

Some Theoretical Properties

Under some mild regularity conditions, the adaptive convex clustering estimates $\hat{\beta}^{*(N)}(glm)$ have good properties in estimation and selection consistency if λ_N is chosen appropriately.

Theorem

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

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

Review: ADMM-based Computational Algorithm

- Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011; Parikh and Boyd, 2014), is a well-established method for solving convex optimization problems.
  - Work well for the quadratic objective function under linear models.
  - Estimate parameters edge by edge.
  - Not very stable for the nonlinear objective function under GLMs.

Proposed Algorithm: Iterative Weighted Least Squares (IWLS)

Based on Newton's Method:

At iteration $t+1$:

$$\begin{cases} \widehat{\boldsymbol{\beta}}_i^{t+1} = \widehat{\boldsymbol{\beta}}_i^t + \left(\boldsymbol{x}_i \hat{p}_i^t (1-\hat{p}_i^t) \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i (y_i - \hat{p}_i^t) \\ \hat{p}_i^t = \frac{\exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_i^t)}{1+\exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_i^t)} \end{cases}$$

$$\Longrightarrow \quad \widehat{\boldsymbol{\beta}}_i^{t+1} = \left(\boldsymbol{x}_i \hat{\pi}_i^t \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i \hat{\pi}_i^t \hat{z}_i^t$$

where, $\hat{z}_i^t = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_i^t + \frac{y_i - \hat{p}_i^t}{\hat{p}_i^t (1-\hat{p}_i^t)}$, $\hat{\pi}_i^t = \hat{p}_i^t (1-\hat{p}_i^t)$

The Network Lasso in each iteration becomes,

Minimize $\sum_{i\in\mathcal{V}} \boldsymbol{f}_i^{t+1} + \lambda \sum_{(j,k)\in\mathcal{E}} \widetilde{w}_{jk} \left\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_k^{t+1}\right\|_2$,

where, $\boldsymbol{f}_i^{t+1} = \hat{\pi}_i^t \left(\hat{z}_i^t - \boldsymbol{x}_i^T \boldsymbol{\beta}_i^{t+1}\right)^2$.

Simulation Settings

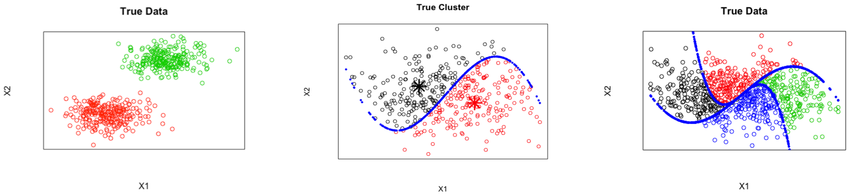
- At each observation i: predictors x_{i1}, x_{i2}, x_{i3}, responses $y_i \in \{1, -1\}$
- Initial construction: 5 nearest neighbors determined by the Euclidian distance of (X_1, X_2)
- Data is randomly split into training (80%) and testing (20%) datasets
- Choose λ to maximize the prediction accuracy (AUC) through 5-fold CV, while the cutting point c is determined such that the sum of sensitivity and specificity is maximized
- Consider 3 different data scenarios
- Each simulation setting is repeated for 50 iterations

Simulated Data

(D1) **Separated Clusters with $K = 2$:** $N = 500$ observations of (X_1, X_2) are simulated shown in Figure (a). The two clusters are clearly separated from each other.

(D2) **Adjacent Clusters with $K = 2$:** $N = 500$ observations of (X_1, X_2) are simulated and the data is split by a nonlinear function of X_1 and X_2, which is the blue curve shown in Figure (b).

(D3) **Adjacent Clusters with $K = 4$:** $N = 900$ observations of (X_1, X_2) are simulated shown in Figure (c). The data is split by two nonlinear functions of X_1 and X_2 (two blue curves).



(a) Separated, $K = 2$ (b) Adjacent, $K = 2$ (c) Adjacent, $K = 4$

Methods & Algorithms in Comparison

(M1) **Optimal Model**: Fit logistic regression model under each true segment.

(M2) **Global Model**: Fit one logistic regression model.

(M3) **K-Means**: Cluster data by K-means according to X_1, X_2 and fit logistic regression under each cluster.

