DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1176
June 2, 2014

Variable Selection via Penalized Likelihood

Zhigeng Geng[1]
Department of Statistics,
University of Wisconsin, Madison, WI

[1] Research supported in part by NIH Grant EY09946 and NSF Grant DMS-1308877.

VARIABLE SELECTION VIA PENALIZED LIKELIHOOD

by

Zhigeng Geng

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 06/02/2014

The dissertation is approved by the following members of the Final Oral Committee:
Grace Wahba, IJ Schoenberg-Hilldale Professor, Statistics
Sijian Wang, Assistant Professor, Statistics
Wei-Yin Loh, Professor, Statistics
Stephen Wright, Professor, Computer Sciences
Karl Rohe, Assistant Professor, Statistics

ACKNOWLEDGMENTS

It was a summer afternoon, warm and sunny, when I first met my advisor, Grace Wahba, in her office. At that time, I just came to the United States from China for a few weeks. I was quite nervous and could not even speak English fluently in front of such a famous statistician. But Grace was like your own grandmother and the summer weather here in Madison, so nice and comfortable to speak with. It is my great honor and privilege to work with her and learn from her so closely since then. Grace is a brilliant and passionate statistician and she has very positive life attitude. She has set up the best role model for me to follow in my career. I would like to express my deepest gratitude to Grace, for her introduction into the field of penalized likelihood methods, and her guidance, support, encouragement and patience through the entire course of my PhD study.

I am thankful to my coadvisor, Professor Sijian Wang. He is extremely smart and quick-minded. Discussions with Sijian have always been inspiring and challenging. He greatly stimulated my interests in variable selection and guided the direction of my research. This thesis would not be possible without discussions with and insights from Sijian.

I want to thank Professor Stephen Wright, Professor Wei-Yin Loh, and Professor Karl Rohe for their service in my thesis committee. I would like to thank Professor Stephen Wright for his guidance in computation. I was in his optimization class and I learned a lot about optimization from him. One of the proposed models in this thesis is solved with efficient optimization algorithm following the idea in his paper. I am thankful to Professor Wei-Yin Loh and Professor Karl Rohe for their insightful questions and valuable suggestions which helped improve the thesis greatly. I was in Professor Loh's decision tree class and it turns out to be one of the most useful classes I have ever taken. Professor Rohe is one of the Thursday Group meeting hosts and I learned a lot about social network from his presentations.

I want to thank Professor Ming Yuan for his sharp questions and helpful suggestions. I want to thank Professor Menggang Yu for his contribution to my research and our paper. I want to thank many other the professors who contributed to my

education, including Professor Ming Yuan, Professor Zhengjun Zhang, Professor Rick Nordheim, Professor Jun Shao, Professor Kam-Wah Tsui, Professor Chunming Zhang and Professor Sunduz Keles.

Our Thursday Group meeting has been a great experience during the course of my study, and I would like to thank the former and current Thursday Group meeting members: Xiwen Ma, Bin Dai, Shilin Ding, Tai Qin, Jing Kong, Luwan Zhang, Han Chen and Shulei Wang. I would like to thank my friends in Madison and the fellow students in the Statistics Department: Yi Chai, Minjing Tao, Xiao Guo, Ji Liu, Xu Xu, Bin Zhu, Yajie Zhao, Qiurong Cui, Vincent Chan, Zhuang Wu, Jiwei Zhao, Guangde Chen, Jiajie Chen, Xinxin Yu, Rui Tang, Wenwen Zhang, Lie Xiong, Chen Zuo, Hao Zheng, Sheng Zhang, Yaoyao Xu, Ning Leng, Xu He, and many many others to name. All these wonderful friends make my stay in Madison enjoyable.

And last, but not least, I would like to thank my family, for their unconditional support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Variable selection via penalized likelihood plays an important role in high dimensional statistical modeling and it has attracted great attention in recent literature. This thesis is devoted to the study of variable selection problem. It consists of three major parts, all of which fall within the framework of penalized least squares regression setting.

In the first part of this thesis, we propose a family of nonconvex penalties named the K-Smallest Items (KSI) penalty for variable selection, which is able to improve the performance of variable selection and reduce estimation bias on the estimates of the important coefficients. We fully investigate the theoretical properties of the KSI method and show that it possesses the weak oracle property and the oracle property in the high-dimensional setting where the number of coefficients is allowed to be much larger than the sample size. To demonstrate its numerical performance, we applied the KSI method to several simulation examples as well as the well known Boston housing dataset. We also extend the idea of the KSI method to handle the group variable selection problem.

In the second part of this thesis, we propose another nonconvex penalty named Self-adaptive penalty (SAP) for variable selection. It is distinguished from other existing methods in the sense that the penalization on each individual coefficient takes into account directly the influence of other estimated coefficients. We also thoroughly study the theoretical properties of the SAP method and show that it possesses the weak oracle property under desirable conditions. The proposed method is applied to the glioblastoma cancer data obtained from The Cancer Genome Atlas.

In many scientific and engineering applications, covariates are naturally grouped. When the group structures are available among covariates, people are usually interested in identifying both important groups and important variables within the selected groups. In statistics, this is a group variable selection problem. In the third part of this thesis, we propose a novel Log-Exp-Sum(LES) penalty for group variable selection. The LES penalty is strictly convex. It can identify important groups as well as select important variables within the group. We develop an efficient group-level coordinate descent algorithm to fit the model. We also derive non-asymptotic error bounds and asymptotic group selection consistency for our method in the high-dimensional setting. Numerical results demonstrate the good performance of our method in both variable selection and prediction. We applied the proposed method to an American Cancer Society breast cancer survivor dataset. The findings are clinically meaningful and may help design intervention programs to improve the qualify of life for breast cancer survivors.

1 INTRODUCTION

Variable selection through optimizing the penalized ordinary least squares has been an active research area in the past decade. With proper choices of selection methods and under appropriate conditions, we are able to build consistent models to select variables and estimate coefficients simultaneously, to avoid model overfitting, and to obtain satisfactory prediction accuracy. In my thesis, we consider the variable selection problem under the usual regression setting: we have training data, $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$, where $\mathbf{x}_i$ is a vector of values with p covariates and y_i is the response. To model the association between response and covariates, we consider the following linear regression:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_j's are regression coefficients. Without loss of generality, we assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation, so the intercept term can be removed in the above regression model.

For the purpose of variable selection, we consider the penalized ordinary least squares (OLS) estimation:

$$\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{1.2}$$

where λ is a non-negative tuning parameter and $p_\lambda(\cdot)$ is a sparsity-induced penalty function which may or may not depend on λ.

Many methods have been proposed and their properties have been thoroughly studied, for example, see LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic-Net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), COSSO (Lin and Zhang, 2006), SICA (Lv and Fan, 2009), MCP (Zhang, 2010), truncated L_1 (Shen et al., 2011) and SELO (Dicker et al., 2011).

We review some of the popular penalties here.

LASSO (Tibshirani, 1996) is one of the the most popular and widely studied methods. It penalizes the L_1-norm of the coefficients:

$$\lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) = \lambda \sum_{i=1}^{p} |\beta_i|. \tag{1.3}$$

Meinshausen and Bühlmann (2006), Zhao and Yu (2006) and Zhang and Huang (2008) studied the conditions under which LASSO can achieve model selection consistency in high dimensional setting. Fan and Lv (2011) recently showed that, as a boundary case of the folded-concave penalty, LASSO possesses the weak oracle property. This weak oracle property was introduced by Lv and Fan (2009). It means that when $n, p \to \infty$, with probably tending to 1, the penalty will identify the sparsity structure of true coefficients, and the L_∞ loss on the nonzero elements of true coefficients is consistent under a rate slower than $\sqrt{n}$.

Zou (2006) proposed the adaptive LASSO, which is a weighted version of LASSO:

$$\lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) = \lambda \sum_{i=1}^{p} w_i |\beta_i|. \tag{1.4}$$

When dimension p is fixed, Zou (2006) showed that with a properly chosen set of weights w_i's, adaptive LASSO has the oracle property. Their results were extended by Huang et al. (2008) in the high dimensional setting where dimension $p = O(exp(n^\alpha))$, for some $\alpha \in (0, 1)$. When $p < n$, a typical weight in use is $w_i = |\widehat{\beta}_i^{ols}|^{-1}$, where $\widehat{\beta}_i^{ols}$ is the unpenalized OLS estimator.

SCAD (Fan and Li, 2001) is another popular method. The penalty function $p_\lambda(\cdot)$ is the continuous differentiable function defined by its derivative on $[0, \infty)$:

$$p'_\lambda(\theta) = I(\theta \leqslant \lambda) + \frac{max\{a\lambda - \theta, 0\}}{(a-1)\lambda} I(\theta > \lambda). \tag{1.5}$$

Here $a > 2$ is another tuning parameter. Then, the penalty function $p_\lambda(\cdot)$ can be obtained as:

$$p_\lambda(|\beta_i|) = (|\beta_i| - \frac{|\beta_i|^2}{2a\lambda}) * I(|\beta_i| \leqslant a\lambda) + \frac{1}{2}a\lambda * I(|\beta_i| > a\lambda). \tag{1.6}$$

Fan and Li (2001) showed that there exists a local optimizer of the SCAD penalty which has the oracle property in the finite-dimensional setting, i.e., it performs as well as if the true model is specified in advance. Their results were extended by Fan and Peng (2004) where dimension p is allowed to grow at the rate $o(n^{1/5})$. Fan and Lv (2011) further extended the results and showed, under desirable conditions, there exists a local optimizer of SCAD which possesses the oracle property even when $p = O(exp(n^\alpha))$, for some $\alpha \in (0, 1)$.

Zhang (2010) proposed the MCP method which is given by:

$$p_\lambda(|\beta_i|) = (|\beta_i| - \frac{|\beta_i|^2}{2\gamma\lambda}) * I(0 \leqslant |\beta_i| \leqslant \gamma\lambda) + \frac{\gamma\lambda}{2} I(|\beta_i| > \gamma\lambda) \tag{1.7}$$

Here $\gamma > 0$ is another tuning parameter. Zhang (2010) showed that, with probability tending to 1, MCP can select the correct model if λ and γ satisfy certain conditions. The results from Fan and Lv (2011) is applicable to MCP as well. Therefore, under desirable conditions, there exists a local optimizer of MCP which possesses the oracle property even when $p = O(exp(n^\alpha))$, for some $\alpha \in (0, 1)$.

