VARIABLE SELECTION METHODS FOR STRUCTURED DATA

by

Chen Cheng

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 10/05/2015

The dissertation is approved by the following members of the Final Oral Committee:

Chunming Zhang, Professor, Statistics
Xiaoxia Shi, Assistant Professor, Economics
Kam-Wah Tsui, Professor, Statistics
Zhengjun Zhang, Professor, Statistics
Jun Zhu, Professor, Statistics

To my family

Acknowledgments

First, I would like to express my sincerest appreciation and gratitude to my advisor, Professor Chunming Zhang, for her invaluable guidance and persistent help in my Ph.D. study. I could not feel more fortunate to be a Ph.D. student of hers, since I learn so much from her, not only on statistics but also on personality. She has guided me through each specific research questions, and impacted me greatly on being a better researcher. Her great sense of research direction and rich experiences in statistical theories has inspired me, and her support and encouragement plays an essential role throughout my Ph.D. study.

Besides, I want to thank Professor Xiaoxia Shi, Professor Kam-Wah Tsui, Professor Zhengjun Zhang, Professor Jun Zhu and Professor Yingqi Zhao for kindly serving as my committee members. Id like to thank Professor Zhengjun Zhang especially for encouraging me and sharing his experience with me in research and teaching. I am very grateful to Professor Jun Zhu for her valuable suggestions and comments which help me a lot to improve the thesis. I also want to thank Yi Chai, Lilun Du, Xiao Guo for being my wonderful research colleagues with all the help and supports. I would like to send my thanks to the Department of Statistics at the University of

Wisconsin-Madison, for the diversified and rigid training that not only builds me up a solid background, but also benefits my research and future career in every way. Thanks go to all the faculty members and staff.

This dissertation is dedicated to my parents. Their selfless love supports me every day in my life.

Contents

List of Tables

List of Figures

Abstract

This dissertation focuses on developing regularization models for structured data.

Chapter 1 reviews classic regularization methods in the literature. Chapter 2 introduces the penalized Group-Bregman Divergence (BD) model. It investigates new aspects of variable selection for group-structured data by relaxing the restrictions on data distribution while achieving sparsity at both group and within-group level. Theoretical results show both asymptotic and non-asymptotic properties. Numerical studies are illustrated by both simulation and real data analysis. Chapter 3 describes future application of penalized Group-BD model on spike train data in neuronscience, and a continuous methodology for comparison. To study the functional connectivity, the models are applied to each neuron and build the network by finding out a subset of neurons out of a large pool that have either excitation or inhibition effect on the target one.

Chapter 1

Introduction

1.1 Literature Review of Regularization Methods

Model selection is a popular topic aiming at selecting important variables that contribute to prediction of the response variable. In practice, usually a large pool of variables are considered but only a few of them are significant. We call this sparsity. For example in genetic study, information of thousands of genes is collected but only a few pathways and genes are related to the feature of interest. To discover sparsity in selected model which cannot be achieved by the classic ordinary least square regression, the penalized method, known as regularization, is proposed to overcome the problem of over-parameterization. It estimates coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ of variables by minimizing the target function:

$$T(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + P_\kappa(\boldsymbol{\beta}),$$

where $L(\boldsymbol{\beta})$ is the loss function measuring how fit the model is, and $P_\kappa(\boldsymbol{\beta})$ is the penalty function which assesses the physical plausibility of $\boldsymbol{\beta}$ (Zhang et al., 2010). Here, κ is a tuning parameter that regularizes the penalty.

The classic Lasso method applies the L_1 norm to coefficients as the penalty.

$$P_\kappa(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1.$$

The singularity at 0 makes it possible to select a sparse model. But desirable oracle properties such as model selection consistency do not hold for Lasso generally (Fan and Li, 2001). Fan and Li (2001) proposed the smoothly clipped absolute deviation penalty known as SCAD with penalty function.

$$P_\kappa(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\kappa(|\beta_j|)$$

where

$$p'_\kappa(|\beta|) = I(|\beta| \leq \kappa) + \frac{(a\kappa - |\beta|)_+}{(a-1)\kappa} I(|\beta| > \kappa),$$

for some $a > 2$. It satisfies the three properties—sparsity, unbiasedness, and continuity (Fan and Lv, 2010), and has oracle properties (Fan and Li, 2001).

Both Lasso and SCAD consider parameters individually. However, it is not a rare situation where there is group structure among variables. For example, when a feature is categorical, we introduce dummy variables to consider differences between levels. Then those for different levels of the same categorical feature can be considered as a group. It is also natural in areas such as genetic study. Genes from the same

pathway or probes from the same gene form a group naturally. There are also many other situations in which variables come in natural groups. To formulate the group structure, assume that we have n observations, p variables, and K groups. Then the parameter vector can be written as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp_k})^T$ is the sub-vector for the kth group. To take into account the group structure, several methods were proposed using different penalty functions in the literature. Yuan and Lin (2006) proposed the Group Lasso criterion by using the L_2 norm within each group and the L_1 norm between groups.

$$P_\kappa(\boldsymbol{\beta}) = \kappa \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_2.$$

Appropriate tuning parameter κ can set some group coefficient vector $\boldsymbol{\beta}_k = \mathbf{0}$ because of the singularity of the penalty at $\boldsymbol{\beta}_k = \mathbf{0}$. Within each group the L_2 norm is applied, so if one variable is selected to be significant, then all the other variables in the same group will be selected as well. There are also generalizations of the Group Lasso such as the Composite Absolute Penalty (CAP) (Zhao et al., 2009), which applies L_{γ_k} norm within group k, then take L_{γ_0} norm of the K group norms:

$$P_\kappa(\boldsymbol{\beta}) = \kappa \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_{\gamma_k}^{\gamma_0},$$

where $\gamma_i \geq 1$ are constants for $i = 0, 1, \cdots, K$. These methods share the same limitation: they can achieve sparsity at the group level but fail to discover sparsity within groups. However in real problems, within group sparsity might also be desirable.

To achieve variable selection between and within groups simultaneously, different

penalty structures are considered. For example Sparse Group Lasso (Simon et al., 2013) uses a convex combination of the Group Lasso penalty and the Lasso penalty.

$$P_\kappa(\boldsymbol{\beta}) = \alpha\kappa\|\boldsymbol{\beta}\|_1 + (1-\alpha)\kappa\sum_{k=1}^{K}\sqrt{p_k}\|\boldsymbol{\beta}_k\|_2,$$

where $\alpha \in [0, 1]$.

In Chapter 2 we will apply the Hierarchical Lasso (Zhou and Zhu, 2010) as the penalty part. It decomposes each parameter into two hierarchies: group index and individual index. The jth parameter in group k, β_{kj}, is written as: $\beta_{kj} = d_k\alpha_{kj}$, where $d_k > 0$ is the group index and α_{kj} is the individual index. Then the penalty function can be expressed as

$$P_\kappa(\boldsymbol{\beta}) = \sum_{k=1}^{K} d_k + \kappa\sum_{j=1}^{p_k}|\alpha_{kj}|.$$

This penalty form has several advantages. First, it can remove insignificant groups and select variables within groups flexibly. Second, compared with methods such as the Sparse Group Lasso, it avoids introducing extra tuning parameter to the model. Although it is not convex, we can still construct a quick iterative algorithm with a simple analytic form for each iteration. It's not difficult to check that the Hierarchical Lasso has sparsity and is nearly unbiased under the criteria in Fan and Lv (2010). It might violate the continuity property, but the discrepancy is negligible especially when the dimension is high.

The similarity among these group variable selection methods is that they apply certain norm to each group respectively and then consider a different norm of the

resulting group norms.

1.2 Loss Function

Compared to the study of the penalty function, the loss function part is somehow ignored. Most results are based on the square error loss which is suitable for Normal response but not for problems such as classification. Likelihood based functions are also widely used as the loss which requires the distribution to be fully specified. When the distribution of the data cannot be determined easily, it is desirable to develop properties of group variable selection methods adjusted to a wider range of loss functions. In Zhang et al. (2010) the author incorporate a class of loss functions called Bregman Divergence with component-wise penalties such as Lasso and SCAD. This loss function family includes many commonly used loss functions such as the square error loss and log-likelihood loss for exponential family, but not limited to these. It also applies to situations where the distribution of the observation is unknown or not fully specified. Zhang et al. (2009) gives new aspects of Bregman Divergence applied on non-parametric models as well.

In Chapter 2 we incorporate the Bregman Divergence loss family with the Hierarchical Lasso penalty to achieve group variable selection under a wider scope. We call the corresponding method the penalized Group-BD. In Chapter 3 we consider the application of penalized Group-BD on spike train data. Also we consider a piecewise exponential distribution and use the corresponding log-likelihood function as the loss function. Piecewise exponential distribution is widely used in survival analysis to

model time-to-event data. Rajaram et al. (2005) proposed a more detailed piecewise exponential distribution model designed for Poisson network, which accord with the spike train data.

Chapter 2

The Penalized Group Bregman Divergence Method

2.1 Overview

To achieve variable selection for data with group structure, we have two main concerns: what kind of penalty function to use so that sparsity can be well detected at both group and within-group level, and what kind of loss function to use to adapt to break the restriction of data distribution.

For the penalty part, assume there are K groups and p_k variable in each group. The original penalty considering the group structure is the Group Lasso (Yuan and Lin, 2006) using $\kappa \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_2$ as the penalty, where $\boldsymbol{\beta}_k$ is the coefficient vector for group k and $\kappa > 0$ is the tuning parameter. The non-differentiability of $||\boldsymbol{\beta}_k||_2$ at and only at $||\boldsymbol{\beta}_k||_2 = \mathbf{0}$ provides two bits of information. One is this penalty can achieve

sparsity by eliminating insignificant groups. The other is variables in the same group can only be kept or excluded together. In other words, if a group is selected, all the variables in that group will be selected. However, in real application, there might also be insignificant variables in significant groups. For example, for the spike train data, the exact length of history we should trace back to is unknown. So if we could allow sparsity within each group as well, we could consider a comparatively longer history without losing information or over fitting. There are penalties proposed in literature to achieve sparsity within group as well. Here we compare two popular ones with clear and intuitive penalty form: Sparse Group Lasso (SGL) (Simon et al., 2013) and Hierarchical Lasso (HL) (Zhou and Zhu, 2010). The SGL has penalty function $\kappa \sum_{k=1}^{K} [\alpha ||\boldsymbol{\beta}_k||_1 + (1-\alpha)\sqrt{p_k}||\boldsymbol{\beta}_k||_2]$ where $\alpha \in [0,1]$ is a index controlling the balance between group-level sparsity and within-group sparsity. The HL considers penalty function $\kappa \sum_{k=1}^{K} \sqrt{||\boldsymbol{\beta}_k||_1}$. It is equivalent to a more intuitive form $\kappa \sum_{k=1}^{K} \{d_k + \sum_{j=1}^{p_k} \alpha_{kj}\}$ if each coefficient β_{kj} is decomposed into two parts: group index d_k and individual index α_{kj}. Figure 2.1 gives a simple illustration of the penalty function and contour within one group for both methods. We can see that within one group, the SGL uses elastic net penalty (Zou and Hastie, 2005) which is a convex combination of the L_1 and the L_2 penalty, while the HL uses square root of the L_1 penalty. Figure 2.2 illustrates the penalty function with different groups. From the graphs we can clearly see that both methods can select variables at both group and within-group level, and the HL tend to encourage more sparsity both within and between groups under certain loss. In general it is hard to tell which method is always superior to the other, but under certain circumstances one might give a

better result. We choose the HL over the SGL based on the following considerations. First, the SGL has a index α. It can be pre-determined depending on how we want to distribute the sparsity. However for real problems, it is unknown and people usually use grid search over $[0, 1]$ to get optimal result. Thus compared to the HL, the SGL has one more parameter to tune. Second, consider the penalty function within each group denoted by$Pen_k(\boldsymbol{\beta}_k)$, we can see that $|dPen_k^{SGL}(\boldsymbol{\beta}_k)/d\boldsymbol{\beta}_k|$ is bounded away from 0 when $||\boldsymbol{\beta}_k|| \to \infty$, while $|dPen_k^{HL}(\boldsymbol{\beta}_k)/d\boldsymbol{\beta}_k| \to 0$ when $||\boldsymbol{\beta}_k|| \to \infty$. So when the true parameter values are large, HL tend to provide better estimates than the the SGL. Although the HL is no longer a convex penalty, the decomposition into group index and individual index makes it possible to implement through simple iterative algorithm.

For the loss function part, to broaden the scope, we consider the Bregman Divergence (BD) (Bregman, 1967), which is a family of generalized distance measures. It takes the following form:

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu)$$

where $q(\cdot)$ is called the generating function of $Q(\cdot, \cdot)$. Assume that $q(\cdot)$ is concave with the derivative $q'(\cdot)$. It is easy to see that the concavity of $q(\cdot)$ ensures that $Q(\cdot, \cdot)$ is a non-negative function. When $q(\cdot)$ is strictly concave, $Q(\nu, \mu) = 0$ if and only if $\nu = \mu$. So if we assign ν and μ as the observation and estimate respectively, $Q(\nu, \mu)$ acquires favorable properties as a loss function.

By using different generating function $q(\cdot)$, the Bregman Divergence adjusts to various types of error measures including many popular ones in the literature (Zhang

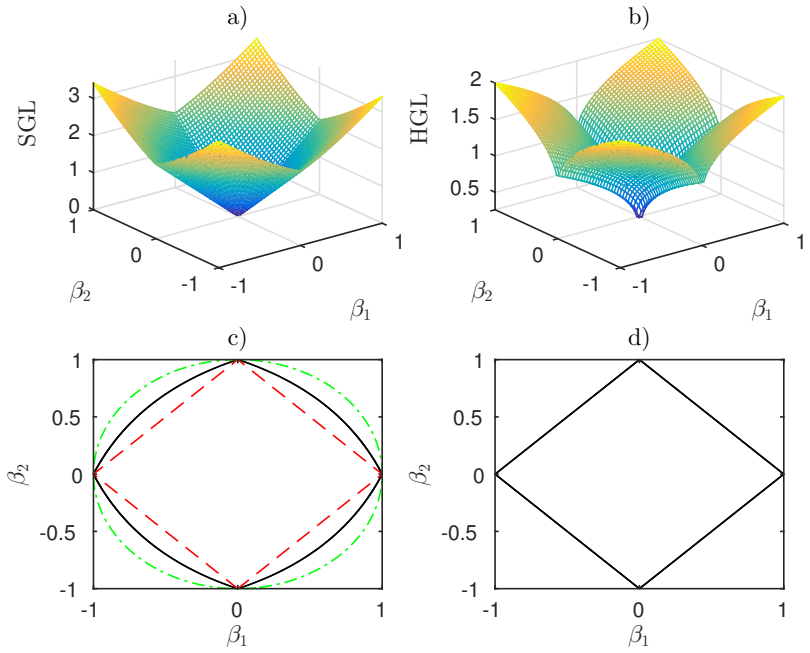

Figure 2.1: *β_1 and β_2 belong to one group. a) and b): penalty functions for SGL ($\alpha = 0.5$) and HL; c): contour of the penalty function for SGL (black solid line), Lasso (red dashed line) and ridge (green dash-dot line); d): contour of HL penalty, same as Lasso.*

et al., 2011). For example, when $q(\mu) = a\mu - \mu^2$ with a constant a, $Q(y, \mu) = (y - \mu)^2$ which is the quadratic loss. When $q(\mu) = \min\{\mu, 1-\mu\}$, $Q(y, \mu) = I\{y \neq I(\mu > 0.5)\}$ which is the misclassification loss for binary responses. $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu) \log(1-\mu)\}$ gives the Bernoulli deviance loss $Q(y, \mu) = -\{y \log(\mu) + (1-y) \log(1 - \mu)\}$ while $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$ yields the exponential loss $Q(y, \mu) = \exp\{-(y - 1/2) \log(\mu/(1-\mu))\}$ (Hastie et al., 2001). If $q(\mu) = \mu - \mu \log(\mu)$, then we have the quasi-likelihood loss $Q(y, \mu) = y\{\log(y) - \log(\mu)\} - (y - \mu)$. With the necessary and sufficient conditions provided in Zhang et al. (2009), a given loss function Q, such as

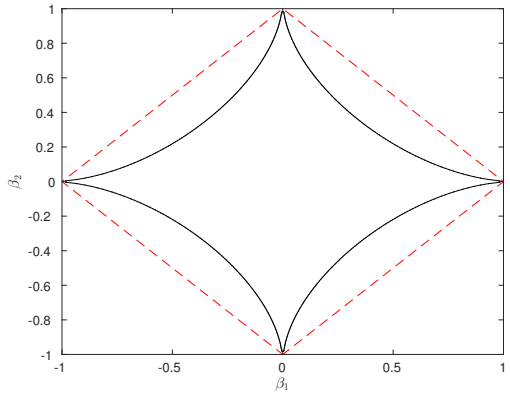

Figure 2.2: *Contour of the penalty function for SGL (red dashed line) and HL (black solid line) when β_1 and β_2 belong to different groups.*

Kullback-Leibler divergence and many margin-based loss functions, can also be shown to belong to the BD family (Zhang et al., 2010). The choice of the BD function is mainly decided by the data type instead of the exact distribution. As long as a the chosen BD function is proper, it provides as good performance no matter which exact one is used. Thus we no longer have the concern of illy assumed distribution.

2.2 The Penalized Group-BD Estimator

2.2.1 Model Specification

Let $\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ be observations from some underlying population $(\boldsymbol{X}, Y)$ where $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ is the vector of explanatory variables and Y is the scalar response. Assume the general linear model $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}) = F^{-1}(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}) = F^{-1}\big(\beta_0 + \sum_{k=1}^{K} d_k \sum_{j=1}^{p_k} \alpha_{kj} x_{kj}\big)$ where β_0 is the intercept and $F(\cdot)$

is a link function. For simplicity of notation, we consider the intercept term is included in $\boldsymbol{\beta}$ and $\boldsymbol{X}$ unless otherwise specified.

Combining BD with the Hierarchical Lasso penalty, the target problem is to minimize the following criterion function:

$$T(\boldsymbol{d}, \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, m(\boldsymbol{X}_i)) + \sum_{k=1}^{K} d_k + \kappa \sum_{j=1}^{p_k} |\alpha_{kj}| \tag{2.1}$$

We call the corresponding estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{gBD}}$ the penalized Group-BD estimator:

$$\widehat{\boldsymbol{\beta}}_{kj}^{\mathrm{gBD}} = \widehat{d}_k \widehat{\alpha}_{kj},$$

where $\widehat{d}_k$ and $\widehat{\alpha}_{kj}$, $k = 1, \ldots, K$, $j = 1, \ldots, p_k$ are minimizers of (2.1)

Theorem 2.1 below shows the equivalence between the optimization problem above and that expressed in the form of the original coefficients $\{\beta_{kj} : k = 1, \ldots, K, j = 1, \ldots, p_k\}$.