(M4) **ADMM**: Fit convex clustering of GLM by ADMM algorithm.

(M5) **IWLS**: Fit convex clustering of GLM by IWLS algorithm.

(M6) **Adaptive IWLS**: Fit adaptive convex clustering of GLM by IWLS algorithm.

Note that M6 is the proposed approach.

Simulation Results for D1

1. ADMM

Alternating Direction Method of Multipliers

- ❑ Works well for linear regression model
- ❑ Estimate parameters one by one
- ❑ Not stable for nonlinear regression model

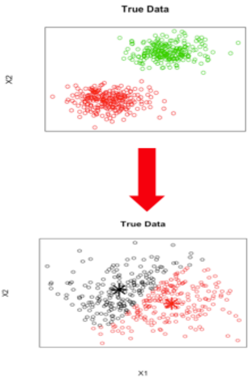
2. IWLS

Iterative Weighted Least Square

- ❑ Linearization of logistic objective function
- ❑ Estimate parameters as a whole set
- ❑ Can be faster in computation

Segment		True	ADMM	IWLS
1	b_0	-1	-0.75 (0.185)	-1.05 (0.220)
	b_1	2.5	1.88 (0.208)	2.53 (0.387)
2	b_0	1.5	0.87 (0.375)	1.52 (0.285)
	b_1	-3.5	-2.02 (0.709)	-3.51 (0.586)
Time(min)			2.22 (0.97)	1.51 (0.214)

Simulation Results Comparison for D1 & D2

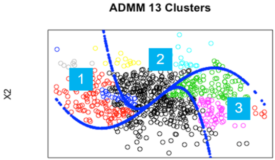


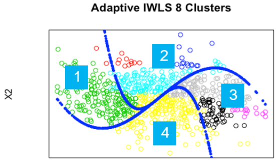
Segment		True	GLM ADMM	ADMM	GLM IWLS	IWLS
1	b_0	-1	-1.02 (0.249)	-0.75 (0.185)	-1.02 (0.249)	-1.05 (0.220)
	b_1	2.5	2.53 (0.353)	1.88 (0.208)	2.53 (0.353)	2.53 (0.387)
2	b_0	1.5	1.56 (0.364)	0.87 (0.375)	1.56 (0.364)	1.52 (0.285)
	b_1	-3.5	-3.61 (0.621)	-2.02 (0.709)	-3.61 (0.621)	-3.51 (0.586)

Segment		True	GLM ADMM	ADMM	GLM IWLS	IWLS
1	b_0	-1	-2.57 (17.56)	-0.40 (0.878)	-1.0 (0.437)	-0.41 (0.236)
	b_1	2.5	0.31 (41.12)	1.15 (1.742)	2.61 (0.572)	1.18 (0.329)
2	b_0	1.5	11.76 (28.60)	0.94 (0.524)	1.57 (0.580)	0.63 (0.230)
	b_1	-3.5	-29.36 (78.24)	-1.92 (0.947)	-3.74 (1.378)	-1.27 (0.477)

- "Wrong" pairs are connected across two segments.
- Using fixed penalty weight in convex clustering leads to serious shrinkage problems.

Simulation Results for D3





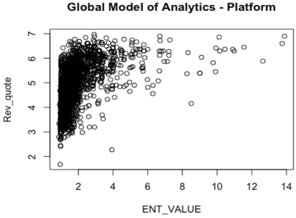
	Classification error	F norm	AUC	F1 score	Precision	Recall
Optimal	0.22 (0.04)	12.12 (5.51)	0.87 (0.03)	0.78 (0.05)	0.79 (0.05)	0.78 (0.08)
Global	0.46 (0.04)	70.14 (0.90)	0.54 (0.04)	0.59 (0.12)	0.54 (0.08)	0.70 (0.20)
K means	0.43 (0.04)	67.08 (1.40)	0.61 (0.04)	0.55 (0.10)	0.59 (0.07)	0.56 (1.97)
ADMM	0.51 (0.04)	46.56 (3.08)	0.49 (0.05)	0.66 (0.07)	0.49 (0.04)	0.98 (0.15)
IWLS	0.46 (0.04)	69.46 (0.71)	0.53 (0.05)	0.58 (0.09)	0.53 (0.06)	0.63 (0.18)
Adaptive IWLS	**0.31** (0.04)	**53.46** (12.02)	**0.75** (0.05)	**0.68** (0.06)	**0.70** (0.05)	**0.69** (0.13)

Note that the Frobenius norm, $\mathrm{Fnorm} = \sqrt{\sum_{i=1}^{n} \sum_{j=0}^{p} \left(b_{ij} - \beta_{ij}\right)^2}$, measures the estimation accuracy of coefficients.