In many regression problems, the predictor are naturally divided into meaningful groups based on some prior domain knowledge. For example, when analyzing genomic data, one can group genes into known biological pathways, or group the SNPs within the intragenic and regulatory regions of a given gene into a group and perform genebased association analysis. Besides the group structure based on the domain knowledge, there are also many group structures based on model. For example, in ANOVA factor analysis, a factor may have several levels and can be expressed via several dummy variables, then the dummy variables corresponding to the same factor form a natural group. Similarly, in additive models, each original predictor may be expanded into different order polynomials or a set of basis functions, then these polynomials (or basis functions) corresponding to the same original predictor form a natural group. We are interested in identifying important groups and important variables within the selected groups that are related to the responses. In statistics, this is a group variable selection problem.

When the group structure is assumed, we incorporate the group information and consider a slight modification of the linear regression problem (1.1):

$$y_i = \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj} \beta_{kj} + \epsilon_i, \; i = 1, \ldots, n, \tag{1.8}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_{kj}'s are regression coefficients. We denote $\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kp_k})'$ to be the vector of regression coefficients for covariates in the kth group. We again assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation.

For the purpose of group variable selection, we consider the penalized ordinary least square (OLS) estimation:

$$\min_{\beta_{kj}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj} \beta_{kj} \Big)^2 + \lambda \sum_{k=1}^{K} p(\boldsymbol{\beta}_k), \tag{1.9}$$

where $p(\cdot)$ is a sparsity-induced penalty function, λ is a non-negative tuning parameter and K is the total number of groups.

Several methods have addressed the group variable selection problem in literature.

Yuan and Lin (2006) proposed the following group LASSO penalty which is to penalize the L_2-norm of the coefficients within each group:

$$p(\boldsymbol{\beta}_k) = \sqrt{\beta_{k1}^2 + \cdots + \beta_{kp_k}^2}. \tag{1.10}$$

Due to the singularity of $\|\boldsymbol{\beta}_k\|_2$ at $\boldsymbol{\beta}_k = \mathbf{0}$, some estimated coefficient vector $\hat{\boldsymbol{\beta}}_k$ will be exactly zero and hence the corresponding kth group is removed from the fitted model.

Zhao et al. (2006) proposed penalizing the L_∞-norm of $\boldsymbol{\beta}_k$:

$$p(\boldsymbol{\beta}_k) = \max\{\beta_{k1}, \ldots, \beta_{kp_k}\}. \tag{1.11}$$

The L_∞-norm of $\boldsymbol{\beta}_k$ is also singular at $\boldsymbol{\beta}_k = 0$. Therefore, some estimated coefficient vector $\hat{\boldsymbol{\beta}}_k$ will be exactly zero.

We can see that both L_2-norm and L_∞-norm are singular when the whole vector $\boldsymbol{\beta}_k$ is zero. Therefore, some estimated coefficient vector $\hat{\boldsymbol{\beta}}_k$ will be exactly zero and hence the corresponding kth group will be removed from the fitted model. This is the reason that the L_2-norm and L_∞-norm methods can effectively remove unimportant groups. But a possible limitation with these two methods is that they select variables in an "all-in-all-out" fashion, i.e., when one variable in a group is selected, all other variables in the same group are also selected. In other words, they cannot conduct the within group variable selection. This is because once a component of $\boldsymbol{\beta}_k$ is non-zero, the two norm functions are no longer singular. Hence they cannot conduct the within group variable selection.

In many practical problems, however, people want to keep the flexibility of selecting variables within a group. For example, when a group of genes is related to certain kind of disease, it does not necessarily mean all the individual genes in the same group are related to this disease. People may want to not only remove unimportant groups effectively, but also identify important individual genes within important groups as well. To achieve the goal, Huang et al. (2009) and Zhou and Zhu (2010) independently proposed the group bridge penalty and the hierarchical LASSO penalty.

Huang et al. (2009) proposed the following group bridge penalty:

$$p(\boldsymbol{\beta}_k) = \Big(|\beta_{k1}| + \cdots + |\beta_{kp_k}|\Big)^\gamma, \tag{1.12}$$

where $0 < \gamma < 1$ is another tuning parameter.

Zhou and Zhu (2010) independently proposed a hierarchical LASSO penalty.

This penalty decomposes $\beta_{kj} = \gamma_k \theta_{kj}$ and considers:

$$p(\boldsymbol{\beta}_k) = |\gamma_k| + \sum_{j=1}^{p_k} |\theta_{kj}|. \tag{1.13}$$

When the groups are not overlapped, the hierarchical LASSO penalty is equivalent to the group bridge penalty with $\gamma = 0.5$. We can see that these two penalties are singular at both $\boldsymbol{\beta}_k = \mathbf{0}$ and $\beta_{kj} = 0$ and hence is able to conduct both group selection and within group selection. However, one possible drawback of the two methods is that their penalty functions are no longer convex. This non-convexity may cause numerical problems in practical computation, especially when the numbers of groups and covariates are large.

Simon et al. (2012) proposed the sparse group LASSO penalty:

$$p(\boldsymbol{\beta}_k) = s\sqrt{\beta_{k1}^2 + \cdots + \beta_{kp_k}^2} + (1-s)\sum_{j=1}^{p_k} |\beta_{kj}|, \tag{1.14}$$

where $0 < s < 1$ is another tuning parameter. We can see that, by mixing the LASSO penalty and group LASSO penalty, the sparse group LASSO penalty is convex and is able to conduct both group and within group selection.

The rest of this thesis is organized as follows.

In Chapter 2, we propose a family of nonconvex penalties named the K-Smallest Items (KSI) penalties for variable selection. It is intuitive and distinguished from other existing methods in the sense that it has the flexibility of penalizing only a few coefficients with the smallest magnitudes and placing no penalty at all on the other larger coefficients. Thus it is able to improve the performance of variable selection and reduce estimation bias on the estimates of the important coefficients. The theoretical properties of the KSI method is fully investigated and it is shown to possess the weak oracle property and the oracle property in the high-dimensional setting where the number of coefficients is allowed to be much larger than the sample size. This KSI method is applied to several numerical examples as well as the well known Boston housing dataset. We further extend the idea of KSI penalty to handle group variable selection problem and study the group selection consistency and estimation consistency of the KSI estimators.

In Chapter 3, we propose a nonconvex penalty named Self-adaptive penalty (SAP) for variable selection, which is able to reduce estimation bias. It is distinguished from other existing methods in the sense that the penalization on each individual coefficient takes into account directly the influence of other coefficient estimators. We also thoroughly study the theoretical properties of the SAP method and show that it possesses the weak oracle property under desirable conditions. The proposed method is applied to the glioblastoma cancer data obtained from The Cancer Genome Atlas (TCGA).

In Chapter 4, we propose a new Log-Exp-Sum (LES) penalty for group variable selection. This new penalty is convex, and it can perform variable selection at both group level and within-group level. The theoretical properties of our proposed method are thoroughly studied. We establish both the finite sample error bounds and asymptotic group selection consistency of our LES estimator. The proposed method is applied to the ACS breast cancer survivor dataset.

Finally, we conclude the thesis in Chapter 5.

2 VARIABLE SELECTION VIA THE K-SMALLEST ITEMS PENALTIES FAMILY

2.1 Motivation

In this chapter, we consider the variable selection problem under the usual regression setting: we have training data, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i$ is a vector of values with p covariates and y_i is the response. To model the association between response and covariates, we consider the following linear regression:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_j's are regression coefficients. Without loss of generality, we assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation, so the intercept term can be removed in the above regression model.

For the purpose of variable selection, we consider the penalized ordinary least squares (OLS) estimation:

$$\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.2}$$

where λ is a non-negative tuning parameter and $p_\lambda(\cdot)$ is a sparsity-induced penalty function which may or may not depend on λ.

We have reviewed some of the most popular variable selection methods in the Chapter 1, including LASSO, SCAD and MCP. One possible limitation of these methods is that, the penalty function $p_\lambda(\cdot)$ is applied to every single regression coefficient β_j regardless of its magnitude. This may introduce unnecessary estimation bias to the estimates of the important coefficients. In many real problems, however, we may want to have the flexibility of placing more penalty on the unimportant variables, while putting less or even no penalty on the important ones. Doing so will help to remove the unimportant variables and reduce the estimation bias on the estimates of the important coefficients.

The weighted versions of these variable selection methods, adaptive LASSO, for example, will help to alleviate the estimation bias issue to some extend. However, they usually lead to a two-step procedure, where some properly chosen weights are first obtained, then a penalized ordinary least squares model is fitted. This may require extra computation efforts. Moreover, they do not completely eliminate the bias issue because the weight obtained from the traditional ordinary least squares estimator is usually not zero.

For the purpose of variable selection and estimation bias reduction, we propose a family of nonconvex penalties named the K-Smallest Items (KSI) penalties for variable selection. It is intuitive and distinguished from other existing methods in the sense that it has the flexibility of penalizing only K coefficients with the smallest magnitudes and placing no penalty at all on the other larger coefficients. Here these K smallest coefficients in absolute value are automatically selected and penalized by the KSI penalties, and no“two-step procedure” is needed. Thus it is able to improve the performance of variable selection and reduce estimation bias on the estimates of the important coefficients.

The chapter is organized as follows. In Section 2.2, we propose the KSI penalties family and present the corresponding algorithm. In Section 2.3, we investigate the theoretical properties of our proposed method and show that it possesses the weak oracle property and the oracle property in the high-dimensional setting where the number of coefficients is allowed to be much larger than the sample size. In Section 4, we present the simulation results. In Section 2.5, we apply the proposed method to the well known Boston housing dataset. In Section 2.6, we discuss the extension of the KSI method to handle group variable selection problem. Finally we conclude this chapter with Section 2.7.