Theorem 2.1. *If* $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ *is a local minimum of* (2.1)*, where* $d_k \geq 0, k = 1, \ldots, K$*, then* $\widehat{\boldsymbol{\beta}}$*, where* $\widehat{\beta}_{kj} = \widehat{d}_k \widehat{\alpha}_{kj}$*, is a local minimum of*

$$T(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, F^{-1}(\boldsymbol{X}_i \boldsymbol{\beta})) + 2\sqrt{\kappa} \sum_{k=1}^{K} \sqrt{|\beta_{k1}| + |\beta_{k2}| + \cdots + |\beta_{kp_k}|}. \tag{2.2}$$

On the other hand, if $\widehat{\boldsymbol{\beta}}$ *is a local minimum of* (2.2)*, then we define* $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$*, where* $\widehat{d}_k = 0, \widehat{\alpha}_k = 0$ *if* $\|\widehat{\boldsymbol{\beta}}_k\|_1 = 0$*, and* $\widehat{d}_k = (\kappa \|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}, \widehat{\alpha}_k = \widehat{\boldsymbol{\beta}}_k / (\kappa \|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$ *if* $\|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0$*. Then the so-defined* $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ *is a local minimum of* (2.1)*.*

The equivalence of the penalty term in (2.1) and (2.2) shows that the penalized Group-BD estimate is identifiable. As a comparison, (2.1) gives a form more convenient for computation while the form in (2.2) will be used in theoretical results.

2.2.2 Non-asymptotic Error Bounds

Bickel et al. (2009) shows some finite-sample error bounds for Lasso and Dantzig estimator under the Restricted Eigenvalue (RE) assumption. Incorporating the group structure to the RE assumption, we show some similar non-asymptotic bounds for the Group-BD estimator.

Let $c_{\max} = \max_{k=1,\ldots,K}(\|\widehat{\boldsymbol{\beta}}\|_1)^{1/2}$ and $c_{\min} = \min_{k=1,\ldots,K}(\|\widehat{\boldsymbol{\beta}}\|_1)^{1/2}$. Although the continuity property is violated for the Hierarchical Lasso penalty, it makes the assumption acceptable that the estimator is bounded away from 0. The RE assumption with group structure is constructed as follows:

Group RE **Assumption:** Let $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$. For a positive integer s and any $\delta \in \mathbb{R}^p$, the following condition holds:

$$\zeta(s) \triangleq$$

$$\min_{G \subset \{1,\ldots,K\}, |G| \le s} \; \min_{\forall \delta \neq 0, \sum_{k \notin G} \|\delta_k\|_1 \le \sum_{k \in G}(1+2\frac{c_{\max}}{c_{\min}})\|\delta_k\|_1} \frac{2\|\mathbf{X}\delta\|_2}{\sqrt{n}\sqrt{\sum_{k \in G} p_k (1+\frac{c_{\max}}{c_{\min}})^2 \|\delta_k\|_2^2}}$$

$$> 0.$$

Let $q_j(y, \theta) = (\partial^j / \partial \theta^j) Q(y, F^{-1}(\theta))$. Then we have Theorem 2.2.

Theorem 2.2. *Consider the model in* (2.2). *Let* $\boldsymbol{\beta}^0$ *be the vector of true values for the coefficients and* $\widehat{\boldsymbol{\beta}}$ *be the solution to* (2.2). *Assume that* $\|\mathbf{X}\|_\infty \le c_1$ *and for any*

$\boldsymbol{\beta}$ and $(y, \boldsymbol{x})$, $q_2(y, \boldsymbol{x}^T\boldsymbol{\beta}) > c_2$ where c_1 and c_2 are constants. Let $G(\boldsymbol{\beta}) = \{1 \le k \le K | \exists j, 1 \le j \le p_k, s.t. \beta_{kj} \neq 0\}$ be the support of $\boldsymbol{\beta}$. Assume the Group RE *assumption holds with $\zeta = \zeta(s)$ where $s = |G(\boldsymbol{\beta}^0)|$. Let $A > 1$ and $\gamma = Ac_1\{2\log(p)\}^{1/2}/n$. Let $W = (n)^{-1/2}\sum_{i=1}^n q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta})$. Assume that the tail behavior of W satisfies that for any $w > \gamma/c_1$, $P(W > w) \le P(Z > w)$ where Z follows a standard normal distribution. Then with probability at least $1 - p^{1-A^2}$, the following inequalities hold:*

$$\frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \le \frac{16s\gamma^2}{\zeta^2 c_1^2}, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \le \frac{16s\gamma}{\zeta^2 c_1}, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \le (2\sqrt{s}+1)\frac{8\sqrt{s}\gamma}{\zeta^2 c_1}. \quad (2.3)$$

The first part in (2.3) shows the upper bound for prediction error and the rest two show bounds for variable selection loss using both L_1 and L_2 norm.

2.2.3 Adaptive Penalized Group-BD Estimator

The original hierarchical penalty forces the coefficients to be equally penalized which seems to be unfair in some sense. In order to further improve the performance of the penalized Group-BD estimates, we adopt the idea of adding different weights to coefficients in the penalty term known as the adaptive method (Zhou and Zhu, 2010; Zou, 2006). The minimization problem is specified as follows:

$$\begin{aligned} \min_{\boldsymbol{\beta}} T(\boldsymbol{\beta}) &= \min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^n Q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta})) \\ &\quad +2\sqrt{\kappa}\sum_{k=1}^K \sqrt{w_{k1}|\beta_{k1}| + w_{k2}|\beta_{k2}| + \cdots + w_{kp_k}|\beta_{kp_k}|}, \end{aligned}$$

where $\{w_{kj}\}$ are pre-specified weights. Intuitively, if a variable is significant, which means its effect is not negligible, then we should not penalize it too much. In other words the corresponding weight should be small. On the other side, coefficients for insignificant variables should be heavily penalized by large weights. In practice estimates from some simple or quick methods can be used as the weights, for example the ordinary least square estimate:

$$w_{kj} = \frac{1}{|\widehat{\beta}_{kj}^{\text{ols}}|^{\gamma}},$$

or ridge estimate for high dimensional cases:

$$w_{kj} = \frac{1}{|\widehat{\beta}_{kj}^{\text{ridge}}|^{\gamma}}.$$

where γ is a positive constant.

The adaptive Lasso enjoys consistency and oracle properties (Zou, 2006) that the original Lasso does not have under general assumptions. Similarly, for the penalized Group-BD estimator with properly chosen weights, the adaptive estimator can achieve desirable asymptotic properties which will be discussed in Section 2.3.

2.3 Asymptotic Properties

2.3.1 Problem Setup

To consider asymptotic properties, in this section we assume that the dimension n is diverging. Now we have p_n variables in total and p_{nk} variables in the kth group for

$k = 1, \ldots, K$. For every n, the observations $(\boldsymbol{X}_{ni}, Y_{ni}), i = 1, 2, \ldots, n$ are i.i.d. from population $(\boldsymbol{X}_n, Y_n)$. The subscript n shows that the term now is changing with n. Then the problem of interest becomes

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{\text{gBD}} &= \arg\min_{\boldsymbol{\beta}_n} T_n(\boldsymbol{\beta}_n) = \arg\min_{\boldsymbol{\beta}_n} \frac{1}{n}\sum_{i=1}^{n} Q(Y_{ni}, F^{-1}(\boldsymbol{X}_{ni}^T\boldsymbol{\beta})) \\ &\quad +2\sqrt{\kappa_n}\sum_{k=1}^{K}\sqrt{w_{n,k1}|\beta_{n,k1}| + w_{n,k2}|\beta_{n,k2}| + \cdots + w_{n,kp_k}|\beta_{n,kp_k}|}. \quad (2.4)\end{aligned}$$

Let $\boldsymbol{\beta}_n^0$ denotes the true parameter vector and $\boldsymbol{\beta}_{nk}^0$ denotes the corresponding sub-vector for group k. Without loss of generality, we assume that within group k, the first s_{nk} entries of $\boldsymbol{\beta}_{nk}^0$ are non-zero while the rest are zero for $k = 1, \ldots, K$. Then denote $s_n = \sum_{k=1}^{K} s_{nk}$ as the total number of truly significant variables, $\mathcal{S} = \{(k, j) : \beta_{n,kj}^0 \neq 0\}$ as the index set of truly significant parameters of $\boldsymbol{\beta}_n^0$, and $\mathcal{S}^c$ as the index set of truly insignificant parameters. Let $\boldsymbol{\beta}_n^{(\text{I})} = \boldsymbol{\beta}_{n,\mathcal{S}}$ be the sub-vector of $\boldsymbol{\beta}_n$ corresponding to the significant index set and $\boldsymbol{\beta}_n^{(\text{II})}$ contains the remaining entries. Also we define $\boldsymbol{X}^{(\text{I})}$ and $\boldsymbol{X}^{(\text{II})}$ as entries of $\boldsymbol{X}$ corresponding to $\boldsymbol{\beta}_n^{(\text{I})}$ and $\boldsymbol{\beta}_n^{(\text{II})}$ respectively. For notational simplicity, we consider that the intercept term is included in $\boldsymbol{\beta}_n^{(\text{I})}$ and $\boldsymbol{X}^{(\text{I})}$ unless otherwise specified.

Define

$$\begin{aligned}H_n &= -E\Big\{\frac{q''(m(\boldsymbol{X}_{n1}))}{F'(m(\boldsymbol{X}_{n1}))^2}\boldsymbol{X}_{n1}^{(\text{I})}\boldsymbol{X}_{n1}^{(\text{I})T}\Big\}, \\ \Omega_n &= E\Big\{\text{var}(Y_{n1}|\boldsymbol{X}_{n1})\frac{q''(m(\boldsymbol{X}_{n1}))^2}{F'(m(\boldsymbol{X}_{n1}))^2}\boldsymbol{X}_{n1}^{(\text{I})}\boldsymbol{X}_{n1}^{(\text{I})T}\Big\}.\end{aligned}$$

To achieve asymptotic properties, we make some general assumptions:

A1. $E(\boldsymbol{X}_n) = \mathbf{0}$, $\sup_n \|\boldsymbol{X}_n\|_\infty < \infty$, and $\sup_n E(Y_n^2) < \infty$.

A2. $\boldsymbol{\beta}_n^0$ is sparse and $\sup_n \|\boldsymbol{\beta}_n^0\|_1 < \infty$.

A3. The generating function $q(\cdot)$ and the link function $F(\cdot)$ are smooth, $q''(\cdot) < 0$ and $F'(\cdot) \neq 0$.

A4. $\inf_{n\in\mathbf{N},1\leq j\leq p_n} E\{\mathrm{var}(Y_n \mid \boldsymbol{X}_n)\boldsymbol{X}_{nj}^2\} > 0$, and $\sup_n E\{|Y_n - m(\boldsymbol{X}_n)|^l\} \leq l!M^l$ for some constant $M > 0$ with $l = 3, 4, 5, \ldots$.

A5. The eigenvalues of H_n and Ω_n are uniformly bounded away from 0.

We define

$$\begin{aligned} w_{\max}^{(\mathrm{I})} &= \max\{w_{n,kj}, k = 1, \ldots, K \text{ and } j = 1, \ldots, s_k\}, \\ w_{\min}^{(\mathrm{II})} &= \min\{w_{n,kj}, k = 1, \ldots, K \text{ and } j = s_k, \ldots, p_k\}, \end{aligned}$$

where $w_{\max}^{(\mathrm{I})}$ is the maximal weight of truly significant coefficients and $w_{\min}^{(\mathrm{II})}$ is the minimal weight of insignificant ones.

2.3.2 Consistency

Theorem 2.3 shows that the adaptive the penalized Group-BD estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD}}$ is $(n/p_n)^{1/2}$ consistent.

Theorem 2.3. *Under assumptions* A1 − A4*, if* $p_n^4/n \to 0$*,* $\kappa_n \to 0$*, and*

$$\kappa_n (w_{\max}^{(\mathrm{I})})^{1/2} = O_p(n^{-1/2}),$$

then there exists a local minimum $\widehat{\boldsymbol{\beta}}_n$ *of* $T_n(\boldsymbol{\beta}_n)$ *such that*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\| = O_p((p_n/n)^{1/2}).$$

Here we need the divergence rate of dimension p_n to be slower compared with n (i.e. $p_n^4/n \to 0$). A natural consideration is whether we could extend the result to cases with higher dimensions. It turns out that if we post constrains on both $w_{\max}^{(\mathrm{I})}$ and $w_{\min}^{(\mathrm{II})}$, we can relax p_n to a higher rate. If we also assume relatively high weights for insignificant variables, they will be more likely shrunk to zero, which helps to achieve the consistency of the entire estimator.

Theorem 2.4. *Under assumptions* A1 − A5*, suppose* $K^3/ns_n^2 = O(1)$*,* $s_n^4/n \to 0$*,* $s_n(p_n - s_n) = O(n)$*,* $(n/s_n)^{1/2} \min_{\beta_{nj}^0 \neq 0} |\beta_{nj}^0| \to \infty$*,* $\kappa_n \to 0$*,* $\kappa_n^{-1}(w_{\min}^{(\mathrm{II})})^{-1/2} = o_P(1)$ *and* $\kappa_n(w_{\max}^{(\mathrm{I})})^{1/2} = O_P(1/n)$*, then there exists a local minimum* $\widehat{\boldsymbol{\beta}}_n$ *of* $T_n(\boldsymbol{\beta}_n)$ *such that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\| = O_p(\sqrt{s_n/n})$.

In sparse cases, usually s_n is much smaller than p_n, so the result of Theorem 2.4 shows a much faster rate of consistency and the diverging rate of p_n is relaxed to be comparable with n.

2.3.3 Oracle Properties

Theorem 2.5. *Assume assumptions* A1 − A5*.*

(I). *For any* $(n/s_n)^{1/2}$ *local minimum* $\widehat{\boldsymbol{\beta}}_n$ *satisfies*

$$\Pr(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{II})} = \mathbf{0}) \to 1.$$

(II). *Moreover, assume that* $s_n^5/n \to 0$ *and that the eigenvalues of* Ω_n *are uniformly bounded away from 0, then we have*

$$\sqrt{n}A_n\Omega_n^{-1/2}H_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \to N(0, G)$$

for any $s \times (s_n + 1)$ *matrix* A_n *such that* $A_n A_n^T \to G$ *with* G *being an* $s \times s$ *semi-positive definite matrix.*

Theorem 2.5 shows that the penalized Group-BD estimator could achieve the oracle properties that insignificant parameters are estimated to be exactly zero with probability tending to 1 and estimators for significant parameters have the asymptotic normal distribution with the same means and variances as if the zero coefficients were known in advance (Zhang et al., 2010). Also the asymptotic normality does not require on the specific distribution of $Y_n \mid \boldsymbol{X}_n$, which is desirable when there is no adequate information about the distribution.

2.3.4 Hypothesis Testing

We consider the following hypothesis testing about significant coefficients $\boldsymbol{\beta}_n^{0(\mathrm{I})}$:

$$H_0: \ A_n\boldsymbol{\beta}_n^{0(\mathrm{I})} = \mathbf{0}, \quad \text{versus} \quad H_1: \ A_n\boldsymbol{\beta}_n^{0(\mathrm{I})} \neq \mathbf{0}, \tag{2.5}$$

where A_n is a given $s \times (s_n + 1)$ matrix such that $A_n A_n^T = G$, and G is a $s \times s$ positive definite matrix. This form of hypotheses (2.5) can be adjusted to various simultaneous tests about linear combinations of parameters.

According to Theorem 2.5, the Group-BD estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD(I)}}$ has the asymptotic covariance matrix $V_n = H_n^{-1}\Omega_n H_n^{-1}$. Define

$$\begin{aligned}\widehat{H}_n &= \frac{1}{n}\sum_{i=1}^{n} q_2(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD(I)}})\boldsymbol{X}_{ni}^{(\mathrm{I})}\boldsymbol{X}_{ni}^{(\mathrm{I})T},\\ \widehat{\Omega}_n &= \frac{1}{n}\sum_{i=1}^{n} q_1^2(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD(I)}})\boldsymbol{X}_{ni}^{(\mathrm{I})}\boldsymbol{X}_{ni}^{(\mathrm{I})T}.\end{aligned}$$

Then we can get

$$\widehat{V}_n = \widehat{H}_n^{-1}\widehat{\Omega}_n\widehat{H}_n^{-1}.$$

Corollary 2.6. *Under assumptions* A1 – A5, $\widehat{V}_n$ *is a consistent estimator of* V_n *in the sense that* $\|\widehat{V}_n - V_n\| \to 0$ *in probability.*

Based on the estimator of the covariance matrix, we considered a generalized Wald-type test statistic (Zhang et al., 2010):

$$W_n = n(A_n\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD(I)}})^T(A_n\widehat{H}_n^{-1}\widehat{\Omega}\widehat{H}_n^{-1}A_n^T)^{-1}(A_n\widehat{\boldsymbol{\beta}}_n^{\mathrm{gBD(I)}}). \tag{2.6}$$

Under the null hypothesis, W_n is asymptotically χ^2 distributed as stated in Theorem 2.7.

Theorem 2.7. *Under assumptions* A1 – A5, *if* $s_n^5/n \to 0$, *then under* H_0 *in* (2.5),

$$W_n \xrightarrow{d} \chi_s^2. \tag{2.7}$$

It can also be shown that W_n has the asymptotic power 1, since $W_n \to \infty$ in probability at the rate n against any fixed alternatives.

2.4 Implementation

Recently, the Coordinate Descent (CD) algorithm was widely applied to the penalized linear and logistic regression (Friedman et al., 2007, 2010; Wu and Lange, 2008) and showed computational superiority especially in sparse cases. Here we apply this algorithm to the penalized Group-BD model.

According to (2.1), when d_k is fixed, it becomes a Lasso problem and when α_{kj} is fixed, it is a non-negative garrote problem (Zhou and Zhu, 2010). Define $q_j(y,\theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$, $j = 0, 1, \ldots$. The BD loss function can be approximated by

$$\frac{1}{2}\sum_{i=1}^{n} t_i(Z_i - \beta_0 - x_i^T\boldsymbol{\beta})^2,$$

where $t_i = q_2(y_i, \beta_0 + x_i^T\boldsymbol{\beta})/n$ and $Z_i = (\beta_0 + x_i^T\boldsymbol{\beta}) - q_1(y_i, \beta_0 + x_i^T\boldsymbol{\beta})/q_2(y_i, \beta_0 + x_i^T\boldsymbol{\beta})$ (Zhang et al., 2011).