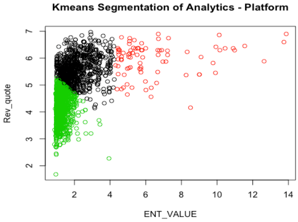
Application on IT Pricing Data

- Two sets of data are considered based on the product categories: Brand1 (Analytics platform) and Brand2 (Security)
- Total number of RFQs: $N_1 = 2682$, $N_2 = 2642$.
- Three independent features X_1, X_2, X_3 are generated by Xue et al. (2015) describing the characteristics of each bundle.
- The full data is split into 80% training and 20% testing.
- Research Goal: Predict the purchase likelihood for each RFQ while taking into account the heterogeneity in model performance

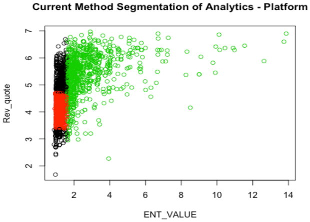
Results for Brand1



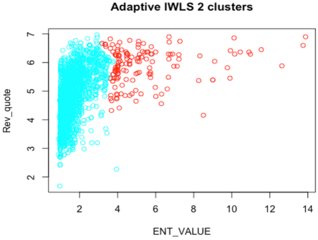
(a) Global-Brand1



(b) Kmeans-Brand1



(c) Current-Brand1



(d) Adaptive-Brand1

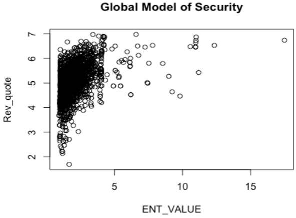
Results for Brand1

	Classification Error	F1 score	AUC	Precision	Recall
Global	0.4	0.53	0.653	0.47	0.62
K means	0.39	0.55	0.652	0.48	0.65
Current	0.44	0.49	0.584	0.43	0.58
Adaptive IWLS	0.4	0.55	0.659	0.47	0.67

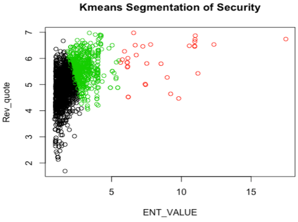
✓ **Compare with Global and K means**

✓ **Slightly better than Current**

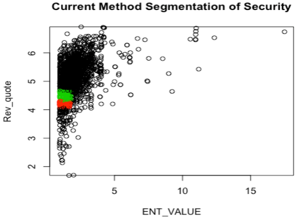
Results for Brand2



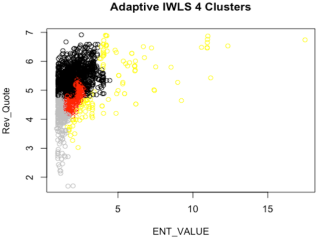
(a) Global-Brand2



(b) Kmeans-Brand2



(c) Current-Brand2



(d) Adaptive-Brand2

Results for Brand2

	Classification Error	F1 score	AUC	Precision	Recall
Global	0.40	0.41	0.614	0.32	0.57
K means	0.41	0.41	0.614	0.32	0.56
Current	0.47	0.37	0.542	0.28	0.56
Adaptive IWLS	0.32	0.39	0.624	0.37	0.41

✓ **20% improve from Global**

✓ **32% improve from Current**

Discussion

- Propose adaptive convex clustering method for GLMs, which can perform segmentation and model fitting simultaneously.
- The IWLS-based algorithm is developed to achieve better convergency properties in parameter estimation.
- Future research directions include how to obtain a relatively balanced segmentation structure for business data.
- Stability of cross-validation for binary data will be further explored.

List of 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.

Hestenes, M. R. (1969), "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, 4, 303–320.