2.2 Method

KSI Penalties Family

We propose the following K-Smallest Items (KSI) penalties family:

$$\lambda * \sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|). \tag{2.3}$$

Here λ is a nonnegative tuning parameter. K is a positive integer between 1 to p which controls the number of regression coefficients one wants to penalize. In practice, K is treated as another tuning parameter. $\beta_{[j]}$'s are the ordered covariates of $\boldsymbol{\beta}$ satisfies: $0 \leqslant |\beta_{[p]}| \leqslant |\beta_{[p-1]}| \leqslant \ldots \leqslant |\beta_{[1]}|$. And $p_\lambda(\cdot)$ is a sparsity-induced penalty function which may or may not depend on λ. Hereafter we assume $p_\lambda(\cdot)$ satisfies the following condition as is suggested in Fan and Lv (2011).

Condition $(C1)$. For any given λ, $p_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $p'_\lambda(t) \triangleq \partial p_\lambda(t)/\partial t$ with $p'_\lambda(t) > 0$, for $t \in (0, \infty)$. In addition, $p'_\lambda(t)$ is increasing in $\lambda \in (0, \infty)$ and $p'_\lambda(0+)$ is independent of λ. Furthermore, $p''_\lambda(t) \triangleq \partial^2 p_\lambda(t)/\partial t^2$ is continuous with respect to t almost everywhere.

From (2.3), it is easy to see that the KSI penalties family only penalizes the K coefficients with the smallest magnitudes and puts no penalty on the other larger coefficients. When $K = p$, i.e., p-smallest items are penalized, the KSI method reduces to the usual case as is in (2.2). More flexibility is introduced into (2.2) when K is set to be $K \neq p$. With some properly chosen values of (λ, K) and under desirable conditions, the KSI method is able to remove the unimportant variables while introduce small or even no estimation bias to the estimates of the important coefficients. This idea can be applied to many exciting penalty functions, such as LASSO and SCAD. Therefore, the KSI penalties family covers a large collection of potential members.

The number of nonzero elements in the solution of the KSI method is at least $p - K$. This is readily justified by the observation that $\beta_{[1]}, \ldots, \beta_{[p-K]}$ are nonzeros because they are not penalized. In fact, $\beta_{[1]}, \ldots, \beta_{[p-K]}$ are the ordinary least squares estimates given the values of $\beta_{[p-K+1]}, \ldots, \beta_{[p]}$.

The KSI penalties family is generally nonconvex because of the ordering constraint in $\beta_{[j]}$'s, even if $p_\lambda(|\cdot|)$ is convex. To see this, consider the case when $p_\lambda(|\cdot|) = |\cdot|$. If $K = p$, the KSI penalty is the well known convex LASSO penalty. If $K < p$, however, it is not necessary convex as is shown in the following example:

Example 2.1: Let $\hat{\boldsymbol{\beta}} = argmin_{\beta_1,\beta_2}\frac{1}{2}(1.5-\beta_1)^2 + \frac{1}{2}(1-\beta_2)^2 + |\beta_{[2]}|$. It is easy to verify $\hat{\boldsymbol{\beta}} = (1.5, 0)^T$ is the global minimizer; while $\hat{\boldsymbol{\beta}} = (0.5, 1)^T$ is a strict local minimizer which is different from the global minimizer.

Algorithm

In this section, we present the algorithm for solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_j} Q(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|). \tag{2.4}$$

We will apply gradient projection algorithm (Goldstein, 1964; Levitin and Polyak, 1966) which combines a proximal step with a gradient step.

The details of the algorithm are as follows:

Step 1 (Initialization): Set iteration number $m = 0$, initialize $\boldsymbol{\beta}$ with some initial value $\boldsymbol{\beta}^{(0)}$.

Step 2 (Gradient Step): Let $\mathbf{u} = \boldsymbol{\beta}^{(m)} - s * \mathbf{X}'(\mathbf{X} * \boldsymbol{\beta}^{(m)} - \mathbf{Y})/n$. s here is some prespecified step length.

Step 3 (Sort $\mathbf{u}$): Sort $\mathbf{u}$ according to the absolute value. Let idx_1 be set of the indices for the K-smallest elements of $\mathbf{u}$ in absolute value, and idx_2 be the rest.

Step 4 (Proximal Step and Update $\boldsymbol{\beta}^{(m+1)}$): For each index j in set idx_1, solve the following single variable optimization problem:

$$\beta_j^{(m+1)} = \arg\min_v \frac{1}{2}(v - u_j)^2 + s * \lambda * p_\lambda(|v|). \tag{2.5}$$

For each index j in set idx_2, we simply set $\beta_j^{(m+1)} = u_j$.

Step 5 (Termination): Terminate the iteration if $\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|$ is small, or the changes of the objective function value $|Q(\boldsymbol{\beta}^{(m+1)}) - Q(\boldsymbol{\beta}^{(m)})|$ is small. Otherwise, set $m \leftarrow m+1$, and repeat from step 2.

Note that in Step 4, there is explicit solution for the single variable optimization problem for both LASSO and SCAD penalty. In general, (2.5) should be an easy optimization problem so we skip the details for solving it here. The validity of Step 3 and Step 4 in the algorithm is guaranteed by the following proposition. After Step 4, we have $|\beta_h^{(m+1)}| \geqslant |\beta_i^{(m+1)}|$, $\forall\, h \in idx\_2$ and $\forall\, i \in idx\_1$.

Proposition 2.1. *Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$, then Step 3 and Step 4 presented in the algorithm solve the following optimization problem:*

$$\min_{\beta_j} \frac{1}{2}\sum_{j=1}^{p}(\beta_j - u_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|). \tag{2.6}$$

The proof is given in the Appendix for completeness.

Pathwise Optimization and Warm Starts

In practice, we usually need to compute $\hat{\boldsymbol{\beta}}$ for many different pairs of regularization parameter values (λ, K). Following Friedman et al. (2010a), we apply the idea of solving $Q(\boldsymbol{\beta})$ along a path of values for (λ, K) and using the previous estimates as warm starts. To be specific, for a fixed K and a decreasing sequence $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$, we first solve $Q_{(\lambda_1,K)}(\boldsymbol{\beta})$ with parameters (λ_1, K) using a reasonable starting point $\boldsymbol{\beta}^{warm}$. Suppose $\hat{\boldsymbol{\beta}}_1$ is the current minimizer of $Q_{(\lambda_1,K)}(\boldsymbol{\beta})$. At the next step, $\hat{\boldsymbol{\beta}}_1$ is used as a warm start when we solve $Q_{(\lambda_2,K)}(\boldsymbol{\beta})$ with parameters (λ_2, K). And so on and so forth.

This scheme exploits the idea of warm starts along the decreasing sequence of λ, and leads to a faster, more stable and even more accurate algorithm. We have simulation examples where it is faster to compute the path down to a small λ and a smaller objective function value $Q(\hat{\boldsymbol{\beta}})$ can be obtained, than starting directly at that value for λ.

Choice of $\boldsymbol{\beta}^{warm}$

The choice of a reasonable starting point $\boldsymbol{\beta}^{warm}$ at the very beginning of the algorithm deserves some discussions. According to our simulation experience, by setting $\boldsymbol{\beta}^{warm} = \mathbf{0}$, the algorithm will generally select less variables along the path of λ than otherwise. It also works well in terms of prediction in simulations.

On the other hand, in the real data analysis, by applying Sure Independence Screening (SIS) method (Fan and Lv, 2008a), we can sometimes produce an optimum estimator which gives better prediction performance on future data. To be more specific, for a given K, we first set $\boldsymbol{\beta}^{warm} = \mathbf{0}$. We then apply SIS to select $p - K$ variables and call the index set for these $p - K$ variables $\mathcal{A}$. We calculate the ordinary least squares estimates based on the submatrix $\mathbf{X}_{\mathcal{A}}$ and call it $\boldsymbol{\beta}_{SIS}^{ols}$, where $\boldsymbol{\beta}_{SIS}^{ols} \in \mathbb{R}^{p-K}$. Finally, we set the corresponding values in $\boldsymbol{\beta}^{warm}$ to be $\boldsymbol{\beta}_{\mathcal{A}}^{warm} = \boldsymbol{\beta}_{SIS}^{ols}$.

In order to be consistent, in both simulation section and realy data analysis section, the $\boldsymbol{\beta}^{warm}$'s are obtained using the SIS approach.

2.3 Theoretical Results

In this section, we study the theoretical properties of the KSI estimators. It is generally difficult to study the global minimizer of a nonconvex function, so as is common in literature, we focus on local minimizer. In the remaining of this section, we first present several conditions which characterize the behavior of a local minimizer. We then show under several assumptions, there exists a local minimizer of KSI method such that the nonasymptotic weak oracle property is satisfied. Finally we show there exists a local minimizer of KSI method such that the oracle property is satisfied.

Characterization of KSI estimators

Because it is hard to deal with the order constraint in the penalty function, we first connects the KSI method with its un-ordered counterpart. Throughout this chapter, we denote $Q(\boldsymbol{\beta})$ to be the penalized likelihood using KSI penalties family; denote $C(\boldsymbol{\beta})$ to be its corresponding un-ordered counterpart, where $C(\boldsymbol{\beta})$ simply penalizes on the last K covariates of $\boldsymbol{\beta}$ regardless of their values. To be specific, we have:

$$Q(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|); \tag{2.7}$$

$$C(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_j|). \tag{2.8}$$

The following proposition connects the KSI method to its un-ordered counterpart $C(\boldsymbol{\beta})$.

Proposition 2.2.

(a) If $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\boldsymbol{\beta})$, and $\hat{\boldsymbol{\beta}}$ satisfies $|\hat{\beta}_h| \geqslant |\hat{\beta}_i|$, for $\forall\, h, i$, where $1 \leqslant h \leqslant p - K$ and $p - K + 1 \leqslant i \leqslant p$. Then $\hat{\boldsymbol{\beta}}$ is a local minimizer of $C(\boldsymbol{\beta})$.

(b) If $\hat{\boldsymbol{\beta}}$ is a local minimizer of $C(\boldsymbol{\beta})$, and $\hat{\boldsymbol{\beta}}$ satisfies $|\hat{\beta}_h| > |\hat{\beta}_i|$, for $\forall\, h, i$, where $1 \leqslant h \leqslant p - K$ and $p - K + 1 \leqslant i \leqslant p$. Then $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\boldsymbol{\beta})$.