Thus the algorithm is summarized as follows:

1. Initialize d_k, α_{kj} and $\beta_{kj} = d_k\alpha_{kj}$;

2. Compute Z_i and t_i as described above. Set $\widetilde{Z}_i = Z_i - (\sum_{i=1}^n t_i Z_i)/(\sum_{i=1}^n t_i)$ and $\widetilde{x}_i = x_i - (\sum_{i=1}^n t_i x_i)/(\sum_{i=1}^n t_i)$ to eliminate the intercept in the model. Now the problem is

$$\min_{d_k,\alpha_{kj}} \left\{ \frac{1}{2}\sum_{i=1}^{n} t_i(\widetilde{Z}_i - \widetilde{x}_i^T\boldsymbol{\beta})^2 + \sum_{k=1}^{K} d_k + \kappa\sum_{k=1}^{K} w_{kj}\sum_{j=1}^{p_k}|\alpha_{kj}| \right\}.$$

3. Let $x_{i,kj}^d = d_k x_{i,kj}$, $k = 1, \ldots, K$, $j = 1, \ldots, p_k$.

Inner loop for α_{kj}:

a) Update α_{kj} by

$$\alpha_{kj}^{\text{new}} = \arg\min_{\alpha_{kj}} \left\{ \frac{1}{2} \sum_{i=1}^{n} t_i \Big(\widetilde{Z}_i - \sum_{k,j} \widetilde{x}_{i,kj}^{d} \alpha_{kj} \Big)^2 + \kappa \sum_{k=1}^{K} w_{kj} \sum_{j=1}^{p_k} |\alpha_{kj}| \right\}.$$

b) If $\max(|\alpha_{kj}^{\text{new}} - \alpha_{kj}|)$ is small enough, let $\alpha_{kj} = \alpha_{kj}^{\text{new}}$ and stop. Otherwise, let $\alpha_{kj} = \alpha_{kj}^{\text{new}}$ and go back to Step 3(a).

4. Let $\widetilde{x}_{i,k}^{\alpha} = \sum_{j=1}^{p_k} \alpha_{kj} x_{i,kj}$, $k = 1, \ldots, K$.

   Inner loop for d_k:

   a) Update d_k by

$$d_k^{\text{new}} = \arg\min_{d_k \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{n} t_i \Big(\widetilde{Z}_i - \sum_{k=1}^{K} \widetilde{x}_{i,k}^{\alpha} d_k \Big)^2 + \sum_{k=1}^{K} d_k \right\}.$$

   b) If $\max(|d_k^{\text{new}} - d_k|)$ is small enough, $d_k = d_k^{\text{new}}$ and stop. Otherwise, let $d_k = d_k^{\text{new}}$ and go back to Step 4(a)

5. Let $\beta_{kj}^{\text{new}} = d_k \alpha_{kj}$. If $\|\boldsymbol{\beta}^{\text{new}} - \boldsymbol{\beta}\|_1$ is small enough, stop. Otherwise, let $\boldsymbol{\beta} = \boldsymbol{\beta}^{\text{new}}$ and go back to Step 2.

See Appendix 2 for the main matlab codes of this algorithm.

2.5 Simulation

In this section we show that under different data types, the group methods maintain the advantage when variables are generated in groups by comparing Ridge, SCAD, Lasso, and adaptive Lasso under BD loss together with the penalized Group-BD and the adaptive penalized Group-BD with ridge regression weights, and also oracle method for which consider indexes of significant variables are known. We also show that a specific choice of the BD loss doesn't effect the results significantly.

The response variables are generated from the following regression model:

$$E(Y_i \mid \boldsymbol{X}_i) = F^{-1}(\theta_i)$$

where $\theta_i = \beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}$ and F is a known link function.

For the explanatory variables, groups of size 4 or 3 are generated independently. Within each group, covariates are normal with mean 0, variance 1 and correlation coefficient $\rho = 0.5$. $\boldsymbol{X}_k = (X_{k,1}, \ldots, X_{k,p_k})^T \sim N(\mathbf{0}, \Sigma_{p_k})$ for $k = 1, \ldots, K$ where $\Sigma_{p_k}(j,j) = 1$ and $\Sigma_{p_k}(j,l) = \rho$ for $j \neq l$ and $j,\ l = 1, \ldots, p_k$.

For the purpose of sparsity, we consider only 3 significant groups while other groups have coefficients 0. The following true parameter values are considered:(2, 0, 0, -2.5), (2.5, 0, 0, -2), (2, 0, 1.5). Thus we have the true parameter vector

$$\boldsymbol{\beta}_{\text{true}} = (2, 0, 0, -2.5, 2.5, 0, 0, -2, 2, 0, 1.5, 0, \ldots, 0)^T.$$

To compare methods under different dimensionality, we considered three different

combination of sample size n and number of parameters p:

- Low-dimensional case: $n = 400$, $p = 56$;
- High-dimensional case: $n = 200$, $p = 200$;
- Ultra-high-dimensional case: $n = 350$, $p = 2600$.

To select the tuning parameter for methods with penalization, another n observations are generated as the tuning observations for grid search. And 10,000 testing observations are used for the comparison results. All the results are the medians of 100 simulation runs.

2.5.1 Over-dispersed Poisson Responses

First, we consider over-dispersed Poisson responses. Y_i's are counts satisfying $\text{var}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = 2m(\boldsymbol{x}_i)$ where $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ with a logarithm link $F(\theta) = \log(\theta)$. The negative quasi-likelihood is utilized as the BD loss.

Table 2.1 evaluates different methods through estimation accuracy and ability to discover sparsity at both the group and individual variable level. Some observations can be made. First, in estimation accuracy the penalized Group-BD and the adaptive penalized Group-BD achieve lowest L_1 error. Second, while all methods perform well in discovering significant variables, the penalized Group-BD and the adaptive penalized Group-BD achieve best in discovering insignificant groups and variables. And the low standard deviation shows stability of the two methods. Also as the dimension increases, the performance of the penalized Group-BD and the adaptive

Table 2.1: *Simulation results for over-dispersed Poisson responses under quasi-likelihood loss*

(n,p)	Method	$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{\text{true}}\|_1$	CZ	CNZ	CG	WG
(400,56)	Oracle	0.7098 (0.26)	50.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.6592 (0.16)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	13.00 (0.0)
	SCAD	1.1478 (0.47)	44.00 (4.5)	6.00 (0.0)	3.00 (0.0)	5.00 (2.7)
	Lasso	2.1523 (0.89)	38.00 (6.7)	6.00 (0.0)	3.00 (0.0)	8.00 (2.8)
	ALasso	1.2218 (0.58)	46.00 (4.1)	6.00 (0.0)	3.00 (0.0)	3.00 (2.6)
	GLasso	4.2969 (2.13)	47.00 (6.4)	6.00 (0.3)	3.00 (0.1)	1.00 (2.4)
	gBD	1.2295 (0.91)	48.00 (2.7)	6.00 (0.2)	3.00 (0.1)	0.00 (0.8)
	AgBD	1.0297 (0.71)	49.00 (2.0)	6.00 (0.2)	3.00 (0.1)	0.00 (0.6)
(200,200)	Oracle	1.0201 (0.41)	194.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.9642 (0.32)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	3.3035 (1.16)	175.00 (7.8)	6.00 (0.0)	3.00 (0.0)	16.00 (5.8)
	Lasso	5.4964 (2.28)	172.00 (11.9)	6.00 (0.0)	3.00 (0.0)	18.00 (7.6)
	ALasso	3.6827 (1.68)	184.00 (7.3)	6.00 (0.0)	3.00 (0.0)	9.00 (5.7)
	GLasso	7.0201 (2.49)	170.50 (21.5)	6.00 (0.8)	3.00 (0.3)	9.00 (9.0)
	gBD	2.0109 (1.66)	191.00 (3.3)	6.00 (0.4)	3.00 (0.2)	0.00 (1.2)
	AgBD	1.7069 (1.37)	192.00 (2.3)	6.00 (0.4)	3.00 (0.2)	0.00 (0.8)
(350,2600)	Oracle	0.7210 (0.31)	2594.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.9116 (0.31)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	3.6715 (1.36)	2547.00 (18.2)	6.00 (0.0)	3.00 (0.0)	44.00 (16.4)
	Lasso	6.1387 (1.80)	2556.00 (15.5)	6.00 (0.0)	3.00 (0.0)	36.00 (14.0)
	ALasso	4.4927 (2.06)	2574.00 (12.6)	6.00 (0.2)	3.00 (0.0)	19.00 (11.9)
	GLasso	10.3222 (1.654)	2588.00 (7.7)	4.00 (1.8)	3.00 (0.3)	0.00 (3.3)
	gBD	1.0831 (1.33)	2593.00 (1.5)	6.00 (0.4)	3.00 (0.2)	0.00 (0.8)
	AgBD	0.9747 (0.51)	2594.00 (1.2)	6.00 (0.0)	3.00 (0.0)	0.00 (0.7)

ALasso: adaptive Lasso.; gBD: the penalized Group-BD; AgBD: the adaptive penalized Group-BD; CZ: number of correctly discovered zero coefficients; CNZ: number of correctly selected non-zero coefficients; CG: number of correctly selected significant groups. WG: number of falsely discovered groups. Values in parentheses are standard deviations.

penalized Group-BD remain stable while others' are weaken obviously, which shows the superiority of the group penalty under high dimensionality.

To compare the methods under different level of group-structures we also consider various level of within-group correlation ρ. A illustration of estimation error of adaptive Group-BD and SCAD is shown in Figure 2.3. Since higher correlation indicates a stronger group structure, the adaptive Group-BD has more advantage over non-group method SCAD.

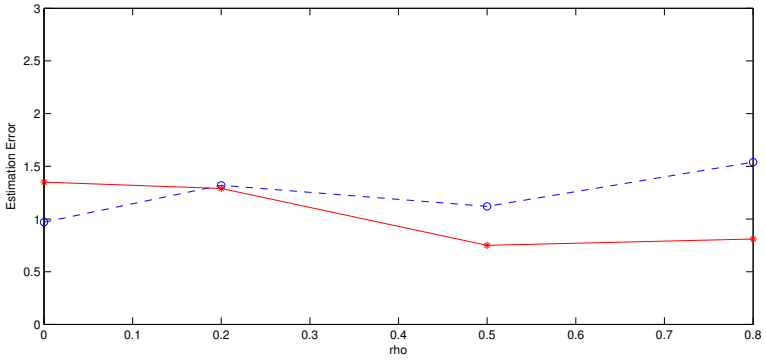

Figure 2.3: *Estimation error under different levels of group structure. Red solid line: SCAD; Blue dashed line: Adaptive Group BD*

2.5.2 Zero-inflated Poisson Responses

Here we consider the zero-inflated Poisson distribution with the distribution function $P(Y=0) = \pi + (1-\pi)e^{-\lambda}$ and $P(Y=j) = (1-\pi)\lambda^j e^{-\lambda}/j!,\ j = 1, 2, \ldots$. We choose $\pi = 0.1$. Then we compare the results using the quasi-likelihood loss pretending the distribution is unknown with those using exact negative log-likelihood loss. From Table 2.2 we can see that the BD loss provides similar results, which means that even

if we don't know the exact distribution, a properly chosen BD loss can perform just as well.

2.5.3 Bernoulli Responses

To consider Bernoulli responses, we use the logit link $F(\theta) = \log\{\theta/(1-\theta)\}$. Both the Bernoulli deviance loss $Q(Y,\mu) = -2\{Y\log(\mu) + (1-Y)\log(1-\mu)\}$ and the Exponential loss $Q(Y,\mu) = \exp\{-(Y-0.5)\log[\mu/(1-\mu)]\}$ are considered as the BD loss. And we compare different methods by the misclassification rate.

From Table 2.3 and Table 2.4, we can see that with more limited information from the binary responses, the advantage of the penalized Group BD and the adaptive penalized Group BD in misclassification rate might not be as significant, but they still outperform others and show obvious advantage in discovering insignificant variables. Also by comparing two tables we can observe that the two loss function, both belong to the BD family, give very similar results.

In some cases, the adaptive method (for either Lasso or the penalized Group-BD) might be slightly worse than the unadapted method. That might be caused by unstable performance of Ridge regression which is used as the adaptive weights under high dimension. If a more sophisticated method is used for the weights, the performance might be improved. However it violates the convenience purpose of using adaptive methods, and under high dimension the unadapted penalized Group-BD method already outperforms other methods and has performance close to oracle results.

Table 2.2: *Simulation results for Zero-inflated Poisson responses with both negative log-likelihood loss and quasi-likelihood loss*

(n,p)	Method	$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{\text{true}}\|_1$	CZ	CNZ	CG	WG
Negative-likelihood loss						
(400,56)	Oracle	1.1422 (0.63)	50.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.6568 (0.12)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	13.00 (0.0)
	SCAD	2.1357 (1.17)	42.00 (5.9)	6.00 (0.0)	3.00 (0.0)	6.00 (3.2)
	Lasso	3.7398 (1.73)	33.00 (8.1)	6.00 (0.0)	3.00 (0.0)	10.00 (2.6)
	ALasso	2.2829 (1.45)	44.00 (6.1)	6.00 (0.0)	3.00 (0.0)	5.00 (3.2)
	GLasso	9.5152 (2.95)	49.00 (4.8)	3.50 (1.8)	2.00 (0.9)	0.00 (2.0)
	gBD	2.0912 (1.37)	47.00 (4.2)	6.00 (0.1)	3.00 (0.1)	0.00 (1.3)
	AgBD	1.7783 (1.26)	48.00 (3.5)	6.00 (0.1)	3.00 (0.1)	0.00 (1.1)
(200,200)	Oracle	1.3947 (0.79)	194.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.9691 (0.28)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	4.4003 (1.92)	173.00 (9.6)	6.00 (0.0)	3.00 (0.0)	18.00 (7.0)
	Lasso	7.3767 (3.20)	164.50 (15.0)	6.00 (0.1)	3.00 (0.0)	23.00 (8.3)
	ALasso	4.8948 (2.54)	180.00 (10.0)	6.00 (0.1)	3.00 (0.0)	12.00 (7.1)
	GLasso	9.4098 (2.47)	190.00 (19.2)	5.00 (1.8)	3.00 (0.9)	1.00 (8.1)
	gBD	2.9411 (2.46)	190.00 (5.8)	6.00 (0.3)	3.00 (0.1)	0.00 (2.0)
	AgBD	2.4578 (1.99)	191.00 (4.2)	6.00 (0.2)	3.00 (0.1)	0.00 (1.5)
(350,2600)	Oracle	1.9056 (0.52)	2594.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	14.3080 (0.03)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	649.00 (0.0)
	SCAD	10.1400 (2.32)	2518.00 (23.0)	6.00 (0.0)	3.00 (0.0)	68.00 (21.1)
	Lasso	13.9591 (2.54)	2527.00 (17.0)	6.00 (0.0)	3.00 (0.0)	61.00 (16.0)
	ALasso	12.7141 (4.16)	2548.00 (15.8)	6.00 (0.0)	3.00 (0.0)	40.00 (14.4)
	GLasso	12.9657 (3.95)	2584.50 (11.2)	4.00 (1.7)	0.50 (1.0)	1.00 (1.3)
	gBD	6.7392 (1.29)	2588.00 (4.3)	6.00 (0.0)	3.00 (0.0)	1.00 (2.2)
	AgBD	6.2040 (2.49)	2588.00 (4.5)	6.00 (0.0)	3.00 (0.0)	1.00 (3.6)
Quasi-likelihood loss						
(400,56)	Oracle	0.8214 (0.44)	50.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.6535 (0.13)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	13.00 (0.0)
	SCAD	1.2833 (0.76)	44.00 (5.0)	6.00 (0.0)	3.00 (0.0)	5.00 (3.0)
	Lasso	2.6912 (1.28)	36.50 (7.6)	6.00 (0.0)	3.00 (0.0)	9.00 (3.0)
	ALasso	1.5623 (0.90)	46.00 (4.6)	6.00 (0.0)	3.00 (0.0)	3.00 (2.9)
	GLasso	4.3636 (1.89)	45.00 (8.1)	6.00 (0.3)	3.00 (0.1)	2.00 (3.1)
	gBD	1.4192 (1.03)	48.00 (3.3)	6.00 (0.1)	3.00 (0.1)	0.00 (1.0)
	AgBD	1.1598 (0.90)	49.00 (2.6)	6.00 (0.1)	3.00 (0.1)	0.00 (0.8
(200,200)	Oracle	1.1323 (0.57)	194.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	12.9331 (0.28)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	3.1013 (1.49)	176.00 (9.2)	6.00 (0.0)	3.00 (0.0)	15.50 (6.8)
	Lasso	6.0943 (2.42)	168.00 (12.7)	6.00 (0.0)	3.00 (0.0)	21.00 (7.9)
	ALasso	4.0678 (1.93)	182.00 (8.5)	6.00 (0.1)	3.00 (0.0)	11.00 (6.6)
	GLasso	6.9529 (2.73)	165.00 (24.2)	6.00 (0.8)	3.00 (0.4)	12.00 (9.8)
	gBD	2.3385 (1.96)	191.00 (4.3)	6.00 (0.4)	3.00 (0.2)	0.00 (1.5)
	AgBD	1.9492 (1.68)	192.00 (3.5)	6.00 (0.4)	3.00 (0.2)	0.00 (1.3)
(350,2600)	Oracle	1.1702 (0.36)	2594.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	14.2841 (0.04)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	649.00 (0.0)
	SCAD	9.8301 (2.22)	2533.00 (22.7)	6.00 (0.0)	3.00 (0.0)	56.00 (19.4)
	Lasso	12.3820 (1.84)	2548.00 (15.2)	6.00 (0.7)	3.00 (0.0)	43.50 (14.4)
	ALasso	11.2678 (1.65)	2552.50 (15.3)	6.00 (0.7)	3.00 (0.0)	37.50 (14.6)
	GLasso	12.1504 (2.90)	2587.00 (10.3)	4.50 (1.9)	0.50 (0.3)	1.00 (1.1)
	gBD	7.1955 (0.88)	2590.50 (3.6)	6.00 (0.0)	3.00 (0.0)	1.00 (1.9)
	AgBD	7.2277 (1.37)	2591.00 (5.6)	6.00 (0.0)	3.00 (0.0)	1.00 (3.7)

Table 2.3: *Simulation results for Bernoulli responses with Deviance Loss*

(n, p)	Method	MR	CZ	CNZ	CG	WG
(400,56)	Oracle	0.1267 (0.00)	50.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.2471 (0.01)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	13.00 (0.0)
	SCAD	0.1273 (0.01)	48.00 (4.0)	6.00 (0.2)	3.00 (0.0)	2.00 (2.2)
	Lasso	0.1345 (0.01)	42.00 (9.8)	6.00 (0.0)	3.00 (0.0)	6.00 (4.1)
	ALasso	0.1311 (0.01)	48.00 (7.5)	6.00 (0.0)	3.00 (0.0)	2.00 (3.8)
	GLasso	0.1336(0.01)	37.00(11.5)	6.00 (0.0)	3.00 (0.0)	5.00(4.1)
	gBD	0.1289 (0.00)	49.00 (2.8)	6.00 (0.0)	3.00 (0.0)	0.00 (1.2)
	AgBD	0.1285 (0.00)	49.00 (2.2)	6.00 (0.0)	3.00 (0.0)	0.00 (1.1)
(200,200)	Oracle	0.1431 (0.01)	194.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.3165 (0.01)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	0.1983 (0.03)	190.00 (4.4)	5.00 (1.6)	3.00 (0.7)	4.00 (4.3)
	Lasso	0.2081 (0.02)	168.00 (20.8)	6.00 (1.4)	3.00 (0.6)	22.00 (14.6)
	ALasso	0.1988 (0.02)	178.00 (14.0)	6.00 (1.2)	3.00 (0.5)	14.00 (11.1)
	GLasso	0.2053(0.02)	142.00(31.7)	6.00 (1.4)	3.00 (0.6)	20.00(13.1)
	gBD	0.1568 (0.02)	192.00 (2.6)	6.00 (1.0)	3.00 (0.4)	1.00 (1.6)
	AgBD	0.1551 (0.02)	193.00 (1.7)	6.00 (0.9)	3.00 (0.4)	0.00 (1.2)
(350,2600)	Oracle	0.1360 (0.00)	2594.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.5058 (0.02)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	649.00 (0.0)
	SCAD	0.1870 (0.03)	2590.00 (4.5)	5.00 (1.6)	3.00 (0.6)	4.00 (4.4)
	Lasso	0.2152 (0.01)	2581.00 (34.7)	4.00 (0.9)	3.00 (0.6)	11.50 (31.1)
	ALasso	0.2106 (0.01)	2584.00 (23.8)	4.00 (0.9)	3.00 (0.6)	10.00 (21.8)
	GLasso	0.1764(0.02)	2547.00(38.5)	6.00 (0.0)	3.00 (0.0)	12.50(17.4)
	gBD	0.1425 (0.02)	2592.00 (5.2)	6.00 (0.6)	3.00 (0.3)	0.00 (3.1)
	AgBD	0.1455 (0.02)	2592.00 (3.1)	6.00 (0.8)	3.00 (0.3)	0.00 (2.0)

MR: misclassification rate.