Jiang, H., Lozano, A. C., and Liu, F. (2012), "A Bayesian Markov-switching model for sparse dynamic network estimation." in *SDM*, SIAM, pp. 506–515.

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2010), "A hierarchical Bayesian framework for constructing sparsity-inducing priors," *arXiv preprint arXiv:1009.1914*.

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

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.

Thank You
For
Your Attention!

Any Questions?

Bayesian Justification of Adaptive Convex Clustering

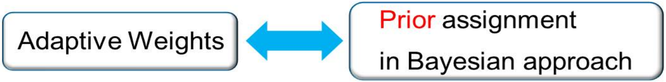
Adaptive Weights ⟷ Prior assignment in Bayesian approach

- Denote $\boldsymbol{\beta}_e = (\beta_{e0}, \cdots, \beta_{ep})^T = \boldsymbol{\beta}_j - \boldsymbol{\beta}_k$, $e = 1, \cdots, E, (j,k) \in \mathcal{E}$
- The hierarchical priors given to the coefficient difference on each $\boldsymbol{\beta}_e$ are (Lee et al., 2010; Jiang et al., 2012),

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

where, $G(a,b)$ denotes the Gamma distribution, and $IG(a,b)$ represents the Inverse Gamma distribution.

Bayesian Justification of Adaptive Convex Clustering



- Denote $\boldsymbol{\beta}_e = (\beta_{e0}, \cdots, \beta_{ep})^T = \boldsymbol{\beta}_j - \boldsymbol{\beta}_k$, $e = 1, \cdots, E, (j,k) \in \mathcal{E}$
- The hierarchical priors given to the coefficient difference on each $\boldsymbol{\beta}_e$ are (Lee et al., 2010; Jiang et al., 2012),

$$\boldsymbol{\beta}|\sigma_1^2, \cdots, \sigma_E^2 \sim N(0, \Sigma_{\boldsymbol{\beta}}), \quad \boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_N^T\right)^T,$$
$$\sigma_e^2|\tau_e \sim G(\frac{p+1}{2}, 2\tau_e^2),$$
$$\tau_e|a_e, b_e \sim IG(a_e, b_e),$$

where, $G(a,b)$ denotes the Gamma distribution, and $IG(a,b)$ represents the Inverse Gamma distribution.

Bayesian Justification of Adaptive Convex Clustering

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

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

- $\mathcal{N}(i)$ denotes the neighbors of node i, and $\sigma^2_{(i,j)} = \sigma^2_{(j,i)}$
- The corresponding iterative procedure to solve for $\boldsymbol{\beta}$ is,

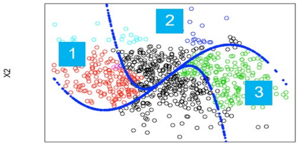
$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i\in\mathcal{V}} f_i(\boldsymbol{\beta}_i; y_i, \boldsymbol{x}_i) + \sum_{e\in\mathcal{E}} w_e^{(t+1)} ||\boldsymbol{\beta}_e^{(t)}||_2,$$

where,

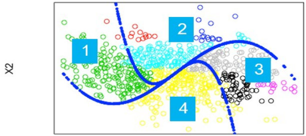
$$w_e^{(t+1)} = \frac{a_e + p}{||\boldsymbol{\beta}_e^{(t)}||_2 + b_e}$$

Simulation Results for D3

Adaptive ADMM 6 Clusters



Adaptive IWLS 8 Clusters



Segment		True	GLM (Adap ADMM)	Adaptive ADMM	GLM (Adap IWLS)	Adaptive IWLS
1	b_0	**-1**	-0.584	-0.299	**-0.615**	**-0.548**
	b_1	**2.5**	2.290	1.341	**2.050**	**1.861**
2	b_0	**1.5**	0.235	0.188	**1.500**	**0.965**
	b_1	**-3.5**	-1.468	-0.866	**-1.959**	**-1.685**
3	b_0	**0.5**	0.289	0.222	**0.329**	**0.307**
	b_1	**1.5**	1.150	0.850	**1.566**	**1.309**
4	b_0	**-0.5**	0.235	0.188	**-0.087**	**-0.062**
	b_1	**-1.5**	-1.468	-0.886	**-1.284**	**-1.150**
...			...	...	...	...