The proof is given in the Appendix for completeness.

According to this proposition, if the order in $\hat{\boldsymbol{\beta}}$ is assumed, a local minimizer of $Q(\boldsymbol{\beta})$ is readily a local minimizer of $C(\boldsymbol{\beta})$. On the other hand, a strict inequality condition is needed for a local minimizer of $C(\boldsymbol{\beta})$ to be a local minimizer of $Q(\boldsymbol{\beta})$. The strict inequality in part (b) ensures that for any vector $\boldsymbol{\beta}$ within a small open ball centered at $\hat{\boldsymbol{\beta}}$, the order of this $\boldsymbol{\beta}$'s covariates is still valid, i.e., the first $p - K$

covariates of β in absolute value are strictly greater than the last K covariates of β in absolute value. Therefore $\sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|) = \sum_{j=p-K+1}^{p} p_\lambda(|\beta_j|)$, and $Q(\boldsymbol{\beta})$ has the same form as $C(\boldsymbol{\beta})$. Then $\hat{\boldsymbol{\beta}}$ is indeed a local minimizer of $Q(\boldsymbol{\beta})$. If the strict inequality fails, a local minimizer of $C(\boldsymbol{\beta})$ is not necessary a local minimizer of $Q(\boldsymbol{\beta})$ as is shown in the following example.

Example 2.2: Let $C(\boldsymbol{\beta}) = \frac{1}{2}(1-\beta_1)^2 + \frac{1}{2}(2-\beta_2)^2 + |\beta_2|$ and $Q(\boldsymbol{\beta}) = \frac{1}{2}(1-\beta_1)^2 + \frac{1}{2}(2-\beta_2)^2 + |\beta_{[2]}|$. Then $\hat{\boldsymbol{\beta}} = (1,1)^T$ is the unique global minimizer of $C(\boldsymbol{\beta})$. However, it is easy to verify $\hat{\boldsymbol{\beta}}$ is not a local minimizer of $Q(\boldsymbol{\beta})$.

In spirit of Theorem 1 in Fan and Lv (2011), we derive the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}$ to be the solution to minimization problem of $Q(\boldsymbol{\beta})$. By the relation of $Q(\boldsymbol{\beta})$ and $C(\boldsymbol{\beta})$ presented in proposition 2.2, we first derive the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}$ to be the local minimizer of $C(\boldsymbol{\beta})$.

Theorem 2.3. *Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$. Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T \in \mathbb{R}^p$, where $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^s$ are the nonzero components of $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}}_2 = \mathbf{0} \in \mathbb{R}^{p-s}$ are the zero components of $\hat{\boldsymbol{\beta}}$. s here is the number of nonzero components of $\hat{\boldsymbol{\beta}}$ and we assume $s \geqslant p-K$. Let X_1 and X_2 denote the submatrices of X formed by first s columns and last $p-s$ columns of X, respectively.*

Then $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $C(\boldsymbol{\beta})$ if the following conditions are satisfied:

$$\frac{1}{n}X_1^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1) - \nabla J_\lambda(\hat{\boldsymbol{\beta}}_1) = 0; \tag{2.9}$$

$$\frac{1}{n}||X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)||_\infty < \lambda p_\lambda'(0+); \tag{2.10}$$

$$\frac{1}{n}X_1^T X_1 + \nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_1) > 0. \tag{2.11}$$

Here $\nabla J_\lambda(\hat{\boldsymbol{\beta}}_1) = (0, \ldots, 0, \lambda p_\lambda'(|\hat{\beta}_{p-K+1}|)sgn(\hat{\beta}_{p-K+1}), \ldots, \lambda p_\lambda'(|\hat{\beta}_s|)sgn(\hat{\beta}_s))^T \in \mathbb{R}^s$; and $\nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_1)$ is a $s \times s$ diagonal matrix given by:

$$diag(0, \ldots, 0, \lambda p_\lambda''(|\hat{\beta}_{p-K+1}|), \ldots, \lambda p_\lambda''(|\hat{\beta}_s|)). \tag{2.12}$$

On the other hand, if $\hat{\boldsymbol{\beta}}$ is a local minimizer to $C(\boldsymbol{\beta})$, then it must satisfy (2.9) - (2.11) *with strict inequalities replaced by nonstrict inequalities.*

The proof is given in the Appendix for completeness.

By Theorem 2.3 and Proposition 2.2, we can derive the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}$ to be the local minimizer to $Q(\hat{\boldsymbol{\beta}})$.

Proposition 2.4. *Assume the same notations as in Theorem 2.3 and $p'_\lambda(\cdot)$ satisfies condition $(C1)$. If: (a). $\hat{\boldsymbol{\beta}}$ satisfies $|\hat{\beta}_h| > |\hat{\beta}_i|$, for $\forall\ h, i$, where $1 \leqslant h \leqslant p-K$ and $p-K+1 \leqslant i \leqslant p$; and (b). conditions (2.9) - (2.11) are satisfied. Then $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q(\boldsymbol{\beta})$. On the other hand, assume: (c). $\hat{\boldsymbol{\beta}}$ satisfies $|\hat{\beta}_h| \geqslant |\hat{\beta}_i|$, for $\forall\ h, i$, where $1 \leqslant h \leqslant p-K$ and $p-K+1 \leqslant i \leqslant p$; and (d). $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\boldsymbol{\beta})$. Then $\hat{\boldsymbol{\beta}}$ must satisfy (2.9) - (2.11) with strict inequalities replaced by nonstrict inequalities.*

The proof of Proposition 2.4 is given in the Appendix for completeness.

It is well known that Karush-Kuhn-Tucker (KKT) conditions characterize the behavior of the minimizer of a convex function. When it comes to a nonconvex function, KKT conditions do not apply directly. Proposition 2.4 presented here plays the role as the KKT conditions to characterize the behavior of a local minimizer of the nonconvex KSI method. It is also worth noting that the necessary condition for a local minimizer and sufficient condition for a strict local minimizer differ slightly (nonstrict versus strict inequalities). The strict inequality is needed in order to ensure the optimizer is indeed a strict local minimizer.

Nonasymptotic Weak Oracle Properties

In this section, we study the nonasymptotic property of the KSI method. The weak oracle property was introduced by Lv and Fan (2009). It means that when $n \to \infty$, with probably tending to 1, the penalty will identify the sparsity structure of the true coefficients, and the L_∞ loss on the nonzero elements of the true coefficients is consistent under a rate slower than $\sqrt{n}$. Fan and Lv (2011) showed that there exists a nonconvex SCAD estimator which satisfies the weak oracle property when n is sufficiently large and $log(p) = O(n^{1-2t})$, for some $t \in (0, \frac{1}{2})$. We will extend their arguments and show that a similar KSI estimator exists for our proposed method.

It is worth noting that under the rate $log(p) = O(n^{1-2t})$, the number of predictors p is allowed to be much larger than the number of observations n, i.e., $p >> n$. This situation arises in many practical applications and is of great interests to many researchers.

Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be the true regression coefficients in model (2.1). Assume $\boldsymbol{\beta}^* = ({\boldsymbol{\beta}_1^*}^T, {\boldsymbol{\beta}_2^*}^T)^T$, where $\boldsymbol{\beta}_1^* \in \mathbb{R}^s$ are the nonzero components of $\boldsymbol{\beta}^*$, and $\boldsymbol{\beta}_2^* = \mathbf{0} \in \mathbb{R}^{p-s}$ are the zero components of $\boldsymbol{\beta}^*$. Denote X_1 and X_2 to be the submatrices of X formed by first s columns and last $p-s$ columns of X, respectively.

We have the following theorem:

Theorem 2.5. *(Weak Oracle Property) Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$. Under the conditions:*

- $(C2)$ *Let* $d_n = \frac{1}{2}\min\{|\beta_j^*| \Big| \beta_j^* \neq 0\} > n^{-\gamma}\log n$, *for some* $\gamma \in (0, \frac{1}{2})$;
- $(C3)$ $\|(X_2^T X_1)(X_1^T X_1)^{-1}\|_\infty \leqslant \min\{C\frac{p'_\lambda(0+)}{p'_\lambda(d_n)},\ O(n^{\frac{1}{2}-t})\}$, *for some* $C \in (0,1)$ *and* $t \in (\gamma, \frac{1}{2})$;
- $(C4)$ $\|(X_1^T X_1)^{-1}\|_\infty = o(n^\alpha\sqrt{\log n})$, *where* $\alpha < \min\{-\frac{1}{2} - \gamma,\ t - \gamma - 1\}$;
- $(C5)$ $\exists\ \lambda_n$, *such that* $\lambda_n p'_{\lambda_n}(d_n) = O(n^{-\alpha-\gamma-1}\sqrt{\log n})$ *and* $n^{-t}\sqrt{\log n}/\lambda_n = o(1)$;
- $(C6)$ $\lambda_n\kappa_0 = o(\tau)$, *where* $\kappa_0 = \max\{-p''_{\lambda_n}(\delta_i) \Big| \boldsymbol{\delta} \in \mathbb{R}^s, \|\boldsymbol{\delta} - \boldsymbol{\beta}_1^*\|_\infty \leqslant n^{-\gamma}\log n\}$, *and* $\tau = \lambda_{min}[\frac{1}{n}X_1^T X_1]$;
- $(C7)$ $|\beta^*_{[p-K_n]}| - |\beta^*_{[p-K_n+1]}| > 2n^{-\gamma}\log n$;
- $(C8)$ *parameter* K_n *satisfies* $K_n \geqslant p - s$;
- $(C9)$ X *is standardized such that the diagonal elements of* $X^T X/n$ *is* 1

Then for $s = o(n)$ *and* $\log(p) = O(n^{1-2t})$, *there exists a strict local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q(\boldsymbol{\beta})$ *such that for sufficiently large* n, *with probability at least* $1 - (sn^{-1} + (p-s)e^{-n^{1-2t}\log n})$, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *satisfies:*

1. *(Sparsity)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

2. *(L_∞ loss)* $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_\infty = O(n^{-\gamma} \log n)$,

where $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^s$ and $\hat{\boldsymbol{\beta}}_2 \in \mathbb{R}^{p-s}$ are the subvectors of $\hat{\boldsymbol{\beta}}$ formed by the first s covariates and the last $p-s$ covariates of $\hat{\boldsymbol{\beta}}$, respectively.