2.6 Real Data

In this section, we applied the adaptive penalized Group-BD method and others for comparison to various data sets.

2.6.1 LSVT Data

Vocal impairment is reported in the vast majority of Parkinson's disease subjects. The extent of vocal impairment can be assessed using sustained vowel phonations

Table 2.4: *Simulation results for Bernoulli responses with Exponential Loss*

(n, p)	Method	MR	CZ	CNZ	CG	WG
(400,56)	Oracle	0.1276 (0.00)	50.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.2474 (0.01)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	13.00 (0.0)
	SCAD	0.1286 (0.01)	50.00 (3.7)	6.00 (0.0)	3.00 (0.0)	2.00 (2.2)
	Lasso	0.1348 (0.01)	42.00 (9.5)	6.00 (0.0)	3.00 (0.0)	6.00 (4.1)
	ALasso	0.1316 (0.01)	48.00 (7.0)	6.00 (0.0)	3.00 (0.0)	2.00 (3.7)
	GLasso	0.1342(0.01)	37.00(11.6)	6.00 (0.0)	3.00 (0.0)	5.00(4.1)
	gBD	0.1290 (0.01)	49.00 (2.9)	6.00 (0.1)	3.00 (0.0)	0.00 (1.3)
	AgBD	0.1287 (0.01)	50.00 (3.0)	6.00 (0.0)	3.00 (0.0)	0.00 (1.4)
(200,200)	Oracle	0.1441 (0.01)	194.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.3167 (0.01)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	49.00 (0.0)
	SCAD	0.2112 (0.03)	190.00 (4.9)	4.00 (1.7)	3.00 (0.8)	4.00 (4.8)
	Lasso	0.2077 (0.02)	169.00 (19.8)	6.00 (1.4)	3.00 (0.6)	21.00 (14.2)
	ALasso	0.2008 (0.02)	179.00 (14.0)	6.00 (1.2)	3.00 (0.5)	14.00 (11.1)
	GLasso	0.2100(0.02)	148.00(34.4)	6.00 (1.4)	3.00 (0.7)	17.00(13.6)
	gBD	0.1543 (0.02)	193.00 (2.0)	6.00 (1.0)	3.00 (0.4)	0.00 (1.4)
	AgBD	0.1574 (0.03)	193.00 (1.7)	6.00 (1.1)	3.00 (0.5)	0.00 (1.3)
(350,2600)	Oracle	0.1368 (0.00)	2594.00 (0.0)	6.00 (0.0)	3.00 (0.0)	0.00 (0.0)
	Ridge	0.5058 (0.02)	0.00 (0.0)	6.00 (0.0)	3.00 (0.0)	649.00 (0.0)
	SCAD	0.2059 (0.03)	2590.00 (5.5)	4.00 (1.7)	3.00 (0.8)	4.00 (5.5)
	Lasso	0.2126 (0.01)	2577.00 (37.9)	4.00 (1.0)	3.00 (0.7)	16.00 (34.0)
	ALasso	0.1996 (0.02)	2565.50 (30.7)	5.00 (1.0)	3.00 (0.2)	27.50 (28.4)
	GLasso	0.1702(0.02)	2552.00(34.8)	6.00 (0.0)	3.00 (0.0)	10.00(14.1)
	gBD	0.1410 (0.02)	2593.00 (2.4)	6.00 (0.7)	3.00 (0.3)	0.00 (1.7)
	AgBD	0.1407 (0.01)	2593.00 (1.4)	6.00 (0.6)	3.00 (0.2)	0.00 (1.0)

(Tsanas et al., 2014). To extract information from vocal signals, various types of dysphonia measures, which are clinical speech signal processing algorithms, are used. In Lee Silverman Voice Treatment (LSVT) program, participants produced sustained vowel /a/ phonations and vocal signals are quantified through dysphonia measures (see Tsanas et al. (2014) for details). LSVT expert clinicians assessed sustained vowel phonations perceptually whether they are "acceptable" or "unacceptable". The aim is to automatically evaluate the phonations based on the dysphonia measures and discover which measures are most informative in assessment. Current analysis

related to variable selection is based on subset selection through cross validation which is already shown to be surpassed by regularization method in the literature and the corresponding result is difficult to interpret. Instead, we apply the adaptive Group-BD method compared with others to give some new investigation of this study.

There are 126 samples with 309 dysphonia measures in 11 types which act as the groups. In the pre-process of the data, we eliminate variables with negligible variation which leaves 10 groups and 243 variables. Then we use the ten-fold cross validation to select tuning parameters and bootstrap sampling to test the results. The misclassification rates on the test data are 0.048, 0.125, and 0.016 for SCAD, adaptive Lasso and the adaptive penalized Group-BD respectively. Among them, adaptive Lasso chooses 20 variables of 6 groups, SCAD chooses 5 variables of 4 groups, and the penalized Group-BD chooses 14 variables of 2 groups. We can see that compared with adaptive Lasso, adaptive penalized Group-BD gives lower misclassification rate with a more sparse model. While the disadvantage of SCAD in misclassification rate may be caused by a even more sparse model, the selected variables distribute out into 4 groups. By contrast, although the penalized Group-BD selects more variables, they focus on two groups which implies that in practice only a few types of dysphonia measures are needed to achieve acceptable assessment accuracy. Among the 11 groups, the signal-to-noise ratio (SNR) type and Mel frequency cepstral coefficients (MFCC) type are selected. SNR measures quantify excessive noise in the phonation suggesting that incomplete vocal fold closure and vocal tremor may be the primary characteristic that clinicians use to perceptually assess whether a phonation is “acceptable” or “unacceptable”. MFCC measures target the placement of the articulators, which is

known to be affected in Parkinson's disease (Tsanas et al., 2014). The result shows that in future phonation assessment, measurements of these two types could provide good prediction performance without aggregating data from all 11 types.

2.6.2 Urban Land Cover Data

This data set contains urban land cover information extracted from high resolution aerial or satellite imagery. The high resolution aerial image is classified into 9 types of urban land cover and multi-scale spectral, size, shape, and texture information are collected. More details can be found in Johnson and Xie (2013) and Johnson (2013). For the purpose of variable selection, we build a multinomial model to discover which are significant for the land cover type. Since Kullback-Leibler divergence also belongs to the BD family, we use it as the loss function for the multinomial response:

$$T(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\Big[-\sum_{k=1}^{K}\mathrm{I}(Y_i = k)\log(\mu_{ik})\Big] + \kappa\Big(\sum_{k=1}^{K-1}\sqrt{||\boldsymbol{\beta}_k||_1}\Big)$$

where $K = 9$ as the number of land cover types, and $\mu_{ij} = \exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)/(1 + \sum_{k=1}^{K-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_k))$ for $j = 1, \ldots, K-1$ and $\mu_K = 1/(1 + \sum_{k=1}^{K-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_k))$, which indicates that the Kth type is used as baseline type.

We average each variable on different coarser scales which results in 168 variables (21 for each type). There are 168 cases included in the data set. We randomly split them 100 times into a training set of 112 samples and a validation set of 56 samples. For each splitting variables selected are recorded. Consider only those that are selected more than 50 times among the 100, we have the following results: From

Table 2.5: *(Urban Land Cover Data) Variable selected for each type of the urban land cover*

Type	Variable selected
car	Standard deviation of Green, Gray-Level Co-occurrence Matrix III
concrete	Compactness,Length/Width, Near Infrared
tree	Gray-Level Co-occurrence Matrix II, Normalized Difference Vegetation Index
building	Gray-Level Co-occurrence Matrix I, Standard deviation of Red
asphalt	Green, Border Index, Gray-Level Co-occurrence Matrix I
grass	Green, Normalized Difference Vegetation Index
shadow	Brightness
soil	Border Index, Green

the results we can observe that for some types such as car, grass and shadow, one kind of spectral, texture and shape variables plays a more important role than other kinds, while for types such as asphalt and tree, spectral, texture and shape variables have mixed effects.

2.6.3 Low Resolution Spectrometer Data

The data set from the Low Resolution Observation (IRAS-LRS) program contains a subset of observations of high intensity sources over two spectral bands between 12 hour and 24 hour right ascension. There are 531 high quality spectra observed. Variables include 44 blue band and 49 red band channels together with astronomical longitude and latitude information, which form 3 nature groups with size 44, 49 and 2 respectively. Spectra are classified by features of the spectrum curves. The classes of spectra are labeled from 00 to 99 where the first digit shows the basic class and the second shows the sub-classes. Here we only consider classification of the basic classes ranging from 0 to 9 and apply BD generation function $q(\mu) = \mu - \mu \log \mu$.

Table 2.6: *(Low Resolution Spectrometer data) Testing error and number of selected variables and groups*

Methods	TE	Var.Selected	Group Selected
Lasso	1.5535 (0.0593)	12.75 (2.3757)	2.42 (0.4960)
SCAD	1.5709 (0.0820)	48.44 (9.7425)	2.75 (0.4352)
gBD	1.5397 (0.0431)	10.41 (2.1932)	2.21 (0.4333)

Results in Table 2.6 show that the penalized Group-BD gives the best result with fewest variables and groups selected.

2.6.4 Agricultural Structure Classification

In this section we apply the penalized Group-BD method on the agricultural structure data of Australia. The purpose is to indicate the agricultural structural change in Australia between 1986-1996 by factors such as demographic, farm financial and household statistics. The agricultural structure to be classified includes 13 categories labeled from 0 to 12.

From the density curve of agricultural structure in Figure 2.4, it can be considered as over-dispersed Poisson variable. Hence we apply the BD generation function $q(\mu) = \mu - \mu \log \mu$. The original data set contains 1330 observations from statistical local areas. After deleting missing values, there remains 1220 areas and 58 attributes. Groups are naturally formed. Variables in the same group are the same measurement or indicator in different years or at different levels. Group sizes vary from 2 to 5. The data set is randomly divided into two parts: test set including 1/3 of all observations and the rest that is used to fit the model. Results based on average of 100 replicates

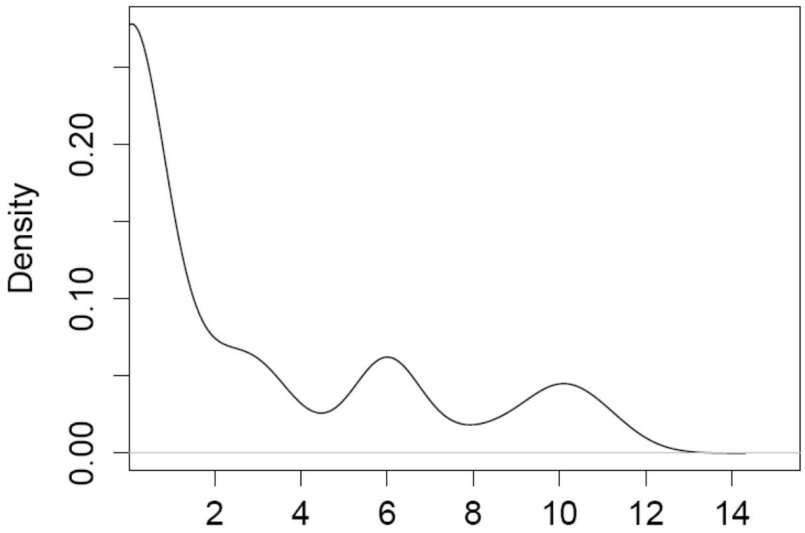

Figure 2.4: *(Agricultural structure data) Density curve.*

are shown in Table 2.7

Table 2.7: *(Agricultural structure data) Testing error and number of selected variables and groups*

Methods	TE	Var.Selected	Group Selected
Lasso	0.6152	20.16	12.41
SCAD	0.6148	9.39	6.94
gBD	0.6085	14.58	7.54

Results show that the penalized Group-BD gives smallest test error. Lasso detects more connections and groups than others but may yield to false discoveries while SCAD yields the most sparse model. The penalized Group-BD should provide a good balance in detecting sparsity. There are 4 groups containing 10 coefficients selected more than 80 times among the 100 replications by all three methods. They are:

farm establishment area indicating the total area of farm establishments, estimated value of median farm in 1986 and 1996, number of farm establishments exceeding an estimated value of agricultural operations of $30,000 and $5,000 in 1986 and 1996, and farm family income. The estimated coefficients are shown in Fig 2.5. We can observe similar patterns among them. Consider other groups in the entire set of results, the penalized Group-BD selects variables in a more stable way. Lasso and SCAD select more different variables in the 100 replicates, usually single ones from various groups.

2.7 Discussion

We have generalized the hierarchical variable selection to a much wider range of data types. Here we no longer need to specify the distribution of the response variable and asymptotic results proved can be applicable to all loss functions that fit into the Bregman Divergence family including many commonly used ones. Numerical results show the advantage of the Hierarchical Lasso under the BD loss for different types of data.

Here we considered known group structures. In the literature, there are already some studies about simultaneously grouping (Bondell and Reich, 2008; Witten et al., 2014). They assume variables in the same group either share similar effect on or association with the response. If it is the case, then an initial grouping information is not need. However, in reality sometimes the group structure comes in a more natural way. When the group structure is based on non-statistical similarities, we need prior

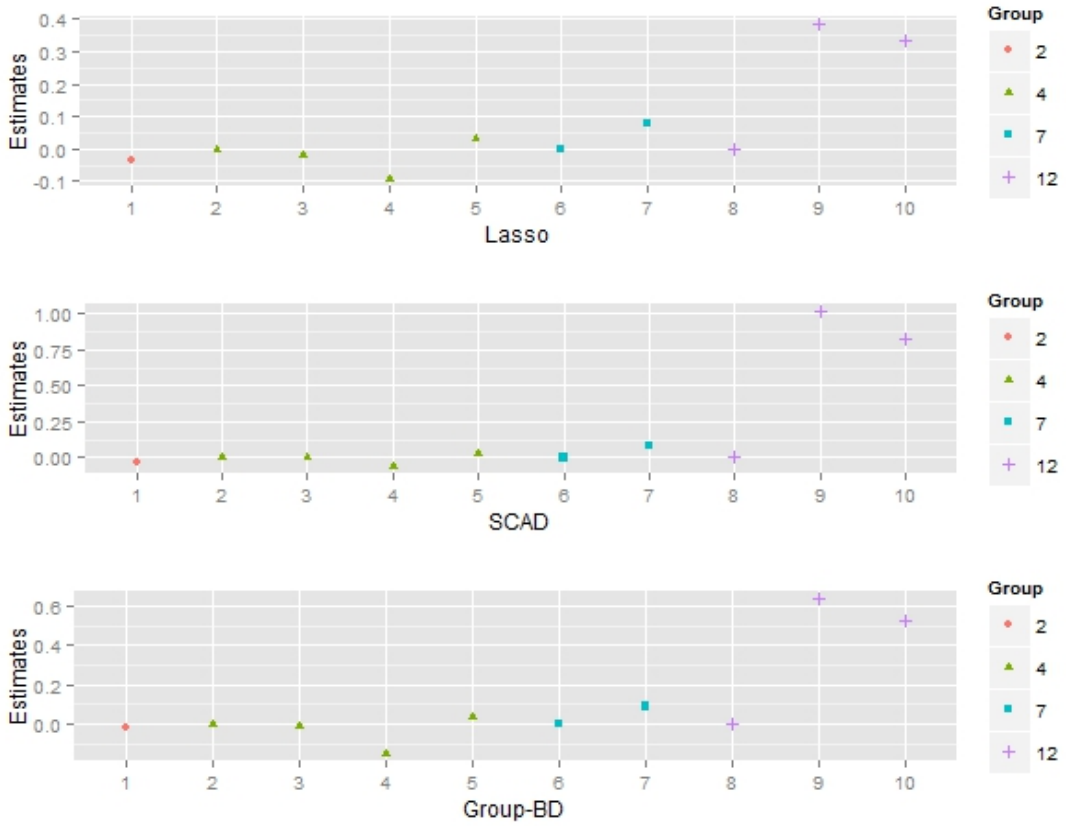

Figure 2.5: *(Agricultural structure data) Estimated coefficients of the 10 coefficients in Group 2, 4, 7 and 12 by Lasso, SCAD and penalized Group-BD.*

information about it to consider group variable selection.

Also we consider group structures without overlapping. However in practice there might be groups sharing the same variables. For example, in gene study some genes may belong to more than one pathways. A natural generalization for overlapping problems is to consider variables in the penalty term for every group they belong to. When the number of shared variables is not diverging with the sample size n, the asymptotic properties still hold for overlapping group structures.

Another possible improvement is the selection of the Bregman Divergence loss. Here we choose the loss based on the data type. For example for the binary response case, both the deviance loss and exponential loss give similar results as illustrated in the Section 2.5. A data driven selection of the BD function $Q(\nu, \mu)$ may be more preferable and less artificial.

Chapter 3

Future Work

3.1 Application to Spike Train Data

The penalized Group-BD method can be applied to structured data in neuronscience. In the study of neuroscience, discovering the interaction network among neurons is an important topic of interest. As electrically excitable cells, neurons send out information as signals under external or internal stimulus. Here we focus on the internal stimulus from other neurons in a certain area of the brain and the interaction among them.

During the experiment process, when a neuron sends out a signal, known as firing a spike, it might inhibit or excite other neurons, which will affect other neurons' firing rates in a certain period of time. A network can be built based on this kind of interaction. While a number of neurons from a certain region of the brain under study might be involved in a designed activity, not every pair of them has non-negligible

effect on each other. Thus sparse modeling is desired to discover significant relations among neurons and eliminate insignificant connections.

For the information to be transmitted, the effect caused by spikes of other neurons might have time lag. The spike rate at a certain time point might be effected by the activity of other neurons in the past. Also, the activity of itself may also effect the firing rate in the future. Thus, we consider both coupling effects from other neurons back up to a certain period of time and autoregressive effect of the target neuron itself.