In Theorem 2.5, condition $(C2)$ requires that the minimum signal of true coefficients $\boldsymbol{\beta}^*$ can not be too small. Condition $(C3)$ is similar to the irrepresentative condition of Lasso (Zhao and Yu, 2006). It essentially requires the correlation between X_1 and X_2 can not be too strong. Condition $(C4)$ essentially requires that matrix $\frac{1}{n}X_1^T X_1$ is not singular and there exists a upper bound for the L_∞-norm of its inverse. Condition $(C5)$ controls the growth rate of tuning parameter λ_n. The existence of λ_n is valid because of $\alpha < t - \gamma - 1$. Condition $(C6)$ is similar to condition $C(4)$. It explicitly requires the smallest eigenvalue of $\frac{1}{n}X_1^T X_1$ do not vanish too fast. Condition $(C7)$ requires that there is a gap between the K regression coefficients that are penalized and the rest $p-K$ regression coefficients that are unpenalized. This is a price to pay when we want to maintain the order constraint in our estimate $\hat{\boldsymbol{\beta}}$. When $K = p$, no order constraint is needed and condition $(C7)$ can be removed. When $K = p - s$, condition $(C7)$ reduces to condition $(C2)$. Condition $(C8)$ requires the number of coefficients being penalized can not be too small. When $K < p - s$, there are exists some $\beta_j^* = 0$ that is unpenalized. Thus, the sparsity conclusion $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ will not be able to hold. Condition $(C9)$ assumes the matrix X has been standardized, which is a common procedure in practice.

The proof is given in the Appendix for completeness.

From Theorem 2.5, under the given regularity conditions, for sufficiently large n and $p = O(exp(n^{1-2t}))$, with large probability, there exists a KSI estimator which satisfies the weak oracle property. The diverging rate of p is allowed to grow up to exponentially fast with respect to n. The estimation consistency rate $O(n^{-\gamma} \log n)$ is slower than $O(\sqrt{n})$. This is in line with SCAD and LASSO penalty as shown in Fan and Lv (2011). Both of the diverging rate of p and the estimation consistency rate depend on the structure of design matrix and the minimum signal of the true coefficients. To obtain better estimation consistency rate, we need to make stronger assumption on the design matrix and minimum signals, as well as slower diverging rate of p. And vice versa.

Oracle Properties

In this section, we study the oracle property (Fan and Li, 2001) of the KSI method. In the previous section when we study the weak oracle properties, we made the assumption that the parameter K satisfies $K \geqslant p - s$. In this section we will show, if the parameter K is correctly chosen such that $K = p - s$, then for $\forall\, p$ and under much more relaxed assumptions than in previous section, there exists a strict local minimizer of the KSI method such that the oracle property is satisfied. It is worth noting that in this case, p is allow to diverge at a rate even faster than $O(exp(n))$. This is a much stronger conclusion than in any existing literature.

Assume the same notations as in the previous section, we have the following theorem:

Theorem 2.6. *Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$. Under the conditions:*

- $(C9)$ *X is standardized such that the diagonal elements of $X^T X/n$ is* 1
- $(C10)$ *$s \leqslant n$ and $\lambda_{min}[\frac{1}{n}X_1^T X_1] \geqslant c > 0$, where c is some positive constant;*
- $(C11)$ $\min\{|\beta_j^*|\Big|\beta_j^* \neq 0\} >> \sqrt{s/n}$;
- $(C12)$ *$K_n = p - s$ and $\lambda_n >> max\{\sqrt{\log(p)}, \sqrt{s/n^2}\max_{||\mathbf{v}||_2=1}||X_2' X_1 \mathbf{v}||_\infty\}$.*

Then for $\forall\, p$, there exists a strict local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ of $Q(\boldsymbol{\beta})$ such that $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$, and $||\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*||_2 = O_p(\sqrt{s/n})$. In fact, $\hat{\boldsymbol{\beta}}_1$ is given by $\hat{\boldsymbol{\beta}}_1 = (X_1' X_1)^{-1} X_1' \mathbf{y}$. Here $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^s$ and $\hat{\boldsymbol{\beta}}_2 \in \mathbb{R}^{p-s}$ are the subvectors of $\hat{\boldsymbol{\beta}}$ formed by the first s covariates and the last $p - s$ covariates of $\hat{\boldsymbol{\beta}}$, respectively.

The proof is given in the Appendix for completeness.

Here condition $(C10)$ requires that matrix $\frac{1}{n}X_1^T X_1$ is not singular and its smallest eigenvalue is bounded below by a positive constant. Condition $(C11)$ requires the smallest signal in $\boldsymbol{\beta}^*$ is much stronger than $\sqrt{s/n}$. This is a price to pay when we want to maintain the order constraint in our estimate $\hat{\boldsymbol{\beta}}$ and distinguish the important coefficients from the unimportant ones. Condition $(C12)$ assumes the parameter K satisfies $K = p - s$. Then with a large enough parameter λ, we will be able to remove the unimportant coefficients and recover the important coefficients. The estimation for the important coefficients work as well as if the correct submodel were known in advance and a ordinary least squares regression were fitting based on the submodel.

Theorem 2.7. *(Oracle Property) Under the conditions of Theorem* 2.6*, with probability tending to* 1*, the strict local minimizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *in Theorem* 2.6 *must satisfy:*

1. *(Sparsity)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$*;*
2. *(Asymptotic normality).*

$$A_n(X_1'X_1)^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \xrightarrow{\mathscr{D}} N(0, \sigma^2 G), \tag{2.13}$$

where A_n *is a* $q \times s$ *matrix such that* $A_n A_n' \to G$*,* G *is a* $q \times q$ *symmetric positive definite matrix.*

The proof is obvious by Theorem 2.6 and the normality assumption in $\boldsymbol{\epsilon}$, so we skip the proof here.

2.4 Simulation Studies

In this section, we perform simulation studies to compare the finite sample performance of LASSO, SCAD, adaptive LASSO, adaptive SCAD, with our KSI-LASSO and KSI-SCAD methods. We consider three examples based on the following linear regression model:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i, \; i = 1, \ldots, n,$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. We chose σ to control the signal-to-noise ratio to be 3. The details of the settings are described as follows.

Example 1: There are $p = 10$ variables and $n = 100$ observations in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{jk} = 0.5^{|j-k|}$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0]^T.$$

Example 2: There are $p = 50$ variables and $n = 100$ observations in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$. The $\mathbf{\Sigma}_{jk}$ was given by: $\mathbf{\Sigma}_{jj} = 1$; $\mathbf{\Sigma}_{jk} = 0.5$, for $1 \leqslant j, k \leqslant 5$ and $j \neq k$; $\mathbf{\Sigma}_{jk} = 0.2$ otherwise. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [1, 1, 1, 2, 2, \underbrace{0, \ldots, 0}_{45}]^T.$$

Example 3: There are $p = 100$ variables and $n = 60$ observations in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$. The $\mathbf{\Sigma}_{jk}$ was given by: $\mathbf{\Sigma}_{jj} = 1$; $\mathbf{\Sigma}_{jk} = 0.5$, for $1 \leqslant j, k \leqslant 5$ and $j \neq k$; $\mathbf{\Sigma}_{jk} = 0.2$ otherwise. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [1, 1, 1, 2, 2, \underbrace{0, \ldots, 0}_{95}]^T.$$

We repeated each simulation example for $1,000$ times. The KSI-LASSO and the KSI-SCAD were fitted using the algorithm described in Section 2.2. The LASSO and the adaptvie LASSO were fitted using the R package "glmnet" (Friedman et al., 2010b). The SCAD and the adaptive SCAD were fitted using the R package "ncvreg". The parameter "a" of both the SCAD and the adaptive SCAD penalties was fixed at 3.7 as suggested in Fan and Li (2001). We used the same $a = 3.7$ in the KSI-SCAD as well. 5-fold Cross Validation (CV) was used to select the best tuning parameters.

In example 1, tuning parameter K for the KSI methods was selected from set $\{ 9, 8, 7, 6, 5 \}$; the weights for the adaptive methods were obtained using the reciprocal of absolute values of ordinary least squares estiamtes. In example 2, K was selected from set $\{ 49, 47, 45, 43, 41 \}$; the weights for the adaptive methods were also obtained using the reciprocal of absolute values of ordinary least squares estiamtes. In example 3, K was selected from set $\{\,95, 90, 85, 80, 75, 70, 65, 60\,\}$. Note that in example 3, p is greater than n. The ordinary least squares estiamates do not exist. We obtained the weights for the adaptive methods using the ridge regression, with the optimum tuning parameter for the ridge regression selected from 5-fold corss validation. In all three examples, the tuning parameter λ was selected from set $\{exp(3 - 0.1 * i)\big|0 \leqslant i \leqslant 110\}$. Following our discussion in section 2.2, for each value of K, the initial warm starting point $\boldsymbol{\beta}^{warm}$ was calculated by SIS method.

To evaluate the variable selection performance of methods, we consider sensitivity (Sens) and specificity (Spec), which are defined as follows:

$$\begin{aligned}\text{Sens } &= \frac{\text{\# of selected important variables}}{\text{\# of important variables}}\\ \text{Spec } &= \frac{\text{\# of removed unimportant variables}}{\text{\# of unimportant variables}}.\end{aligned}$$

For both sensitivity and specificity, a higher value means a better variable selection performance. We also provide the number of nonzero estimated coefficients for readers' interests.

To evaluate the prediction performance of methods, following Tibshirani (1996), we consider the model error (ME) which is defined as:

$$\text{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector, $\boldsymbol{\beta}^*$ is the true coefficient vector, and $\boldsymbol{\Sigma}$ is the covariance matrix of the design matrix $\mathbf{X}$. We would like to acknowledge that the model error is closely related to the predictive mean square error proposed in Wahba (1985) and Leng et al. (2006).

To evaluate the estimation performance of methods, we consider the estimation bias which is given as:

$$\text{BIAS} = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2.$$

A smaller value means better estimation performance.

The simulation results are summarized in Table 2.1. We notice that in all three examples, the KSI methods produced higher specificity, smaller number of nonzero estimated coefficients, smaller model error and smaller bias, than their corresponding original methods. The sensitivity of KSI methods and their corresponding original methods are close to each other and comparable as well.

2.5 Real Data Analysis

In this section, we apply LASSO, SCAD, adaptive LASSO, adaptive SCAD, and KSI-LASSO and KSI-SCAD methods to analyze the well-known Boston housing data from the 1970 census (Harrison Jr and Rubinfeld, 1978). This dataset consists of 506 observations and 13 predictors, with median value of owner-occupied homes (MEDV) being the target variable. The purpose of Harrison Jr and Rubinfeld (1978) was to study the effect of air quality on housing price in the Boston area in the 1970's. Table 2.2 gives the names and meanings of the variables.

Harrison Jr and Rubinfeld (1978) fitted the following linear model:

$$\begin{aligned}\log(\mathrm{MEDV}) = \quad & \beta_0 + \beta_1\mathrm{AGE} + \beta_2\mathrm{B} + \beta_3\mathrm{CHAS} + \beta_4\mathrm{CRIM} + \beta_5\log(\mathrm{DIS}) \\ & +\beta_6\mathrm{INDUS} + \beta_7\log(\mathrm{LSTAT}) + \beta_8\mathrm{NOX}^2 + \beta_9\mathrm{PT} \\ & +\beta_{10}\log(\mathrm{RAD}) + \beta_{11}\mathrm{RM}^2 + \beta_{12}\mathrm{TAX} + \beta_{13}\mathrm{ZN},\end{aligned}$$

whose ordinary least squares estimate, t-statistic and p-value for each predictor are given in Table 2.3.

In order to compare the prediction performance and the variable selection performance of different penalized methods, we created 50 extra noise predictors z_i, $1 \leqslant i \leqslant 50$, and added them onto the original 13 predictors in the Boston housing dataset. These z_i's satisfy normal distribution with mean zero, unit variance and pairwise correlation $\rho = 0.2$. After this data augmentation step, we have obtained a new dataset with 506 subjects and 63 predictors. Among these 63 predictors, 50 are pure noise predictors.

We applied six methods to analyze the augmented Boston housing data: LASSO, SCAD, adaptive LASSO, adaptive SCAD, and our KSI-LASSO and KSI-SCAD methods. We randomly split the augmented dataset into a training set with sample size $n = 337$ and a test set with sample size $n = 169$ (the ratio of two sample sizes is about $2 : 1$). We fitted models on the training set using 5-fold CV. We then evaluated the prediction performances on the test set. We repeated the whole procedure beginning with a new random split for 100 times.

The tuning parameter K was selected from set { 58, 55, 53, 51, 50, 49, 47, 45, 42 }. The tuning parameter λ was selected from set $\{exp(4 - 0.1 * i) \big| 0 \leqslant i \leqslant 130\}$. The weights for the adaptive methods were obtained using the reciprocal of absolute values of ordinary least squares estiamtes. In order to be consistent with the simulation section, for each value of K, we obtained the initial warm starting point $\boldsymbol{\beta}^{warm}$ using the SIS method.

Table 2.4 summarizes the average mean square errors (MSE) on test sets over 100 replicates and the average number of selected variables over 100 replicates for LASSO, SCAD, adaptive LASSO, adaptive SCAD, and the KSI-LASSO and KSI-SCAD methods. We can see that both KSI-LASSO and KSI-SCAD methods produced the smallest MSEs (better prediction performance). We can also see that, on average, both KSI-LASSO and KSI-SCAD methods selected around 15 predictors, which are significantly less than LASSO, adaptive LASSO and SCAD. Because 50 predictors are pure noise predictors generated by ourselves and around 13 predictors are potentially related to the response, the KSI methods actually performed very well in terms of variable selection.

2.6 Extension and Discussion

In the previous sections, we have demonstrated the advantages of the KSI method in the case of variable selection problem. In fact, we could further extend the idea of KSI method to handle group variable selection problem. In this section, we briefly discuss the properties of the KSI methods when they are applied to the group variable selection problem.

To be specific, we consider the following linear regression setup with group structure: we have training data, $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$, where $\mathbf{x}_i$ and y_i are a p-length vector of covariates and response for the ith subject, respectively. We assume the total of p covariates can be divided into G groups. Let the gth group have p_g variables, and we use $\mathbf{x}_{i,(g)} = (x_{i,g1}, \ldots, x_{i,gp_g})^T$ to denote the p_g covariates in the gth group for the ith subject. To model the association between response and covariates, we consider linear regression:

$$y_i = \sum_{g=1}^{G} \sum_{j=1}^{p_g} x_{i,gj} \beta_{gj} + \epsilon_i,\ i = 1, \ldots, n, \tag{2.14}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_{gj}'s are regression coefficients. We denote $\boldsymbol{\beta}_g = (\beta_{g1}, \cdots, \beta_{gp_g})'$ to be the vector of regression coefficients for covariates in the gth group. We assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation.

For the purpose of variable selection, we consider the penalized ordinary least square (OLS) estimation:

$$\min_{\beta_{gj}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{g=1}^{G} \sum_{j=1}^{p_g} x_{i,gj} \beta_{gj} \Big)^2 + J_\lambda(\boldsymbol{\beta}), \tag{2.15}$$

where $J(\boldsymbol{\beta})$ is a sparsity-induced penalty function and λ is a non-negative tuning parameter.

Yuan and Lin (2006) proposed the following group LASSO penalty which is to penalize the L_2-norm of the coefficients within each group:

$$J_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} \sqrt{\beta_{g1}^2 + \cdots + \beta_{gp_g}^2} = \lambda \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_2. \tag{2.16}$$

Wang et al. (2007b) proposed the group SCAD penalty and Breheny and Huang (2009) proposed the group MCP penalty for group variable selection. They have the following form:

$$J_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} p_\lambda(||\boldsymbol{\beta}_g||_2), \tag{2.17}$$

where $p_\lambda(\cdot)$ is the SCAD penalty and the MCP penalty, respectively.

By applying the idea of KSI method, we can introduce more flexibility into the above group variable selection problem, reduce estimation bias and improve prediction performance, just as in the case of variable selection problem. In the remaining of this section, we consider the following penalized regression problem:

$$\min_{\beta_{gj}} Q_g(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{g=1}^{G}\sum_{j=1}^{p_g} x_{i,gj}\beta_{gj}\Big)^2 + \lambda \sum_{g=G-K+1}^{G} p_\lambda(||\boldsymbol{\beta}_{[g]}||_2). \tag{2.18}$$

Here λ is a nonnegative tuning parameter. K is a positive integer between 1 to G which controls the number of groups one wants to penalize. It is treated as another tuning parameter. $\boldsymbol{\beta}_{[g]}$'s are the group coefficients of $\boldsymbol{\beta}$ ordered by the magnitude of their L_2-norms, i.e.: $0 \leqslant ||\boldsymbol{\beta}_{[G]}||_2 \leqslant ||\boldsymbol{\beta}_{[G-1]}||_2 \leqslant \ldots \leqslant ||\boldsymbol{\beta}_{[1]}||_2$. And $p_\lambda(\cdot)$ is a sparsity-induced penalty function which may or may not depend on λ. We assume $p_\lambda(\cdot)$ satisfies the condition (C1) as in Section 2.

Algorithm

In this subsection, we present the algorithm for solving the optimization problem (2.18). This algorithm is similar to the one presented in Section 2.2, the details of which are as follows:

Step 1 (Initialization): Set iteration number $m = 0$, initialize $\boldsymbol{\beta}$ with some initial value $\boldsymbol{\beta}^{(0)}$.

Step 2 (Gradient Step): Let $\mathbf{u} = \boldsymbol{\beta}^{(m)} - s * \mathbf{X}'(\mathbf{X} * \boldsymbol{\beta}^{(m)} - \mathbf{Y})/n$. s here is some prespecified step length.

Step 3 (Sort $\mathbf{u}_g$): For $1 \leqslant g \leqslant G$, sort $\mathbf{u}_g$ according to their L_2-norms. Let idx_1 be the set of indices for the K-smallest elements of $\mathbf{u}_g$ in L_2-norm, and idx_2

be the rest.

Step 4 (Proximal Step and Update $\boldsymbol{\beta}^{(m+1)}$): For each index g in set *idx_1*, solve the following single variable optimization problem:

$$\boldsymbol{\beta}_g^{(m+1)} = \arg\min_{\mathbf{v}} \frac{1}{2}||\mathbf{v} - \mathbf{u}_g||_2^2 + s * \lambda * p_\lambda(||\mathbf{v}||_2) \tag{2.19}$$

For each index g in set *idx_2*, we simply set $\boldsymbol{\beta}_g^{(m+1)} = \mathbf{u}_g$.

Step 5 (Termination): Terminate the iteration if $||\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}||_2$ is small, or the changes of the objective function value $|Q_g(\boldsymbol{\beta}^{(m+1)}) - Q_g(\boldsymbol{\beta}^{(m)})|$ is small. Otherwise, set $m \leftarrow m+1$, and repeat from step 2.

Theoretical Results

In this subsection, we study the theoretical properties of the KSI estimators (2.18). Following the discussions in Section 3, we derive the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}$ to be a solution to minimization problem of $Q_g(\boldsymbol{\beta})$.