In the literature, conventional methods such the cross-correlation (Perkel et al., 1967) and joint peri-stimulus time histogram (Gerstein and Perkel, 1969) study a pair or three neurons only at a time, which ignores potential relation with other neurons, and is inefficient when the number of neurons is high. Zhao et al. (2012) introduced the logistic regression with L_1 penalty to discover functional connectivity structure among neurons by reforming the spike train data into discrete binned counts. With L_1 penalty, sparse network can be modeled. However the effects from a neuron itself and other neurons may come with group structure. To further consider the structural information, the Sparse Group Lasso (SGL) penalty (Simon et al., 2013) is used as the penalty by Zhang et al. (2015), together with the Bregman Divergence (BD) as the loss function part to generalize the logistic regression to a regularization method that can be applied to a much wider range of Poisson process models. That inspires us that the penalized Group-BD can also be applied to the spike train data.

Assume there are C neurons under study. For neuron c, the entire process is divided into small time bins. Within each bin, the number of spikes is counted. Let δ

denote the size of each bin and assume we have T_c time bins in total for neuron c. Let τ_k denote the kth bin, and $N^c_{\tau_k}$ denote the number of spikes of neuron c within time bin τ_k, then the conditional intensity (mean number of spikes in that bin) of the kth bin $\lambda_c(\tau_k|\cdot) = E(N^c_{\tau_k}|\cdot)$ is modeled as

$$F(\lambda_c(\tau_k|N^{1:C}_{\tau_{0:k-1}})) = \beta_{c;0} + \sum_{p=1}^{P} \eta_{c;p} N^c_{\tau_{k-p}} + \sum_{i\neq c} \sum_{q=1}^{Q} \gamma_{c;iq} N^i_{\tau_{k-q}}$$

where $F(\cdot)$ is a link function, $F^{-1}(\beta_{c;0})$ is the baseline firing rate within a bin, $\eta_{c;p}$ is the autoregressive coefficient at time lag p, and $\gamma_{c;iq}$ is the coupling effect coefficient of neuron i at time lag q. Thus the model considers both autoregressive effect of neuron c itself up to P time bins into the past and effects from other neurons up to Q bins back.

From the model above we can see that coefficients of the same neuron can form a group. Let

$$\boldsymbol{\theta}_c = (\beta_{c;0}; \{\eta_{c;p}\}_{p=1,\ldots,P}; \{\gamma_{c;iq}\}_{i\neq c; q=1,\ldots,Q})$$

be the parameter vector. Thus the corresponding penalty of the Hierarchical Lasso (HL) is

$$P_\kappa(\boldsymbol{\theta}_c) = \kappa\Big\{\sqrt{\sum_{p=1}^{P} |\eta_{c;p}|} + \sum_{i\neq c} \sqrt{\sum_{q=1}^{Q} |\gamma_{c;iq}|}\Big\}$$

where κ is the tuning parameter.

Considering model for count data, we apply quasi-likelihood function $Q(y,\mu) = y\{\log(y) - \log(\mu)\} - (y-\mu)$ which belongs to the Bregman Divergence loss function family with generating function $q(\mu) = \mu - \mu\log(\mu)$. Combining with HL penalty,

now the regularization problem becomes:

$$T_c(\boldsymbol{\theta}_c) = \sum_{k=1}^{T_c}[N^c_{\tau_k}\{\log(N^c_{\tau_k}) - \log(\lambda_c(\tau_k|\cdot)\} - (N^c_{\tau_k} - \lambda_c(\tau_k|\cdot))] + P_\kappa(\boldsymbol{\theta}_c).$$

3.2 Piecewise-Exponential Likelihood and Lasso Penalty

Regarding the spike train data introduced in the section above, the penalized Group-BD method discretizes the waiting time between spikes. We can also model the waiting time directly as a continuous approach. For each neuron, there might be excitation or inhibition effects from other neurons before the next spike is fired, which will change the firing rate of the target neuron. So instead of the classic exponential distribution, a piecewise exponential distribution introduced in Rajaram et al. (2005) can be incorporated. The rate of the exponential type distribution is piecewise constant.

We can consider the spike train as a Poisson process characterized by rate $\lambda_c(t)$ changing with time. Here if we consider generalized linear model for the effects on $\lambda_c(t)$ of other neurons' firing behavior, with canonical link function, we have

$$\lambda_c(t; \boldsymbol{\beta}_c; \boldsymbol{x}_c) = \exp(\beta_{c,0} + \sum_{j \neq c} \beta_{c,j} x_{c,j}(t)).$$

where $\boldsymbol{\beta}_c = (\beta_{c,0}, \beta_{c,1}, \ldots, \beta_{c,c-1}, \beta_{c,c+1}, \ldots, \beta_{c,C})^T$ is the parameter vector, and $x_{c,j}(t)$

is related to the empirical rate of neuron j in the following way:

$$x_{c,j}(t) = \log(1 + \widehat{\lambda}_{c,j}(t)).$$

$\widehat{\lambda}_{c,j}(t)$ is the empirical rate of neuron j at time t and is defined as $\widehat{\lambda}_{c,j}(t) = n_{c,j}(t)/\phi$, where $n_{c,j}(t)$ is the number of spikes for neuron j during time interval $[t-\phi, t)$. ϕ works as a bandwidth. Since the empirical rate is piecewise constant, the corresponding rate for the target neuron c is also piecewise constant. Thus the waiting time between two spikes follows a piecewise exponential distribution. For the following notation we omit the subscript c for simplicity, while they are still set under the study of one specific target neuron c.

Assume neuron c has N spikes happening at time $\{t_1, \ldots, t_N\}$. Consider the waiting time between the $\ell - 1$th and ℓth spike denoted as τ_l. Let $\widetilde{t}_{l,1}, \ldots, \widetilde{t}_{l,k_l}$ denote the time point when the vector x_j changes and $\tau_{l,1} = \widetilde{t}_{l,1} - t_{l-1}$, $\tau_{l,k_l+1} = t_l - \widetilde{t}_{l,k_l}$, $\tau_{l,i} = \widetilde{t}_{l,i} - \widetilde{t}_{l,i-1}$, $i = 2, \ldots, k_l$. According to Rajaram et al. (2005), the conditional density follows

$$f(t_l|t_{l-1}) = \lambda_{l,k_l+1}(\prod_{i=1}^{k_l+1} \exp(-\lambda_{l,i}\tau_{l,i}))$$

where $\lambda_{l,i} = \exp(\beta_0 + \sum_{j \neq c} \beta_j x_j(\widetilde{t}_{l,i}))$ as defined above.

Here the time trace-back is implicated when generating the empirical rates and the bandwidth ϕ indicates the length of time back. The coupling effect over time is averaged and presented by a single coefficient β_j while the autoregressive effect is simplified and merged into the baseline firing rate β_0.

However not all neurons have significant coupling effect on the target neuron,

which can still be represented by the general form above if some β_j's can be zero. Conventional Bayesian approaches in parameter estimation and structure learning using the piecewise exponential model such as Laplace Approximation and variational approach (Rajaram et al., 2005) put strong assumptions of sparsity over the network structure, otherwise it will be infeasible to optimize over all possible networks. To achieve sparse modeling in a more efficient and flexible way, we may consider regularization. Combining the log-likelihood based on the piecewise exponential density and Lasso penalty, we have the estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{PEL} &= \arg\min_{\boldsymbol{\beta}} T(\boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{\tau}) \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{l=1}^{N} [\log(\lambda_{l,k_l+1}) - \sum_{i=1}^{k_l+1} \lambda_{l,i}\tau_{l,i}] + \kappa||\boldsymbol{\beta}||_1\end{aligned}$$

where κ is the tuning parameter.

Appendix A

Proofs in Chapter 2

For simplicity in notation, we write $q_j(y,\theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$, and

$$\begin{aligned} q_1(y,\theta) &= (y-\mu)q''(\mu)/F'(\mu), \\ q_2(y,\theta) &= -q''(\mu)/F'(\mu)^2 + (y-\mu)A(\mu), \end{aligned}$$

where $A(\mu) = \{q^{(3)}(\mu)F'(\mu) - q''(\mu)F''(\mu)\}/F'(\mu)^3$.

We first show Lemma A.1, which will be used in Theorem 2.1.

Lemma A.1. *Suppose* $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ *is a local minimizer of*

$$T(\boldsymbol{d}, \boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, m(\boldsymbol{X}_i)) + \sum_{k=1}^{K} d_k + \kappa \sum_{j=1}^{p_k} |\alpha_{kj}|,$$

and $\widehat{\boldsymbol{\beta}}$ *be the Group-BD estimate related to* $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$*, i.e.* $\widehat{\beta}_{kj} = \widehat{d}_k\widehat{\alpha}_{kj}$*. If* $\widehat{d}_k = 0$*, then* $\widehat{\boldsymbol{\alpha}}_k = 0$*; if* $\widehat{d}_k \neq 0$*, then* $\|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0$ *and* $\widehat{d}_k = (\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\beta}}_k/(\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$.

Proof. If $\widehat{d}_k = 0$, then $\widehat{\alpha}_k = 0$ is quite obvious. Similarly, if $\widehat{\alpha}_k = 0$, then $\widehat{d}_k = 0$. Therefore, if $\widehat{d}_k \neq 0$, then $\widehat{\alpha}_k \neq 0$ and $\|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0$.

We prove $\widehat{d}_k = (\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\beta}}_k/(\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$ for $\widehat{d}_k \neq 0$ by contradiction. Suppose $\exists k'$ such that $\widehat{d}_{k'} \neq 0$ and $\widehat{d}_{k'} = (\kappa\|\widehat{\boldsymbol{\beta}}_{k'}\|_1)^{1/2}$. Let $(\kappa\|\widehat{\boldsymbol{\beta}}_{k'}\|_1)^{1/2}/\widehat{d}_{k'} = c$, then $\widehat{\boldsymbol{\alpha}}_k = c\widehat{\boldsymbol{\beta}}_k/(\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$. Suppose $c > 1$. Let $\widetilde{d}_k = \widehat{d}_k$ and $\widetilde{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\alpha}}_k$ for $k \neq k'$ and $\widetilde{d}_{k'} = \delta'\widehat{d}_{k'}$ and $\widetilde{\boldsymbol{\alpha}}_{k'} = \widehat{\boldsymbol{\alpha}}_{k'}/\delta'$, where δ' satisfies $c > \delta' > 1$ and is very close to 1 such that $\|\widetilde{d}_{k'} - \delta'\widehat{d}_{k'}\|_1 + \|\widetilde{\boldsymbol{\alpha}}_{k'} - \widehat{\boldsymbol{\alpha}}_{k'}\|_1 < \delta$ for some $\delta > 0$.

Then we have

$$\begin{aligned} T(\widetilde{\boldsymbol{d}}, \widetilde{\boldsymbol{\alpha}}) - T(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}}) &= \delta'|\widehat{d_{k'}}| + \frac{1}{\delta'}\kappa\|\widehat{\boldsymbol{\alpha}_{k'}}\|_1 - |\widehat{d_{k'}}| - \kappa\|\widehat{\boldsymbol{\alpha}_{k'}}\|_1 \\ &= \Big(\frac{\delta'}{c} + \frac{c}{\delta'} - \frac{1}{c} - c\Big)\sqrt{\kappa\|\widehat{\boldsymbol{\beta}_{k'}}\|_1} \\ &= -\frac{1}{c}(\delta' - 1)(c^2/\delta' - 1)\sqrt{\kappa\|\widehat{\boldsymbol{\beta}_{k'}}\|_1} \\ &< 0. \end{aligned}$$

Therefore, for any $\delta > 0$, we can find $\widetilde{\boldsymbol{d}}, \widetilde{\boldsymbol{\alpha}}$ such that $\|\widetilde{\boldsymbol{d}} - \delta'\widehat{\boldsymbol{d}}\|_1 + \|\widetilde{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}\|_1 < \delta$ and $T(\widetilde{\boldsymbol{d}}, \widetilde{\boldsymbol{\alpha}}) < T(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$. These contradict with $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ being a local minimizer.

Similarly for the case when $c < 1$. Hence, we have the result that if $\widehat{d}_k \neq 0$, then $\widehat{d}_k = (\kappa\|\widehat{\boldsymbol{\beta}_k}\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}_k} = \widehat{\boldsymbol{\beta}_k}/(\kappa\|\widehat{\boldsymbol{\beta}_k}\|_1)^{1/2}$. □

Proof of Theorem 2.1.

Proof. Suppose $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ is a local minimizer of $T(\boldsymbol{d}, \boldsymbol{\alpha})$, we first show that $\widehat{\boldsymbol{\beta}}$, where $\widehat{\beta}_{kj} = \widehat{d}_k\widehat{\alpha}_{kj}$, is a local minimizer of $T(\boldsymbol{\beta})$ as defined in (2.2), i.e. there exists a δ' such that if $\|\Delta\boldsymbol{\beta}\|_1 < \delta'$ then $T(\widehat{\boldsymbol{\beta}} + \Delta\boldsymbol{\beta}) \geq T(\widehat{\boldsymbol{\beta}})$.

We denote $\Delta\boldsymbol{\beta} = \Delta\boldsymbol{\beta}^{(1)} + \Delta\boldsymbol{\beta}^{(2)}$, where $\Delta\boldsymbol{\beta}^{(1)} = 0$ if $\|\widehat{\boldsymbol{\beta}_k}\|_1 = 0$ and $\Delta\boldsymbol{\beta}^{(2)} = 0$ if $\|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0$. We have $\|\Delta\boldsymbol{\beta}\|_1 = \|\Delta\boldsymbol{\beta}^{(1)}\|_1 + \|\Delta\boldsymbol{\beta}^{(2)}\|_1$.

Now we show $T(\widehat{\boldsymbol{\beta}} + \Delta\boldsymbol{\beta}^{(1)}) \geq T(\widehat{\boldsymbol{\beta}})$ if δ' is small enough. By Lemma A.1, we have $\widehat{d}_k = (\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\beta}}_k/(\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$ if $\|\widehat{d}_k\|_1 \neq 0$ and $\widehat{\boldsymbol{\alpha}}_k = 0$ if $\|\widehat{d}_k\|_1 = 0$. Furthermore, let $\widehat{d}'_k = (\kappa\|\widehat{\boldsymbol{\beta}}_k + \Delta\boldsymbol{\beta}_k^{(1)}\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}}'_k = (\widehat{\boldsymbol{\beta}}_k + \Delta\boldsymbol{\beta}_k^{(1)})/(\kappa\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1 + \Delta\boldsymbol{\beta}_k^{(1)})^{1/2}$ if $\|\widehat{d}_k\|_1 \neq 0$. Let $\widehat{d}'_k = 0, \widehat{\boldsymbol{\alpha}}'_k = 0$ if $\|\widehat{d}_k\|_1 = 0$. Then we have $T(\widehat{\boldsymbol{d}'}, \widehat{\boldsymbol{\alpha}'}) = T(\widehat{\boldsymbol{\beta}} + \Delta\boldsymbol{\beta}^{(1)})$ and $T(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}}) = T(\widehat{\boldsymbol{\beta}})$. Hence we only need to show that $T(\widehat{\boldsymbol{d}'}, \widehat{\boldsymbol{\alpha}'}) \geq T(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$. Note that $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ is a local minimizer of $T(\boldsymbol{d}, \boldsymbol{\alpha})$. Therefore there exists a δ such that for any $\boldsymbol{d}', \boldsymbol{\alpha}'$ satisfying $\|\boldsymbol{d}' - \widehat{\boldsymbol{d}}\|_1 + \|\boldsymbol{\alpha}' - \widehat{\boldsymbol{\alpha}}\|_1 < \delta$, we have $T(\boldsymbol{d}', \boldsymbol{\alpha}') \geq T(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$.

Now since

$$\begin{aligned}
|\widehat{d}'_k - \widehat{d}_k| &= \left|\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k + \Delta\boldsymbol{\beta}_k^{(1)}\|_1} - \sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}\right| \\
&\leq \left|\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1 - \kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1} - \sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}\right| \\
&\leq \frac{1}{2}\frac{\kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1 - \kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}} \\
&\leq \frac{1}{2}\frac{\kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa a - \kappa\delta'}} \\
&\leq \frac{1}{2}\frac{\kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa a/2}},
\end{aligned}$$

where $a = \min\{\|\widehat{\boldsymbol{\beta}}_k\|_1 : \|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0\}$ and $\delta' < a/2$.

Furthermore

$$\|\widehat{\boldsymbol{\alpha}}'_k - \widehat{\boldsymbol{\alpha}}_k\|_1 = \left\|\frac{\widehat{\boldsymbol{\beta}}_k + \Delta\boldsymbol{\beta}_k^{(1)}}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k + \Delta\boldsymbol{\beta}_k^{(1)}\|_1}} - \frac{\widehat{\boldsymbol{\beta}}_k}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}}\right\|_1$$

$$
\begin{aligned}
&\leq \Big|\Big|\frac{\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}\|_1}}-\frac{\widehat{\boldsymbol{\beta}}_k}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}\|_1}}\Big|\Big|_1\\
&\quad+\Big|\Big|\frac{\widehat{\boldsymbol{\beta}}_k}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}\|_1}}-\frac{\widehat{\boldsymbol{\beta}}_k}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}}\Big|\Big|_1\\
&\leq \frac{\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa a/2}}\\
&\quad+\frac{\|\widehat{\boldsymbol{\beta}}_k\|_1\Big|\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}\|_1}-\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}\Big|}{\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k+\Delta\boldsymbol{\beta}_k^{(1)}\|_1}\sqrt{\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1}}\\
&\leq \frac{\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa a/2}}+\frac{b}{\sqrt{\kappa a/2}\sqrt{\kappa a}}\Big(\frac{1}{2}\frac{\kappa\|\Delta\boldsymbol{\beta}_k^{(1)}\|_1}{\sqrt{\kappa a/2}}\Big)\\
&\leq \|\Delta\boldsymbol{\beta}_k^{(1)}\|_1\Big(\frac{1}{\sqrt{\kappa a/2}}+\frac{b}{a\sqrt{\kappa a}}\Big),
\end{aligned}
$$

where $b=\max\{\|\widehat{\boldsymbol{\beta}}_k\|_1:\|\widehat{\boldsymbol{\beta}}_k\|_1\neq 0\}$.

Therefore,there exists a small enough δ', if $\|\Delta\boldsymbol{\beta}^{(1)}\|_1<\delta'$ we have $\|\widehat{\boldsymbol{d}'}-\widehat{\boldsymbol{d}}\|_1+\|\widehat{\boldsymbol{\alpha}}'-\widehat{\boldsymbol{\alpha}}\|_1<\delta$. Hence $T(\widehat{\boldsymbol{d}'},\widehat{\boldsymbol{\alpha}}')\geq T(\widehat{\boldsymbol{d}},\widehat{\boldsymbol{\alpha}})$ (due to local minimality) and $T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)})\geq T(\widehat{\boldsymbol{\beta}})$.

Next we show $T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)}+\Delta\boldsymbol{\beta}^{(2)})\geq T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)})$. Note that

$$
T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)}+\Delta\boldsymbol{\beta}^{(2)})-T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)})=\Delta\boldsymbol{\beta}^{(2)T}\nabla\sum_{i=1}^{n}Q(Y_i,\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}^*})+\sum_{k=1}^{K}\sqrt{\kappa\|\Delta\boldsymbol{\beta}^{(2)}\|_1}
$$

where $\widehat{\boldsymbol{\beta}^*}$ is between $\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)}+\Delta\boldsymbol{\beta}^{(2)}$ and $\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)}$. Since $\|\Delta\boldsymbol{\beta}^{(2)}\|_1<\delta'$ is small enough, the second term dominates the first term, hence we have $T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)}+\Delta\boldsymbol{\beta}^{(2)})\geq T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta}^{(1)})$.

Overall, we have that there exists a small enough δ', if $\|\Delta\boldsymbol{\beta}\|_1<\delta'$, then $T(\widehat{\boldsymbol{\beta}}+\Delta\boldsymbol{\beta})\geq T(\widehat{\boldsymbol{\beta}})$, which implies that $\widehat{\boldsymbol{\beta}}$ is a local minimizer of $T(\boldsymbol{\beta})$.

Similarly, we can prove that if $\widehat{\boldsymbol{\beta}}$ is a local minimizer of $T(\boldsymbol{\beta})$, and if we let $\widehat{d_k} = (\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}, \widehat{\boldsymbol{\alpha}_k} = \widehat{\boldsymbol{\beta}}_k(\kappa\|\widehat{\boldsymbol{\beta}}_k\|_1)^{1/2}$ for $\|\widehat{\boldsymbol{\beta}}_k\|_1 \neq 0$ and let $\widehat{\boldsymbol{d}}_k = 0,\ \widehat{\boldsymbol{\alpha}}_k = 0$ for $\|\widehat{\boldsymbol{\beta}}_k\|_1 = 0$, then $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{\alpha}})$ is a local minimizer of $T(\boldsymbol{d}, \boldsymbol{\alpha})$. □

To prove Theorem 2.2, we first prove Lemma A.2.

Lemma A.2. *Under the condition of Theorem* 2.2*, with probability at least* $1 - p^{1-A^2}$*, the following inequalities hold:*

$$\begin{aligned} & \frac{c_1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 + \gamma\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\ \leq\ & \sum_{i=1}^n q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0))^2 + 2\gamma(1 + c_{\max}/c_{\min}) \sum_{k\in G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1. \end{aligned}$$

Proof. Since $\widehat{\boldsymbol{\beta}}$ is a minimizer of (2.2) we have that for any $\boldsymbol{\beta}$

$$\begin{aligned} & \sum_{i=1}^n [Q(Y_i, F^{-1}(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})) - Q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}^0)) + P_\kappa(\widehat{\boldsymbol{\beta}}) \\ \leq\ & \sum_{i=1}^n [Q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta})) - Q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}^0))] + P_\kappa(\boldsymbol{\beta}). \end{aligned}$$

By Taylor expansion we have

$$\begin{aligned} \frac{c_2}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^0\|_2^2 \ \leq\ & \sum_{i=1}^n q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^*)(\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0))^2 \\ \leq\ & \sum_{i=1}^n q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0))^2 \\ & \kappa + \frac{1}{n}\sum_{i=1}^n q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ & + P_\kappa(\boldsymbol{\beta}) - P_\kappa(\widehat{\boldsymbol{\beta}}) \end{aligned}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}^0$ and $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{**}$ is between $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}$.

Since

$$\frac{1}{n}\sum_{i=1}^{n} q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \le \frac{1}{n}\Big|\Big|\sum_{i=1}^{n} q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i\Big|\Big|_\infty \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$$

and

$$\begin{aligned} P\Big(\frac{1}{n}\Big|\Big|\sum_{i=1}^{n} q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i\Big|\Big|_\infty > \gamma\Big) &\le p\cdot P(|W| > \frac{\sqrt{n}\gamma}{c_1}) \\ &\le p^{1-A^2}. \end{aligned}$$

It follows that on the set $\{\frac{1}{n}\|\sum_{i=1}^{n} q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i\|_\infty \le \gamma\}$, with probability at least $1-p^{1-A^2}$, we have

$$\begin{aligned} \frac{c_2}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}^0\|_2^2 &\le \sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 \\ &\quad +\frac{1}{n}\Big|\Big|\sum_{i=1}^{n} q_1(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^0)\boldsymbol{X}_i\Big|\Big|_\infty \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\ &\quad +P_\kappa(\boldsymbol{\beta})-P_\kappa(\widehat{\boldsymbol{\beta}}) \\ &\le \sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 + \gamma\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\ &\quad +P_\kappa(\boldsymbol{\beta})-P_\kappa(\widehat{\boldsymbol{\beta}}). \end{aligned}$$

Adding $\gamma\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$ to both side then we have

$$\begin{aligned} &\frac{c_1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2 + \gamma\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\ \le\ &\sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 + 2\gamma\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 + P_\kappa(\boldsymbol{\beta})-P_\kappa(\widehat{\boldsymbol{\beta}}) \end{aligned}$$

$$
\begin{aligned}
\leq \; & \sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 + 2\gamma \sum_{k\in G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \\
& + \kappa \sum_{k\in G(\boldsymbol{\beta})} \left(\sqrt{\|\boldsymbol{\beta}_k\|_1} - \sqrt{\|\widehat{\boldsymbol{\beta}}_k\|_1}\right) \\
& + 2\gamma \sum_{k\notin G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k\|_1 - \kappa \sum_{k\notin G(\boldsymbol{\beta})} \sqrt{\|\widehat{\boldsymbol{\beta}}_k\|_1}.
\end{aligned}
$$

Since

$$
\begin{aligned}
\sqrt{\|\boldsymbol{\beta}_k\|_1} - \sqrt{\|\widehat{\boldsymbol{\beta}}_k\|_1} = \frac{\|\boldsymbol{\beta}_k\|_1 - \|\widehat{\boldsymbol{\beta}}_k\|_1}{\sqrt{\|\boldsymbol{\beta}_k\|_1} + \sqrt{\|\widehat{\boldsymbol{\beta}}_k\|_1}} \; & \leq \; \frac{\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1}{\sqrt{\|\boldsymbol{\beta}_k\|_1}} \\
& \leq \; \frac{\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1}{c_{\min}}
\end{aligned}
$$

and

$$
2\gamma \sum_{k\notin G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k\|_1 - \kappa \sum_{k\notin G(\boldsymbol{\beta})} \sqrt{\|\widehat{\boldsymbol{\beta}}_k\|_1} \leq 0,
$$

we have

$$
\begin{aligned}
\frac{c_1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2 + \gamma\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \; & \leq \; \sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 \\
& \quad + 2\gamma \sum_{k\in G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \\
& \quad + \frac{\kappa}{c_{\min}} \sum_{k\in G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \\
& = \; \sum_{i=1}^{n} q_2(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}^{**})(\boldsymbol{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0))^2 \\
& \quad + 2\gamma(1 + c_{\max}/c_{\min}) \sum_{k\in G(\boldsymbol{\beta})} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1.
\end{aligned}
$$

□

Proof of Theorem 2.2.

Proof. Let $s = |G(\boldsymbol{\beta}^0)|$, and $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ in Lemma A.2,then

$$\begin{aligned}\frac{c_1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 &\leq 2\gamma(1 + c_{\max}/c_{\min}) \sum_{k \in G(\boldsymbol{\beta}^0)} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_1 \\ &\leq 2\gamma\sqrt{s}\sqrt{\sum_{k \in G(\boldsymbol{\beta}^0)} p_k(1 + c_{\max}/c_{\min})^2 \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_2^2}.\end{aligned}$$

Under the Group RE assumption

$$\sqrt{\sum_{k \in G(\boldsymbol{\beta}^0)} p_k(1 + c_{\max}/c_{\min})^2 \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_2^2} \leq \frac{2\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2}{\zeta\sqrt{n}}.$$

$$\frac{c_1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \leq 2\gamma\sqrt{s}\frac{2\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2}{\zeta\sqrt{n}}.$$

Then we have

$$\frac{1}{\sqrt{n}}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2 \leq \frac{4\sqrt{s}\gamma}{c_1\zeta}$$

$$\frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \leq \frac{16s\gamma^2}{c_1^2\zeta^2}.$$

Also we have

$$\sum_{k \notin G(\boldsymbol{\beta}^0)} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_1 \leq \sum_{k \in G(\boldsymbol{\beta}^0)} (1 + 2c_{\max}/c_{\min})\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_1,$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 2(1 + c_{\max}/c_{\min}) \sum_{k \in G(\boldsymbol{\beta}^0)} \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\|_1$$

$$\begin{aligned} &\leq 2\sqrt{s}\sqrt{\sum_{k\in G(\boldsymbol{\beta}^0)} p_k(1+c_{\max}/c_{\min})^2\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0\|_2^2} \\ &\leq 2\sqrt{s}\frac{2\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2}{\zeta\sqrt{n}} \\ &\leq \frac{16s\gamma}{\zeta^2 c_1}. \end{aligned}$$

Let $G = G(\boldsymbol{\beta}^0)$,

$$\begin{aligned} \|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_{G^c}\|_2 &\leq \|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_{G^c}\|_1 \leq \sum_{k\in G(\boldsymbol{\beta}^0)}(1+2c_{\max}/c_{\min})\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0\|_1 \\ &\leq 2\sqrt{S}\sqrt{\sum_{k\in G} p_k(1+c_{\max}/c_{\min})^2\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0\|_2^2} \\ &\leq \frac{4\sqrt{S}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2}{\zeta\sqrt{n}} \leq \frac{16s\gamma}{\zeta^2 c_1} \end{aligned}$$

By the Group RE assumption and the fact that $p_k \geq 1$, we have

$$\begin{aligned} \|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_G\|_2 &\leq \sqrt{\sum_{k\in G} p_k(1+c_{\max}/c_{\min})^2\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0\|_2^2} \\ &\leq \frac{2\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2}{\zeta\sqrt{n}} \leq \frac{8\sqrt{s}\gamma}{\zeta^2 c_1}. \end{aligned}$$

Then we have

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2 &\leq \|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_{G^c}\|_2 + \|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)_G\|_2 \leq \frac{16s\gamma + 8\sqrt{s}\gamma c_1}{\zeta^2 c_1^2} \\ &= (2\sqrt{s}+1)\frac{8\sqrt{s}\gamma}{\zeta^2 c_1}. \end{aligned}$$

□

Proof of Theorem 2.3.

Proof. We will show that for any given $\epsilon > 0$, there exists a constant C such that

$$P(\inf_{\|\boldsymbol{u}_n\|=C} T_n(\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n) > T_n(\boldsymbol{\beta}_n^0)) \geq 1 - \epsilon,$$

where $\alpha_n = (p_n/n)^{1/2}$. This implies that with probability at least $1 - \epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n : \|\boldsymbol{u}_n\| \leq C\}$. Hence, there exists a local minimizer such that $\|\widehat{\boldsymbol{\beta}_n} - \boldsymbol{\beta}_n^0\| = O_p(\alpha_n)$.

Let $\alpha_n = (p_n/n)^{1/2}$,

$$\begin{aligned} & T_n(\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n) - T_n(\boldsymbol{\beta}_n^0) \\ = & \frac{1}{n}\sum_{i=1}^{n}[Q_n(Y_{ni}, F^{-1}(\boldsymbol{X}_{ni}^T(\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n))) - Q_n(Y_{ni}, F^{-1}(\boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^0))] \\ & + \sum_{k=1}^{K}(P_{\kappa_n,w_n}(\boldsymbol{\beta}_{nk}^0 + \alpha_n \boldsymbol{u}_k) - P_{\kappa_n,w_n}(\boldsymbol{\beta}_{nk}^0)) \\ \triangleq & \text{ (I) + (II).} \end{aligned} \tag{A.1}$$

Let $q_j = (\partial^j/\partial\theta^j)Q_n(y,\theta)$, then by Taylor expansion,

$$\begin{aligned} \text{(I)} = & \frac{\alpha_n}{n}\sum_{i=1}^{n} q_1(Y_{ni}; \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^0)\boldsymbol{X}_{ni}^T\boldsymbol{u}_n + \frac{\alpha_n^2}{2n}\sum_{i=1}^{n} q_2(Y_{ni}; \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^0)(\boldsymbol{X}_{ni}^T\boldsymbol{u}_n)^2 \\ & + \frac{\alpha_n^3}{6n}\sum_{i=1}^{n} q_3(Y_{ni}; \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^*)(\boldsymbol{X}_{ni}^T\boldsymbol{u}_n)^3 \\ \triangleq & \text{ I}_1 + \text{I}_2 + \text{I}_3 \end{aligned} \tag{A.2}$$

where $\boldsymbol{\beta}_n^*$ is between $\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}$ and $\boldsymbol{\beta}_n^0$.

According to the proof in Zhang et al. (2010), when constant C and n are large enough, I_2 dominates I_1 and I_3 uniformly on $\|\boldsymbol{u}_n\| = C$.

Now consider the penalty part. Let s_k be the number of non-zero parameters in the kth group of the true parameters and the first s_k parameters in group k are non-zero. Since $\alpha_n = (p_n/n)^{1/2} \to 0$ as $n \to \infty$, for $\|\boldsymbol{u}_n\|_2 \leq C$, we have $|\beta^0_{n,kj} + \alpha_n \boldsymbol{u}_{n,kj}| \geq |\beta^0_{n,kj}| - |\alpha_n \boldsymbol{u}_{n,kj}| > 0$ for n large enough and $\beta^0_{n,kj} \neq 0$. Then for the kth group,

$$
\begin{aligned}
& P_{\kappa_n, w_n}(\boldsymbol{\beta}^0_{nk} + \alpha_n \boldsymbol{u}_{nk}) - P_{\kappa_n}(\boldsymbol{\beta}^0_{nk}) \\
= \;& \kappa_n \Big(\sqrt{\sum_{j=1}^{p_k} w_{kj} |\beta^0_{kj} + \alpha_n u_{j,kj}|} - \sqrt{\sum_{j=1}^{p_k} w_{kj} |\beta^0_{kj}|} \Big) \\
\geq \;& \kappa_n \Big(\sqrt{\sum_{j=1}^{s_k} w_{kj} |\beta^0_{kj} + \alpha_n u_{j,kj}|} - \sqrt{\sum_{j=1}^{s_k} w_{kj} |\beta^0_{kj}|} \Big) \\
\geq \;& \kappa_n \Big(\sqrt{\sum_{j=1}^{s_k} w_{kj} |\beta^0_{kj}| - \alpha_n \sum_{j=1}^{s_k} w_{kj} |\boldsymbol{u}_{n,kj}|} - \sqrt{\sum_{j=1}^{s_k} w_{kj} |\beta^0_{kj}|} \Big) \\
= \;& \kappa_n \sqrt{\sum_{j=1}^{s_k} w_{kj} |\beta^0_{kj}|} (\sqrt{1 - \gamma_{nk}} - 1)
\end{aligned}
$$

where $\gamma_{nk} = \alpha_n (w_{n,k1} |\boldsymbol{u}_{n,k1}| + \cdots + w_{n,ks_k} |\boldsymbol{u}_{n,kp_k}|) / (w_{n,k1} |\beta^0_{n,k1}| + \cdots + w_{n,ks_k} |\beta^0_{n,ks_k}|)$. For n large enough, we have $0 \leq \gamma_{nk} < 1$ and

$$
\begin{aligned}
\gamma_{nk} &\leq \alpha_n \|\boldsymbol{u}_{nk}\| (w_{n,k1} + \cdots + w_{n,ks_k}) / c_1 (w_{n,k1} + \cdots + w_{n,ks_k}) \\
&= \alpha_n \|\boldsymbol{u}_{nk}\| / c_1 \leq \alpha_n C / c_1 = o_P(1).
\end{aligned}
$$

Therefore by Taylor expansion,

$$\begin{aligned} & P_{\kappa_n,w_n}(\boldsymbol{\beta}^0_{nk}+\alpha_n\boldsymbol{u}_{nk})-P_{\kappa_n,w_n}(\boldsymbol{\beta}^0_{nk}) \\ \geq\ & \kappa\sqrt{\sum_{j=1}^{s_k} w_{kj}|\beta^0_{n,kj}|}\Big(\frac{1+o_p(1)}{2}(-\gamma_{nk})\Big) \\ \geq\ & -\kappa_n\frac{\alpha_n(\sum_{j=1}^{s_k} w_{kj}|u_{n,kj}|)}{\sqrt{\sum_{j=1}^{s_k} w_{n,kj}|\beta^0_{n,kj}|}}\Big(\frac{1+o_p(1)}{2}\Big) \\ \geq\ & -\alpha_n\kappa_n\frac{\|\boldsymbol{u}_{nk}\|\sqrt{w^{(\mathrm{I})}_{\max}s_k}}{2\sqrt{c_1}}(1+o_p(1)). \end{aligned}$$

Therefore

$$\begin{aligned} (\mathrm{II})\ \geq\ & -\alpha_n\kappa_n\Big(\sum_{k=1}^{K}\frac{\|\boldsymbol{u}_{nk}\|\sqrt{w^{(\mathrm{I})}_{\max}s_k}}{2\sqrt{c_1}}\Big)(1+o_p(1)) \\ \geq\ & -\alpha_n\kappa_n\sqrt{w^{(\mathrm{I})}_{\max}}\Big(\|\boldsymbol{u}_n\|\frac{\sqrt{s_n}}{2\sqrt{c_1}}\Big)(1+o_p(1)) \\ \geq\ & -\alpha_n^2\|\boldsymbol{u}_n\|O_P(1) \end{aligned}$$

which is also dominated by I_2 .

Thus $T_n(\boldsymbol{\beta}^0_n+\alpha_n\boldsymbol{u}_n)-T_n(\boldsymbol{\beta}^0_n)>0$ in probability, which complete the proof. □

Proof of Theorem 2.4.

Proof. As in the proof of Theorem 2.3, it's suffices to show that for any given $\varepsilon>0$ there exists a constant C large enough such that

$$P(\inf_{\|\boldsymbol{u}_n\|=C} T_n(\boldsymbol{\beta}^0_n+\alpha_n\boldsymbol{u}_n)>T_n(\boldsymbol{\beta}^0_n))\geq 1-\epsilon. \tag{A.3}$$

Similarly we have $T_n(\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n) - T_n(\boldsymbol{\beta}_n^0) = (\mathrm{I}) + (\mathrm{II})$ and $(\mathrm{I}) = \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3$, where (I), (II), I_1, I_2 and I_3 are defined as in (A.1) and (A.2).