Theorem 2.8. *Without loss of generality, we assume $||\hat{\boldsymbol{\beta}}_g||_2 \neq 0$, for $1 \leqslant g \leqslant G_0$; and $||\hat{\boldsymbol{\beta}}_g||_2 = 0$, for $G_0 + 1 \leqslant g \leqslant G$. G_0 here is the number of selected important groups and we assume $G_0 \geqslant G - K$. Denote index set $H_0 = \{1, 2, \ldots, G_0\}$, and $H_0^c = \{G_0 + 1, \ldots, G\}$. Let X_g denotes the submatrice of X formed by the columns with group index being g. Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$.*

Assume $\hat{\boldsymbol{\beta}}$ satisfies:

$$||\hat{\boldsymbol{\beta}}_{g_1}||_2 > ||\hat{\boldsymbol{\beta}}_{g_2}||_2, \quad \forall g_1, g_2 \; s.t. \; 1 \leqslant g_1 \leqslant G-K, \; G-K+1 \leqslant g_2 \leqslant G_0; \tag{2.20}$$

then $\hat{\boldsymbol{\beta}}$ is a strict local minimizer to $Q_g(\boldsymbol{\beta})$ if the following conditions are satisfied:

$$\frac{1}{n} X_g^T (\mathbf{y} - X_{H_0} \hat{\boldsymbol{\beta}}_{H_0}) = \mathbf{0}, \quad 1 \leqslant g \leqslant G-K; \tag{2.21}$$

$$\frac{1}{n}X_g^T(\mathbf{y} - X_{H_0}\hat{\boldsymbol{\beta}}_{H_0}) - \lambda p'(||\hat{\boldsymbol{\beta}}_g||_2)\frac{\hat{\boldsymbol{\beta}}_g}{||\hat{\boldsymbol{\beta}}_g||_2} = \mathbf{0}, \quad G-K+1 \leqslant g \leqslant G_0; \tag{2.22}$$

$$\frac{1}{n}||X_g^T(\mathbf{y} - X_{H_0}\hat{\boldsymbol{\beta}}_{H_0})||_2 < \lambda p_\lambda'(0+), \quad G_0+1 \leqslant g \leqslant G; \tag{2.23}$$

$$\frac{1}{n}X_{H_0}^T X_{H_0} + \nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_{H_0}) > 0. \tag{2.24}$$

On the other hand, if $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q_g(\boldsymbol{\beta})$ and it satisfies (2.20) with strict inequality replaced by nonstrict inequality, then it must satisfy (2.21) - (2.24) with strict inequalities replaced by nonstrict inequalities.

Here $\nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_{H_0})$ is a block diagonal matrix such that $\nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_{H_0}) = \lambda * diag(A_1, \ldots, A_{G_0})$, where the $A_g's$ are given by:

$$A_g = \mathbf{O}, \quad 1 \leqslant g \leqslant G-K; \tag{2.25}$$

$$A_g = \frac{p_\lambda''(||\hat{\boldsymbol{\beta}}_g||_2)}{||\hat{\boldsymbol{\beta}}_g||_2^2}\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^T + \frac{p_\lambda'(||\hat{\boldsymbol{\beta}}_g||_2)}{||\hat{\boldsymbol{\beta}}_g||_2^3}\Big(||\hat{\boldsymbol{\beta}}_g||_2^2 * \mathbf{I} - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^T\Big), \quad G-K+1 \leqslant g \leqslant G_0. \tag{2.26}$$

The proof follows the same idea as in Theorem 2.3 and Proposition 2.4, so we skip the proof here.

Next, we study the nonasymptotic weak oracle property of the KSI estimators (2.18) under high dimensional setting. We will extend the arguments in Section 3 and show that a similar KSI estimator exists for (2.18).

Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be the true regression coefficients in model (2.14). Without loss of generality, we assume $||\boldsymbol{\beta}_g^*||_2 \neq 0$, for $1 \leqslant g \leqslant G_0$; and $||\boldsymbol{\beta}_g^*||_2 = 0$, for $G_0+1 \leqslant g \leqslant G$. G_0 here is the number of important groups. Denote index set $H_0 = \{1, 2, \ldots, G_0\}$, and $H_0^c = \{G_0+1, \ldots, G\}$. Denote $s_0 = \sum_{g=1}^{G_0} p_g$ to be the total number of covariates in the important groups.

We have the following theorem:

Theorem 2.9. *(Weak Oracle Property for Group Variable Selection) Assume $p_\lambda(\cdot)$ satisfies condition $(C1)$. Under the conditions:*

- $(C9)$ *X is standardized such that the diagonal elements of X^TX/n is* 1*;*

- $(C13)$ *Let $d_n = \frac{1}{2}\min_{1\leqslant g\leqslant G_0}\max_{1\leqslant j\leqslant p_g}|\beta^*_{gj}| > n^{-\gamma}\log n$, for some $\gamma \in (0, \frac{1}{2})$;*

- $(C14)$ $\max_{G_0+1\leqslant g\leqslant G}||(X_g^TX_{H_0})(X_{H_0}^TX_{H_0})^{-1}||_{\infty,2} = \min\{C\frac{p'_\lambda(0+)}{p'_\lambda(f_n)},\ O(n^{\frac{1}{2}-t})\}$*, for some $C \in (0,1)$ and $t \in (\gamma, \frac{1}{2})$, where $f_n = \min_{1\leqslant g\leqslant G_0}||\boldsymbol{\beta}^*_g||_2$, and the matrix norm is defined as $||A||_{\infty,2} = \max_{\mathbf{x}\neq\mathbf{0}}\frac{||A\mathbf{x}||_2}{||\mathbf{x}||_\infty}$;*

- $(C15)$ *$||(X_{H_0}^TX_{H_0})^{-1}||_\infty = o(n^\alpha\sqrt{\log n})$, where $\alpha < \min\{-\frac{1}{2}-\gamma,\ t-\gamma-1\}$;*

- $(C16)$ *$\exists\ \lambda_n$, such that $\lambda_n p'_{\lambda_n}(f_n) = O(n^{-\alpha-\gamma-1}\sqrt{\log n})$ and $n^{-t}\sqrt{L\log n}/\lambda_n = o(1)$, where $L = \max_{1\leqslant g\leqslant G}p_g$;*

- $(C17)$ *$\lambda_n\kappa_0 = o(\tau)$, where $\kappa_0 = \max\{-p''_{\lambda_n}(||\boldsymbol{\delta}_g||_2)\Big|||\boldsymbol{\delta}_g - \boldsymbol{\beta}^*_g||_\infty \leqslant n^{-\gamma}\log n,\ 1 \leqslant g \leqslant G_0\}$, and $\tau = \lambda_{min}[\frac{1}{n}X_{H_0}^TX_{H_0}] > 0$;*

- $(C18)$ *$||\beta^*_{[G-K_n]}||_2 - ||\beta^*_{[G-K_n+1]}||_2 > 2\sqrt{L}n^{-\gamma}\log n$;*

- $(C19)$ *parameter K_n satisfies $K_n \geqslant G - G_0$.*

Then for $s_0 = o(n)$ and $\log(p) = O(n^{1-2t})$, there exists a strict local minimizer $\hat{\boldsymbol{\beta}}$ of $Q_g(\boldsymbol{\beta})$ such that for sufficiently large n, with probability at least $1 - (s_0n^{-1} + (p - s_0)e^{-n^{1-2t}\log n})$, $\hat{\boldsymbol{\beta}}$ satisfies:

1. *(Sparsity) $\hat{\boldsymbol{\beta}}_g = \mathbf{0}, \forall\, g \geqslant G_0 + 1$;*

2. *(L_∞ Loss and Group Selection) $||\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}^*_g||_\infty = O(n^{-\gamma}\log n)$, and $\hat{\boldsymbol{\beta}}_g \neq \mathbf{0}, \forall\, g \leqslant G_0$.*

Theorem 2.9 is similar to Theorem 2.5 in many ways. It shows under the given regularity conditions, for sufficiently large n and $p = O(exp(n^{1-2t}))$, with large probability, there exists a KSI estimator which satisfies the theoretical property. This KSI estimator possesses group selection consistency, meaning that it will select all important groups (groups that contain at least one important variable) while removing all unimportant groups. It also has estimation consistency in the important groups in terms of L_∞ loss. Many of the regularity conditions are similar to the ones in Theorem 2.5. Condition $(C13)$ requires that minimum among the strongest signal $||\boldsymbol{\beta}^*_g||_\infty$ of important groups can not be too small. It is a relaxation over condition $(C2)$, because the smallest signal in $\boldsymbol{\beta}^*_g$ is allowed to be smaller than $n^{-\gamma} \log n$. In condition $(C14)$, a different matrix norm is used compared with condition $(C3)$. $p'_\lambda(f_n)$ in $(C14)$ is a small relaxation over $(C3)$ because the f_n is likely to be larger than the d_n in $(C3)$. Condition $(C15)$ is similar to condition $(C4)$. Condition $(C16)$ controls the growth rate of tuning parameter λ_n. It is a slightly stronger assumption than condition $(C5)$ because the maximum group size L in condition $(C16)$ can not grow too fast. However, we can always divide a large group into several small groups to meet this requirement. Condition $(C17)$ is similar to condition $(C6)$. Condition $(C18)$ requires that there is a gap between the K groups of regression coefficients that are penalized and the rest $G - K$ groups of regression coefficients that are unpenalized. The gap in $(C18)$ is larger than the gap in $(C7)$, so it is a stronger assumption than $(C7)$. Condition $(C19)$ requires the number of groups being penalized can not be too small.

The proof is given in the Appendix for completeness.

Simulation Studies

In this subsection, we perform simulation studies to compare the finite sample performance of group LASSO with our KSI-groupLASSO method. We consider the same setting as in Section 4 Example 1, but with group structure among the covariates of $\boldsymbol{\beta}^*$. To be specific, we have the following group structure:

$$\boldsymbol{\beta}^* = [\underbrace{3, 1.5, 0, 0, 2}_{group1}, \underbrace{0, 0, 0, 0, 0}_{group2}]^T. \tag{2.27}$$

We repeated this simulation example for $1,000$ times. The KSI-groupLASSO was fitted using the algorithm described in subsection 6.1. The group LASSO was fitted using the R package "grplasso". 5-fold Cross Validation (CV) was used to select the best tuning parameters. To evaluate the performance of methods, we consider sensitivity and specificity, model error and estimation bias as in Section 4.

To evaluate the group selection performance, we consider group sensitivity (gSens) and group specificity (gSpec), which are defined as follows:

$$\begin{aligned}\text{gSens} &= \frac{\text{\# of selected important groups}}{\text{\# of important groups}}\\ \text{gSpec} &= \frac{\text{\# of removed unimportant groups}}{\text{\# of unimportant groups}}.\end{aligned}$$

The simulation results are summarized in Table 2.5.