Similar to the proof in Zhang et al. (2010) we have

$$|\mathrm{I}_1| \leq O_P(\alpha_n \sqrt{s_n/n}) \|\boldsymbol{u}_n^{(\mathrm{I})}\| + O_P(\alpha_n n^{-1/2}) \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1,$$

$$\begin{aligned} \mathrm{I}_2 \;\geq\; & -\frac{\alpha_n^2}{2} \boldsymbol{u}_n^{(\mathrm{I})T} E\Big[\frac{q''(m(\boldsymbol{X}_{n1}))}{F'(m(\boldsymbol{X}_{n1}))^2} \boldsymbol{X}_{n1}^{(\mathrm{I})} \boldsymbol{X}_{n1}^{(\mathrm{I})T} \Big] \boldsymbol{u}_n^{(\mathrm{I})} (1 + o_P(1)) \\ & + O_P(\alpha_n^2 s_n/\sqrt{n}) \|\boldsymbol{u}_n^{(\mathrm{I})}\|^2 + O_P(\alpha_n^2/\sqrt{n}) \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^2 \\ & + O_P(\alpha_n^2 \sqrt{s_n}) \|\boldsymbol{u}_n^{(\mathrm{I})}\| \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1, \end{aligned}$$

and

$$\begin{aligned} |\mathrm{I}_3| \;=\; & O_P(\alpha_n^3 s_n^{3/2}) \|\boldsymbol{u}_n^{(\mathrm{I})}\|^3 + O_P(\alpha_n^3 s_n) \|\boldsymbol{u}_n^{(\mathrm{I})}\|^2 \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1 \\ & + O_P(\alpha_n^3 s_n^{1/2}) \|\boldsymbol{u}_n^{(\mathrm{I})}\| \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^2 + O_P(\alpha_n^3) \|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^3. \end{aligned}$$

For the penalty part, in the kth group

$$\begin{aligned} & P_{\kappa_n, w_n}(\boldsymbol{\beta}_{nk}^0 + \alpha_n \boldsymbol{u}_{nk}) - P_{\kappa_n}(\boldsymbol{\beta}_{nk}^0) \\ = \;& \kappa_n \sqrt{\sum_{j=1}^{s_k} w_{n,kj} |\beta_{n,kj}^0 + \alpha_n u_{n,kj}| + \alpha_n \sum_{j=s_k+1}^{p_k} w_{n,kj} |u_{n,kj}|} \\ & - \kappa_n \sqrt{\sum_{j=1}^{p_k} w_{n,kj} |\beta_{n,kj}^0|} \\ \geq \;& \kappa_n \sqrt{\sum_{j=1}^{s_k} w_{n,kj} |\beta_{n,kj}^0| + \alpha_n \sum_{j=s_k+1}^{p_k} w_{n,kj} |u_{n,kj}| - \alpha_n \sum_{j=1}^{s_k} w_{n,kj} |u_{n,kj}|} \end{aligned}$$

$$
\begin{aligned}
& -\kappa_n \sqrt{\sum_{j=1}^{p_k} w_{n,kj} |\beta^0_{n,kj}|} \\
\geq \; & \kappa_n \sqrt{\alpha_n w^{(\mathrm{II})}_{\min} \sum_{j=s_k+1}^{p_k} |u_{n,kj}|} - \kappa_n \sqrt{\alpha_n w^{(\mathrm{I})}_{\max} \sum_{j=1}^{s_k} |u_{n,kj}|} - \kappa_n \sqrt{\sum_{j=1}^{s_k} w_{n,kj} |\beta^0_{n,kj}|}.
\end{aligned}
$$

Since

$$
\begin{aligned}
\kappa_n \sqrt{\alpha_n w^{(\mathrm{I})}_{\max} \sum_{j=1}^{s_k} |u_{n,kj}|} &\leq \kappa_n \sqrt{\alpha_n w^{(\mathrm{I})}_{\max} \|\boldsymbol{u}^{(\mathrm{I})}_{nk}\| \sqrt{s_k}} \\
&\leq O_P\Big(\frac{s_n^{1/4}}{n^{5/4}} s_k^{1/4}\Big) \sqrt{\|\boldsymbol{u}^{(\mathrm{I})}_{nk}\|},
\end{aligned}
$$

$$
\begin{aligned}
\sum_{k=1}^{K} \kappa_n \sqrt{\alpha_n w^{(\mathrm{I})}_{\max} \sum_{j=1}^{s_k} |u_{n,kj}|} &\leq O_P\Big(\frac{s_n^{1/2}}{n^{5/4}}\Big) \sum_{k=1}^{K} \sqrt{\|\boldsymbol{u}^{(\mathrm{I})}_{nk}\|} \\
&\leq O_P\Big(\frac{s_n^{1/2}}{n^{5/4}}\Big) \sqrt{K} \sqrt{\sum_{k=1}^{K} \|\boldsymbol{u}^{(\mathrm{I})}_{nk}\|} \\
&\leq O_P\Big(\frac{s_n^{1/2}}{n^{5/4}}\Big) K^{3/4} \sqrt{\|\boldsymbol{u}^{(\mathrm{I})}_{n}\|} \\
&= O_P\Big(\frac{s_n}{n}\Big) \sqrt{\|\boldsymbol{u}^{(\mathrm{I})}_{n}\|}
\end{aligned}
$$

Also

$$
\sum_{k=1}^{K} \kappa_n \sqrt{\alpha_n w^{(\mathrm{II})}_{\min} \sum_{j=s_k+1}^{p_k} |u_{n,kj}|} \geq \kappa \sqrt{\alpha_n w^{(\mathrm{II})}_{\min} \|\boldsymbol{u}^{(\mathrm{II})}_{n}\|_1},
$$

and

$$
\sum_{k=1}^{K} \kappa_n \sqrt{\sum_{j=1}^{s_k} w_{n,kj} |\beta_{n,kj^0}|} \leq \kappa_n \sqrt{w^{(\mathrm{I})}_{\max}} \sum_{k=1}^{K} \sqrt{\|\boldsymbol{\beta}^{0(\mathrm{I})}_{nk}\|_1}
$$

$$= \quad O_P\Big(\frac{\sqrt{K}}{n}\Big) = o_P(\alpha_n).$$

Therefore

$$\text{(II)} \geq \kappa_n \sqrt{\alpha_n w_{\min}^{(\text{II})} \|\boldsymbol{u}_n^{(\text{II})}\|_1} - O_P\Big(\frac{s_n}{n}\Big)\sqrt{\|\boldsymbol{u}_n^{(\text{I})}\|} - o_P(\alpha_n).$$

Since $\|\boldsymbol{u}_n^{(\text{II})}\|_1 \leq \sqrt{p_n - s_n}\|\boldsymbol{u}_n^{(\text{II})}\| = O_P(\alpha_n^{-1})$ by $s_n(p_n - s_n) = O(n)$ and $\alpha_n = (s_n/n)^{1/2}$, we have

$$\begin{aligned}
O_P(\alpha_n^2/\sqrt{n})\|\boldsymbol{u}_n^{(\text{II})}\|_1^2 &= O_P(\sqrt{\alpha_n}/\sqrt{n})\|\boldsymbol{u}_n^{(\text{II})}\|^{3/2}\sqrt{\|\boldsymbol{u}_n^{(\text{II})}\|_1}, \\
O_P(\alpha_n^3 s_n^{1/2})\|\boldsymbol{u}_n^{(\text{I})}\|\|\boldsymbol{u}_n^{(\text{II})}\|_1^2 &= O_P(\alpha_n^{3/2} s_n^{1/2})\|\boldsymbol{u}_n^{(\text{I})}\|\|\boldsymbol{u}_n^{(\text{II})}\|^{3/2}\sqrt{\|\boldsymbol{u}_n^{(\text{II})}\|_1}, \\
O_P(\alpha_n^3)\|\boldsymbol{u}_n^{(\text{II})}\|_1^3 &= O_P(\sqrt{\alpha_n})\|\boldsymbol{u}_n^{(\text{II})}\|^{5/2}\sqrt{\|\boldsymbol{u}_n^{(\text{II})}\|_1}, \\
O_P(\alpha_n^2\sqrt{s_n})\|\boldsymbol{u}_n^{(\text{I})}\|\|\boldsymbol{u}_n^{(\text{II})}\|_1 &= O_P(\alpha_n^{3/2}\sqrt{s_n})\|\boldsymbol{u}_n^{(\text{I})}\|\sqrt{\|\boldsymbol{u}_n^{(\text{II})}\|}\sqrt{\|\boldsymbol{u}_n^{(\text{II})}\|_1}, \\
O_P(\alpha_n^3 s_n)\|\boldsymbol{u}_n^{(\text{I})}\|^2\|\boldsymbol{u}_n^{(\text{II})}\|_1 &= O_P(\alpha_n^{5/2} s_n)\|\boldsymbol{u}_n^{(\text{I})}\|^2\sqrt{\boldsymbol{u}_n^{(\text{II})}\|}\sqrt{\boldsymbol{u}_n^{(\text{II})}\|_1}.
\end{aligned}$$

With $\|\boldsymbol{u}_n\| = C$ and some fixed positive value $C_0 < C$, now consider two cases: (1) $\|\boldsymbol{u}_n^{(\text{II})}\|_1 \leq C_0$ and (2) $\|\boldsymbol{u}_n^{(\text{II})}\|_1 > C_0$.

In Case (1), we have $\|\boldsymbol{u}_n^{(\text{I})}\| \geq (C^2 - C_0^2)^{1/2}$. Then for any $\varepsilon > 0$, there exists some C large enough with $\|\boldsymbol{u}_n\| = C$ such that with probability $1 - \varepsilon/2$ and n large enough,

$$-\alpha_n^2 \boldsymbol{u}_n^{(\text{I})T} E\Big\{\frac{q''(m(\boldsymbol{X}_{n1}))}{F'(m(\boldsymbol{X}_{n1}))^2}\boldsymbol{X}_{n1}^{(\text{I})}\boldsymbol{X}_{n1}^{(\text{I})T}\Big\}\boldsymbol{u}_n^{(\text{I})}$$

dominates the following terms:

$$O_P(\alpha_n\sqrt{s_n/n})\|\boldsymbol{u}_n^{(\mathrm{I})}\|,$$

$$O_P(\alpha_n^2 s_n/\sqrt{n})\|\boldsymbol{u}_n^{(\mathrm{I})}\|^2,$$

$$O_P(\alpha_n^3 s_n^{3/2})\|\boldsymbol{u}_n^{(\mathrm{I})}\|^3,$$

$$O_P(s_n/n)\sqrt{\|\boldsymbol{u}_n^{(\mathrm{I})}\|}$$

and by $\kappa_n(w_{\min}^{(\mathrm{II})})^{1/2} = O(1)$, with probability $1-\varepsilon/2$ for large n, the term

$$\kappa_n\sqrt{\alpha_n w_{\min}^{(\mathrm{II})}\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1}$$

dominates the following:

$$O_P(\alpha_n n^{-1/2})\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1,$$

$$O_P(\alpha_n^2\sqrt{s_n})\|\boldsymbol{u}_n^{(\mathrm{I})}\|\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1,$$

$$O_P(\alpha^2 n^{-1/2})\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^2,$$

$$O_P(\alpha_n^3 s_n)\|\boldsymbol{u}_n^{(\mathrm{I})}\|^2\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1,$$

$$O_P(\alpha_n^3\sqrt{s_n})\|\boldsymbol{u}_n^{(\mathrm{I})}\|\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^2,$$

$$O_P(\alpha_n^3)\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^3,$$

$$o_P(\alpha_n).$$

Thus $P(\inf_{\|\boldsymbol{u}_n\|=C} T_n(\boldsymbol{\beta}_n^0 + \alpha_n \boldsymbol{u}_n) > T_n(\boldsymbol{\beta}_n^0)) > 1 - \varepsilon$ when n is large enough.

In Case (2), $-\alpha_n^2 \boldsymbol{u}_n^{(\mathrm{I})T} E[q''(m(\boldsymbol{X}_{n1}))\boldsymbol{X}_{n1}^{(\mathrm{I})}\boldsymbol{X}_{n1}^{(\mathrm{I})T}/F'(m(\boldsymbol{X}_{n1}))^2]\boldsymbol{u}_n^{(\mathrm{I})}$ dominates

$$O_P(\alpha_n^2 s_n/n^{1/2})\|\boldsymbol{u}_n^{(\mathrm{I})}\|^2 \quad \text{and} \quad O_P(\alpha_n^3 s_n^{3/2})\|\boldsymbol{u}_n^{(\mathrm{I})}\|^3,$$

while $\kappa_n(\alpha_n w_{\min}^{(\mathrm{II})}\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1)^{1/2}$ dominates

$$O_P(\alpha_n(s_n/n)^{1/2})\|\boldsymbol{u}_n^{(\mathrm{I})}\|, \quad O_P(s_n/n)(\|\boldsymbol{u}_n^{(\mathrm{I})}\|)^{1/2}$$

and other terms as in Case (1).

Combining the conclusions above, we have shown that for any given $\varepsilon > 0$ there is a constant C large enough such that (A.3) holds for large n. Thus we have proved Theorem 2.4. □

Proof of Theorem 2.5.

Proof. I) It's sufficient to show that for any $\boldsymbol{\beta}_n$ satisfying

$$\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P((s_n/n)^{1/2})$$

and $j = s_k + 1, \ldots, p_k$ for any $k = 1, \ldots, K$, with probability tending to 1, we have that:

$$\frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} < 0, \qquad \text{for } \beta_{n,kj} < 0,$$
$$\frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} > 0, \qquad \text{for } \beta_{n,kj} > 0.$$

In other words, we need to show that with probability tending to 1,

$$\max \sup_{j\in\mathcal{D}_2} \Big\{ \frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P(\sqrt{\frac{s_n}{n}}), \beta_{n,kj} < 0 \Big\} < 0,$$

and

$$\min \inf_{j\in\mathcal{D}_2} \Big\{ \frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P(\sqrt{\frac{s_n}{n}}), \beta_{n,kj} > 0 \Big\} > 0. \tag{A.4}$$

For the first part,

$$\begin{aligned} \frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} &= \frac{1}{n}\sum_{i=1}^{n} q_1(Y_{ni}, \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^0) X_{n,i(kj)} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} q_2(Y_{ni}, \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^*)\{\boldsymbol{X}_{ni}^T(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0)\} X_{n,i(kj)} \\ &\quad + \frac{\kappa_n w_{n,kj}\mathrm{sign}(\beta_{n,kj})}{\sqrt{\sum_{l=1}^{p_k} w_{n,kl}|\beta_{n,kl}|}} \end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_n^0$ and $\boldsymbol{\beta}_n$.

Then

$$\begin{aligned} & \max_j \sup \Big\{ \frac{\partial T_n(\boldsymbol{\beta}_n)}{\partial \beta_{n,kj}} : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P(\frac{s_n}{n}), \beta_{n,kj} < 0 \Big\} \\ \leq\ & \max_j \frac{1}{n} q_1(Y_{ni}, \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^0) X_{n,i(kj)} \\ & + \max_j \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P(\sqrt{s_n/n})} \Big\{ \frac{1}{n}\sum_{i=1}^{n} q_2(Y_{ni}, \boldsymbol{X}_{ni}^T\boldsymbol{\beta}_n^*)\{\boldsymbol{X}_{ni}^T(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0)\} X_{n,i(kj)} \Big\} \\ & - \frac{\kappa_n w_{n,kj}\mathrm{sign}(\beta_{n,kj})}{\sqrt{\sum_{l=1}^{p_k} w_{n,kl}|\beta_{n,kl}|}} \\ \triangleq\ & \mathrm{I}_1 + \mathrm{I}_2 - \mathrm{I}_3. \end{aligned}$$

We have $|\mathrm{I}_1| = O_P(n^{1/2})$, $|\mathrm{I}_2| = O_P((s_n p_n/n)^{1/2})$.

Since $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P((s_n/n)^{1/2})$ and $\|\boldsymbol{\beta}_n^0\|_1$ is bounded, $\|\boldsymbol{\beta}_{nk}\|$ is also bounded. Let $w_{\max}^{(\mathrm{II})} = \max\{w_{n,kj}, k = 1, \ldots, K \text{ and } j = s_k, \ldots, p_k\}$, then we have

$$\mathrm{I}_3 \geq \frac{\kappa_n w_{\min}^{(\mathrm{II})}}{\sqrt{w_{\max}^{(\mathrm{II})}}\sqrt{\|\boldsymbol{\beta}_n^0\|_1}} = \kappa_n \sqrt{w_{\min}^{(\mathrm{II})}} O_P(1).$$

Since $(\kappa_n (w_{\min}^{(\mathrm{II})})^{1/2})^{-1} = o_P(1)$, I_3 dominates I_1 and I_2.

Thus

$$\max_j \sup\{\partial T_n(\boldsymbol{\beta}_n)/\partial \beta_{n,kj} : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| = O_P((s_n/n)^{1/2}), \beta_{n,kj} < 0\} < 0$$

with probability tending to 1.

The second part in (A.4) can be proved similarly.

II). From I), we know that with probability tending to 1,

$$\frac{\partial T_n(\boldsymbol{\beta}_n^{(\mathrm{I})}, \mathbf{0})}{\partial \boldsymbol{\beta}_n^{(\mathrm{I})}}\Big|_{\boldsymbol{\beta}_n^{(\mathrm{I})} = \widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}} = 0.$$

Since for any group k and $j = 1, \ldots, s_k$,

$$\frac{\partial P_{\kappa_n, w_n}(\boldsymbol{\beta}_n^{(\mathrm{I})}, \mathbf{0})}{\partial \beta_{n,kj}}\Big|_{\boldsymbol{\beta}_n^{(\mathrm{I})} = \widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}} = \frac{\kappa_n w_{n,kj} \mathrm{sign}(\widehat{\beta}_{n,kj})}{\sqrt{\sum_{l=1}^{s_k} w_{n,kl} |\widehat{\beta}_{n,kl}|}} = O_P(s_n^{3/2}/\sqrt{n}),$$

it follows that for n large enough,

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n} q_1(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} q_2(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\{\boldsymbol{X}_{ni}^{(\mathrm{I})T}(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})})\}\boldsymbol{X}_{ni}^{(\mathrm{I})} \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n} q_3(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{*(\mathrm{I})})\{\boldsymbol{X}_{ni}^{(\mathrm{I})T}(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})})\}^2\boldsymbol{X}_{ni}^{(\mathrm{I})} \\
&\quad + O_P(s_n^{3/2}/\sqrt{n}) \\
&\triangleq K_1 + K_2 + K_3,
\end{aligned}$$

where $\boldsymbol{\beta}_n^{*(\mathrm{I})}$ is between $\boldsymbol{\beta}_n^{0(\mathrm{I})}$ and $\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}$. We can see that

$$\begin{aligned}
& K_2 - H_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \\
= \;& \Big\{\frac{1}{n}\sum_{i=1}^{n} q_2(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}\boldsymbol{X}_{ni}^{(\mathrm{I})T} - H_n\Big\}(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \\
= \;& -\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{q''(m(\boldsymbol{X}_{ni}))}{F'(m(\boldsymbol{X}_{ni}))^2}\boldsymbol{X}_{ni}^{(\mathrm{I})}\boldsymbol{X}_{ni}^{(\mathrm{I})T} - \\
& E[\frac{q''(m(\boldsymbol{X}_{n1}))}{F'(m(\boldsymbol{X}_{n1}))^2}\boldsymbol{X}_{n1}^{(\mathrm{I})}\boldsymbol{X}_{n1}^{(\mathrm{I})T}]\Big\}(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \\
& + \frac{1}{n}\sum_{i=1}^{n}\{Y_{ni} - m(\boldsymbol{X}_{ni})\}A(m(\boldsymbol{X}_{ni})))\boldsymbol{X}_{ni}^{(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})T}(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \\
= \;& O_P(s_n/\sqrt{n})O_P(\sqrt{s_n/n}) = O_P(s_n^{3/2}/n),
\end{aligned}$$

$$\|K_3\| = \Big|\Big|\frac{1}{n}\sum_{i=1}^{n} q_3(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{*(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}\Big|\Big|O_P(s_n^2/n) = O_P(s_n^{5/2}/n),$$

$$H_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) = -\frac{1}{n}\sum_{i=1}^{n} q_1(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})} + O_P(s_n^{5/2}/n).$$

Thus

$$\sqrt{n}A_n\Omega_n^{-1/2}H_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}-\boldsymbol{\beta}_n^{0(\mathrm{I})}) = -\frac{1}{\sqrt{n}}A_n\Omega_n^{-1/2}\sum_{i=1}^{n} q_1(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})} + o_P(1).$$

Let $\boldsymbol{Z}_{ni} = -A_n\Omega_n^{-1/2}q_1(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}/n^{1/2}$. Now we only need to check that $\boldsymbol{Z}_{ni}$ satisfies the Linderberg-Feller conditions of central limit theorem for normality.

Since $\mathrm{var}(q_1(Y_{ni}, \boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})) = \Omega_n$, it's obvious that $\sum_{i=1}^{n}\mathrm{cov}(\boldsymbol{Z}_{ni}) \to G$.

For any $\delta > 0$, we have

$$\begin{aligned}
& E(\|\boldsymbol{Z}_{n1}\|^{2+\delta}) \\
= \;& n^{-(2+\delta)/2}E(\|A_n\Omega_n^{-1/2}\boldsymbol{X}_{n1}^{(\mathrm{I})}q_1(Y_{n1}, \boldsymbol{X}_{n1}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\|^{2+\delta}) \\
\leq \;& n^{-(2+\delta)/2}E\{\|A_n\|_F^{2+\delta}[\|\Omega_n^{-1/2}\boldsymbol{X}_{n1}^{(\mathrm{I})}\|\Big|\frac{q''(m(\boldsymbol{X}_{n1}))}{F'(m(\boldsymbol{X}_{n1}))}(Y_{n1}-m(\boldsymbol{X}_{n1}))\Big|]^{2+\delta}\} \\
\leq \;& c_1 n^{-(2+\delta)/2}E\{(\kappa_{\min}^{-1/2}(\Omega_n)\|\boldsymbol{X}_{n1}^{(\mathrm{I})}\|)^{2+\delta}|Y_{n1}-m(\boldsymbol{X}_{n1})|^{2+\delta}\} \\
\leq \;& c_2 s_n^{(2+\delta)/2}n^{-(2+\delta)/2}E|Y_{n1}-m(\boldsymbol{X}_{n1})|^{2+\delta} \\
\leq \;& 2c_2 s_n^{(2+\delta)/2}n^{-(2+\delta)/2}[E|Y_{n1}|^{2+\delta}+E|m(\boldsymbol{X}_{n1})|^{2+\delta}] \\
= \;& O((s_n/n)^{(2+\delta)/2})
\end{aligned}$$

for some constants c_1 and c_2. Therefore

$$\sum_{i=1}^{n}E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = O\Big(\frac{s_n^{(2+\delta)/1}}{n^{\delta/2}}\Big).$$

Let $\delta = 1$, then $\sum_{i=1}^{n}E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = o(1)$.

□

Proof of Theorem 2.7.

Proof. According to the construction of W_n in (2.6), it is sufficient to show that

$$\sqrt{n}(A_n\widehat{H}_n^{-1}\widehat{\Omega}_n\widehat{H}_n^{-1}A_n^T)^{-1/2}A_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}-\boldsymbol{\beta}_n^{o(\mathrm{I})})\xrightarrow{d}N(\mathbf{0},\mathrm{I}_s).$$

From the proof of Theorem 2.5 Part II), we can get

$$H_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}-\boldsymbol{\beta}_n^{0(\mathrm{I})})=-\frac{1}{n}\sum_{i=1}^{n}q_1(Y_{ni},\boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}+O_P(s_n^{5/2}/n).$$

Since eigenvalues of H_n is bounded uniformly away from 0 and $s_n^5/n\to 0$ as $n\to\infty$,

$$(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}-\boldsymbol{\beta}_n^{0(\mathrm{I})})=-\frac{1}{n}H_n^{-1}\sum_{i=1}^{n}q_1(Y_{ni},\boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}+o_P(1/\sqrt{n}).$$

Let $U_n=A_nH_n^{-1}\Omega_nH_n^{-1}A_n^T$, $\widehat{U}_n=A_n\widehat{H}_n^{-1}\widehat{\Omega}_n\widehat{H}_n^{-1}A_n^T$. Eigenvalues of $H_n^{-1}\Omega_nH_n^{-1}$ are uniformly bounded away from 0, so are those of U_n.

Then we can see that

$$\begin{aligned}\sqrt{n}U_n^{-1/2}A_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})}-\boldsymbol{\beta}_n^{0(\mathrm{I})}) &= \frac{1}{n}U_n^{-1/2}A_nH_n^{-1}\sum_{i=1}^{n}q_1(Y_{ni},\boldsymbol{X}_{ni}^{(\mathrm{I})T}\boldsymbol{\beta}_n^{0(\mathrm{I})})\boldsymbol{X}_{ni}^{(\mathrm{I})}+o_P(1)\\ &\triangleq \sum_{i=1}^{n}\boldsymbol{Z}_{ni}+o_P(1).\end{aligned}$$

Now check the Linderberg-Feller conditions for $\{\boldsymbol{Z}_{ni}\}$. It's obvious that

$$\sum_{i=1}^{n}\mathrm{cov}(\boldsymbol{Z}_{ni})=\mathrm{I}_s.$$

$E(\|\boldsymbol{Z}_{n1}\|^{2+\delta}) = O((s_n/n)^{(2+\delta)/2})$ according to similar arguments used in the proof of Theorem 2.5. It yields that $\sum_{i=1}^{n} E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = O(s_n^{(2+\delta)/2}/n^{\delta/2}) = o(1)$. Hence

$$\sqrt{n}U_n^{-1/2}A_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{0(\mathrm{I})}) \xrightarrow{d} N(\mathbf{0}, \mathrm{I}_s). \tag{A.5}$$

Since $\|A_n\|_F^2 \to tr(G)$ and $\|\widehat{V}_n - V_n\| = o_P(1)$ by Corollary 1, we have $\|\widehat{U}_n - U_n\| = \|A_n(\widehat{V}_n - V_n)A_n^T\| \leq \|\widehat{V}_n - V_n\|\|A_n\|_F^2 = o_P(1)$. By the assumption, the eigenvalues of $\widehat{U}_n$ are uniformly bounded away from 0 and ∞ with probability tending to 1. Consequently,

$$\|\widehat{U}_n^{-1/2}U_n^{1/2} - \mathrm{I}_s\| = o_P(1). \tag{A.6}$$

Combining (A.5), (A.6) and Slutsky's theorem, we get the conclusion that

$$\sqrt{n}\widehat{U}_n^{-1/2}A_n(\widehat{\boldsymbol{\beta}}_n^{(\mathrm{I})} - \boldsymbol{\beta}_n^{o(\mathrm{I})}) \xrightarrow{d} N(\mathbf{0}, \mathrm{I}_s).$$

Thus under the null hypothesis in (2.6), we have the result in (2.7).

□

Appendix B

Major Implementation of the Algorithms in Chapter 2

The penalized Group-BD is implemented in Matlab.

```
function [hat_beta, iter, converge] = glmcd_glasso(x, y, lambda, ...
weights_pen, options, init_beta,type,group,pii)
% <Input>
% x: matrix of variables
% y: responses
% lambda: tuning parameter
% weights_pen: weights for variables
% options: control values
% init_beta: start value of coefficient vector
% type: data type
```

```
% group: group number, sizes and indexes
%
% <Output>
% hat_beta: estimated coefficient vector
% iter: number of iterations taken
% converge: whether the iterations converge
%
% Functions used:
% BD: quadratic approximation of BD loss according to the data type
% soft_thres: soft threshold function
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[n, p] = size(x);
if strcmpi(type,'poi')
    Qtype='quasi-likelihood';
elseif strcmpi(type,'poi_zero')
    Qtype='poi_zero';
elseif strcmpi(type,'norm')
    Qtype='Normal';
elseif strcmpi(type,'bin')&&options.qtype==1
    Qtype='Deviance';
elseif strcmpi(type,'bin') && options.qtype==2
    Qtype='Exponential';
```

```
else
    warning(message('Wrong type.'));
end

if options.standardize
    mean_x = mean(x);                              % row vector
    x_c = bsxfun(@minus, x, mean_x);           % centers x
    norm_x = sqrt(sum(x_c.^2)/n);          % row vector
    x_stand = bsxfun(@rdivide, x_c, norm_x);   % scales x
else
    x_stand = x;
    mean_x = zeros(1, p);                     % row vector
    norm_x = ones(1,p).*(sum(x_stand.^2)>0); % row vector
end

wlambda = lambda*weights_pen;
cut=4*wlambda;
sign=(wlambda < Inf & norm_x' ~= 0);
x_stand=x_stand(:, sign);
[n, p_sig] = size(x_stand);
group_sig=group.ind(sign);

%%%%%% initial values for iteration
```

```
beta0 = init_beta(1) + mean_x*init_beta(2:end);
beta = init_beta(2:end).*norm_x';
beta = beta(sign);
%-----------------------------------------------
dd=ones(group.num,1);
d_spread= zeros(length(group_sig),1);
loc_ind=0;
for j=1:group.num
    d_spread(loc_ind+1:loc_ind+sum(group_sig==j))...
     =dd(j)*ones(sum(group_sig==j),1);
    loc_ind=loc_ind+sum(group_sig==j);
end
converge=false;
run=0;
alpha=beta./d_spread;

Beta0 = beta;
while (~converge && run <= options.maxit && max(abs(beta))<=1e3)
    %%%% Iteration of w and Z
    run = run + 1;

    if strcmp(Qtype,'poi_zero')
        [q1,q2]=BD(Qtype,x_stand,y,[beta0;beta],pii);
```

```
else
    [q1,q2]=BD(Qtype,x_stand,y,[beta0;beta]);
end
w=q2;
theta=beta0+x_stand*beta;
Z=theta-q1./max(q2,options.zero_eps);
weight_loss = w/n;
a00=0;
change=true;
run1=0;
Alpha0=alpha;
wmean_x0= weight_loss'*x_stand/sum(weight_loss);
wcenter_x=x_stand;
active_set = 1:p_sig;
x_d=bsxfun(@times,wcenter_x,d_spread');
wnorm_x = sum(bsxfun(@times, x_d.^2, weight_loss))...
+options.zero_eps;
r_hat=Z-a00-x_d*Alpha0;

while (change && run1 <= options.maxit)
    run1 = run1 + 1;
    change = false;
    non_zero_set = zeros(1,length(active_set));
```

```
i=0;
for j = active_set
    a_j0 = alpha(j);
    alphadiff = (weight_loss.*r_hat)'*x_d(:,j);
    a_j = soft_thres(alphadiff + a_j0*wnorm_x(j), ...
     wlambda(j))/wnorm_x(j);
    a_j_tem=soft_thres(alphadiff + a_j0*wnorm_x(j),...
     0)/wnorm_x(j);
    if abs(a_j_tem)> cut(j)
        if abs(a_j)>cut(j)
            a_j=a_j_tem;
        else
            val_1=0.5*sum(weight_loss.*(r_hat+...
             x_d(:,j)*a_j_tem).^2)/n...
             +wlambda(j)*cut(j);
            val_2=0.5*sum(weight_loss.*(r_hat+...
             x_d(:,j)*a_j).^2)/n...
             +wlambda(j)*a_j;
            if val_1<val_2
                a_j=a_j_tem;
            end
        end
    end
```

```
    if abs(a_j) > options.zero_eps
        i=i+1;
        non_zero_set(1,i) = j;
        %non_zero_set = [non_zero_set, j];
    else
        a_j = 0;
    end

    diff = a_j0-a_j;

    if abs(diff) > options.thresh
        change=true;
    end

    alpha(j) = a_j;
    r_hat = r_hat + x_d(:,j)*diff;
end

if i>0
    active_set=non_zero_set(1,1:i);
else
    active_set=[];
```

```
    end
end
a00=a00-wmean_x0*alpha;

run2=0;
x_alpha=bsxfun(@times,wcenter_x,alpha');
x_A=zeros(n,group.num);
for j=1:group.num
    x_A(:,j)=sum(x_alpha(:,group_sig==j),2);
end
wnorm_x = sum(bsxfun(@times, x_A.^2, weight_loss))...
 +options.zero_eps;
change=true;
D0=dd;

r_hatd=Z-a00-x_A*D0;
active_group = 1:group.num;
while (change && run2<=options.maxit)
    run2=run2+1;
    change=false;
    non_zero_group = zeros(1,length(active_group));
    ii=0;
    for j = active_group
```

```
    d_j0 =dd(j);
    ddiff = (weight_loss.*r_hatd)'*x_A(:,j);
    d_j = (ddiff + d_j0*wnorm_x(j)-1/n)/wnorm_x(j);
    if d_j >0
        ii=ii+1;
        non_zero_group(1,ii) = j;
    else
        d_j = 0;
    end

    diff = d_j0-d_j;

    if abs(diff) > options.thresh
        change=true;
    end

    dd(j) = d_j;
    r_hatd = r_hatd + x_A(:,j)*diff;
end

if ii>0
    active_group=non_zero_group(1,1:ii);
else
```

```
            active_group=[];
        end
    end
    d_spread=zeros(length(group_sig),1);
    loc_ind2=0;
    for j=1:group.num
        d_spread(loc_ind2+1:loc_ind2+sum(group_sig==j))...
         =dd(j)*ones(sum(group_sig==j),1);
        loc_ind2=loc_ind2+sum(group_sig==j);
    end

    beta= alpha.*d_spread;
    converge= (max(abs(beta-Beta0))< options.thresh);
    Beta0=beta;
    %beta0=wmean_Z-wmean_x0*beta;
    beta0=a00;
end

iter=run;

hat_beta = zeros(p,1);

if options.standardize
```

```
    beta_tem = beta./norm_x(sign)';
    hat_a_0 = beta0 - mean_x(sign)*beta_tem;

    hat_beta(sign) = beta_tem;
    hat_beta = [hat_a_0; hat_beta];
else
    hat_beta(sign) = beta;
    hat_beta = [beta0; hat_beta];
end
```

References

Bickel, P., Y. Ritov, and A. Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* 37:1705–1732.

Bondell, H. D., and B. J. Reich. 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics* 64:115–123.

Bregman, L. M. 1967. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R Computational Mathematics and Mathematical Physics* 7:620–631.

Cheng, C., and C. M. Zhang. 2015. Penalized bregman divergence for group variable selection. Submitted.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *Ann. Statist.* 32:407–499.

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96:1348–1360.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society Series B* 70:849–911.

———. 2010. A selective overview of variable selection in high dimensinal feature space. *Statistica Sinica* 20:101–148.

———. 2011. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* 57:5467–5484.

Fan, J., and H. Peng. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* 32:928–961.

Fan, J., and R. Song. 2010. Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.* 38:3567–3604.

Friedman, J., T. Hastie, H. Hoefling, and R. Tibshirani. 2007. Pathwise coordinate optimization. *Ann. of Applied Statist.* 1:302–332.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33:1–22.

Geng, Z., S. Wang, Monahan Yu M., V. O.P., Champion, and G. Wahba. 2015. Group variable selection via convex log-exp-sum penalty withapplication to a breast cancer survivor study. *Biometrics* 71:53–62.

Gerstein, G. L., and D. H. Perkel. 1969. Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* 164:828–830.

H., Wang, Li R., and C. L. Tsai. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94:553–568.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning*. Springer.

Huang, J., S. Ma, H. Xie, and C. Zhang. 2009. A group bridge approach for variable selection. *Biometrika* 96:339–355.

Jiang, Y., and C. M. Zhang. 2013. High-dimensional regression and classification under a class of convex loss functions. *Statistics and Its Interface* 6:285–299.

Johnson, B. 2013. High resolution urban land cover classification using a competitive multi-scale object-based approach. *Remote Sensing Letters* 4:131–140.

Johnson, B., and Z. Xie. 2013. Classifying a high resolution image of an urban area using super-object information. *ISPRS J. of Photogrammetry and Remote Sensing* 83:40–49.

Lv, J., and Y. Fan. 2009. A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* 37:3498–3528.

Perkel, D. H., G. L. Gerstein, and G. P. Moore. 1967. Neuronal spike trains and stochastic point processes: Ii. simultaneous spike trains. *Biophysical Journal* 7: 419–440.

Rajaram, S., T. Graepel, and R. Herbrich. 2005. Poisson-networks: A model for structured point processes. In *Proceedings of the tenth international workshop on artificial intelligence and statistics*.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani. 2013. A sparse-group lasso. *Journal of Computational and Graphical Statistics* 22:231–245.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B* 58:267–288.

Tsanas, A., M. A. Little, C. Fox, and L. O. Ramig. 2014. Objective automatic assessment of rehabilitative speech treatment in parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22:181–190.

Witten, D. M., A. Shojaie, and F. Zhang. 2014. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* 56:112–122.

Wu, T. T., and K. Lange. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann. of Applied Statist.* 1:224–244.

Yuan, M., and Y. Lin. 2006. Model selection and estimation in regression with grouped variable. *Journal of The Royal Statistical Society Series B* 68:49–67.

Zhang, C. M. 2010. Statistical inference of minimum bd estimators and classifiers for varying-dimensional models. *Journal of Multivariate Analysis* 101:1574–1593.

Zhang, C. M., Y. Chai, M. H. Gao, D. M. Devilbiss, and Z. J. Zhang. 2015. Statistical learning of neuronal functional connectivity. Submitted.

Zhang, C. M., Y. Jiang, and Y. Chai. 2010. Penalized bregman divergence for large-dimensional regression and classification. *Biometrika* 97:551–566.

Zhang, C. M., Y. Jiang, and Z. Shang. 2009. New aspects of bregman divergence in regression and classification with parametric and nonparametric estimation. *The Canadian Journal of Statistics* 37:119–139.

Zhang, C. M., Z. Zhang, and Y. Chai. 2011. Penalized bregman divergence estimation via coordinate descent. *JIRSS* 10:125–140.

Zhao, M., A. Batista, J. P. Cunningham, C. Chestek, Z. Rivera-Alvidrez, R. Kalmar, S. Ryu, K. Shenoy, and S. Iyengar. 2012. An l_1-regularized logistic model for detecting short-term neuronal interactions. *Journal of Computational Neuroscience* 32:479–497.

Zhao, P., G. Rocha, and B. Yu. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* 37:3468–3497.

Zhao, P., and B. Yu. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7:2541–2563.

Zhou, N., and J. Zhu. 2010. Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface* 3:557–574.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B* 67:301–320.