We notice that in this example, the KSI-groupLASSO method produced higher specificity, higher group specificity, smaller number of nonzero estimated coefficients, smaller model error and smaller bias, than the group LASSO method.

2.7 Conclusion

In this chapter, we proposed a novel K-Smallest Items (KSI) penalties family for variable selection. We developed an efficient gradient projection algorithm for solving the corresponding optimization problem. We also studied the nonasymptotic weak oracle property and the oracle property for our proposed method in the high-dimensional setting. Numerical results indicate that the proposed new method works well in terms of prediction accuracy, estimation bias reduction and variable selection. We applied the KSI method to analyze well know Boston housing dataset and showed its advantage. We further extended the KSI method to handle group variable selection problem and studied the group selection consistency and estimation consistency of the KSI estimators. We performed numerical experiments to study its performance under the group variable selection problem setting.

Table 2.1: Summary of simulation results over 1000 replicates. “Sens” means sensitivity of variable selection; “Spec” means specificity of variable selection; “Nonzero” means the number of nonzero estimated coefficients; “ME” means model error; “BIAS” means bias. The numbers in parentheses are the corresponding standard errors. “adpLASSO” means adaptive LASSO. “adpSCAD” means adaptive SCAD.

	5-fold CV Tuning				
	Sens	Spec	Nonzero	ME	BIAS
Example 1					
LASSO	1.000	0.601	5.795	0.518	0.683
	(0.000)	(0.008)	(0.058)	(0.011)	(0.017)
adpLASSO	0.999	0.798	4.411	0.432	0.596
	(0.001)	(0.007)	(0.048)	(0.010)	(0.017)
KSI-LASSO	0.998	0.858	3.992	0.369	0.520
	(0.001)	(0.006)	(0.044)	(0.010)	(0.016)
SCAD	0.996	0.810	4.320	0.397	0.570
	(0.001)	(0.007)	(0.051)	(0.013)	(0.022)
adpSCAD	0.997	0.842	4.099	0.456	0.656
	(0.001)	(0.006)	(0.045)	(0.013)	(0.022)
KSI-SCAD	0.998	0.874	3.873	0.372	0.533
	(0.001)	(0.006)	(0.039)	(0.011)	(0.018)
Example 2					
LASSO	0.981	0.874	10.568	1.360	1.483
	(0.002)	(0.003)	(0.153)	(0.023)	(0.027)
adpLASSO	0.886	0.791	13.819	2.568	3.566
	(0.004)	(0.005)	(0.222)	(0.052)	(0.069)
KSI-LASSO	0.971	0.963	6.510	0.891	1.345
	(0.003)	(0.002)	(0.107)	(0.026)	(0.037)
SCAD	0.857	0.894	9.075	2.028	3.309
	(0.005)	(0.002)	(0.107)	(0.031)	(0.050)
adpSCAD	0.813	0.847	10.945	3.356	4.810
	(0.005)	(0.004)	(0.199)	(0.084)	(0.106)
KSI-SCAD	0.967	0.972	6.088	0.872	1.330
	(0.003)	(0.002)	(0.069)	(0.026)	(0.039)
Example 3					
LASSO	0.917	0.916	12.579	2.858	2.888
	(0.004)	(0.002)	(0.237)	(0.054)	(0.062)
adpLASSO	0.921	0.810	22.611	5.728	7.074
	(0.004)	(0.005)	(0.455)	(0.141)	(0.168)
KSI-LASSO	0.960	0.975	7.176	1.780	2.518
	(0.003)	(0.002)	(0.208)	(0.065)	(0.083)
SCAD	0.741	0.947	8.730	4.374	6.008
	(0.006)	(0.001)	(0.124)	(0.065)	(0.107)
adpSCAD	0.822	0.910	12.696	5.867	7.735
	(0.005)	(0.003)	(0.266)	(0.166)	(0.196)
KSI-SCAD	0.961	0.982	6.494	1.702	2.422
	(0.003)	(0.001)	(0.092)	(0.060)	(0.077)

Table 2.2: Variables in Boston Housing Data

Name	Meaning	Name	Meaning
MEDV	median value of homes	LSTAT	% lower status population
AGE	% homes built before 1940	NOX	nitric oxides concentration
B	$(\%Black - 63)^2/10$	PT	pupil teacher ratio
CHAS	1 if on Charles River; 0 o.w.	RAD	accessibility to radial highways
CRIM	per capita crime rate	TAX	full-value property-tax rate
DIS	distances to employment centers	RM	average number of rooms
INDUS	% non-retail business	ZN	% residential land

Table 2.3: Ordinary least squares estimates of the regression coefficients in the Boston Housing Data; t is the $t-$statistic of the corresponding coefficient.

Predictor	$\hat{\beta}$	t	$p-$value	Predictor	$\hat{\beta}$	t	$p-$value
Constant	4.47	28.05	$< 2e-16$	LSTAT	-3.71e-1	-14.84	$< 2e-16$
AGE	9.07e-5	0.17	8.6e-1	NOX^2	-6.38e-1	-5.64	2.9e-8
B	3.64e-4	3.53	4.6e-4	PT	-3.11e-2	-6.21	1.1e-9
CHAS	9.14e-2	2.75	6.1e-3	log(RAD)	9.57e-2	5.00	7.9e-7
CRIM	-1.19e-2	-9.53	$< 2e-16$	RM^2	6.33e-3	4.82	1.9e-6
log(DIS)	-1.91e-1	-5.73	1.8e-8	TAX	-4.20e-4	-3.43	6.6e-4
INDUS	2.39e-4	0.10	9.2e-1	ZN	8.02e-5	0.16	8.7e-1

Table 2.4: Summary of augmented Boston housing data analysis results. Results are based on 100 random splits using 5−fold CV tuning. “MSE” reports the average mean square errors on test sets and "Selection Frequency” reports the average number of selected variables. The numbers in parentheses are the corresponding standard errors. ”adpLASSO” means adaptive LASSO. ”adpSCAD” means adaptive SCAD.

Mean Square Error and Selection Frequency		
	MSE	Selection Frequency
LASSO	0.0389	28.460
	(0.0006)	(0.717)
adpLASSO	0.0392	19.560
	(0.0006)	(0.600)
KSI-LASSO	0.0366	15.180
	(0.0005)	(0.335)
SCAD	0.0382	19.219
	(0.0005)	(0.487)
adpSCAD	0.0396	14.590
	(0.0007)	(0.343)
KSI-SCAD	0.0366	14.850
	(0.0005)	(0.311)

Table 2.5: Summary of simulation results over 1000 replicates. “Sens” means sensitivity of variable selection; “Spec” means specificity of variable selection; “gSens” means sensitivity of group selection; “gSpec” means specificity of group selection; “Nonzero” means the number of nonzero estimated coefficients; “ME” means model error; “BIAS” means bias. The numbers in parentheses are the corresponding standard errors.

	5-fold CV Tuning						
	Sens	Spec	gSens	gSpec	Nonzero	ME	BIAS
Example 1 with group structure							
group LASSO	1.000	0.138	1.000	0.193	9.035	0.545	0.764
	(0.000)	(0.009)	(0.000)	(0.012)	(0.062)	(0.010)	(0.016)
KSI-groupLASSO	1.000	0.556	1.000	0.779	6.105	0.431	0.657
	(0.000)	(0.009)	(0.000)	(0.013)	(0.066)	(0.009)	(0.016)

3 VARIABLE SELECTION VIA THE SELF-ADAPTIVE PENALTY

3.1 Motivation

In this chapter, we consider the variable selection problem under the usual regression setting: we have training data, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i$ is a vector of values with p covariates and y_i is the response. To model the association between response and covariates, we consider the following linear regression:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,\ i = 1, \ldots, n, \tag{3.1}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_j's are regression coefficients. Without loss of generality, we assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation.

For the purpose of variable selection, we consider the penalized ordinary least squares (OLS) estimation:

$$\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 + \lambda J(\boldsymbol{\beta}), \tag{3.2}$$

where $J(\boldsymbol{\beta})$ is a certain sparsity-induced penalty function and λ is a non-negative tuning parameter.

In this chapter, we propose a novel Self-adaptive penalty (SAP) for variable selection. It is a nonconvex penalty and it is able to reduce estimation bias and improve the performance of variable selection and prediction. SAP is motivated by the adaptive LASSO, which is a weighted version of LASSO. The penalization of SAP on each individual coefficient takes into account directly the influence of other estimated coefficients. We transform the SAP model into a weighted LASSO problem and apply the coordinate descent algorithm (Friedman et al., 2007; Wu and Lange, 2008) to fit the model. We thoroughly investigate the theoretical properties of

SAP and show that it possesses the weak oracle property under desirable conditions. The proposed method is applied to the glioblastoma cancer data obtained from The Cancer Genome Atlas (TCGA).

This chapter is organized as follows. In Section 3.2, we propose the Self-adaptive penalty and present the corresponding algorithm. In Section 3.3, we derive its weak oracle property in the high-dimensional setting where the number of coefficients is allowed to be much larger than the sample size. Simulation results are presented in Section 3.4. In Section 3.5, we apply the proposed method to the TCGA glioblastoma dataset. Finally we conclude this chapter with Section 3.6.

3.2 Method

Self-adaptive penalty

We propose the following Self-adaptive penalty (SAP):

$$\lambda J(\boldsymbol{\beta}) \triangleq \lambda \log_a \left(a^{|\beta_1|} + \cdots + a^{|\beta_p|} \right), \tag{3.3}$$

where $a \in (0, 1)$ is another tuning parameter. The SAP is nonconvex, which can be straightforwardly verified by calculating its second derivative.

Consider an arbitrary vector $\boldsymbol{\beta}^0$ which is close to $\boldsymbol{\beta}$, the penalty can be locally approximated using the first order approximation:

$$\log_a(a^{|\beta_1|} + \cdots + a^{|\beta_p|}) \approx \log_a(a^{|\beta_1^0|} + \cdots + a^{|\beta_p^0|}) + \sum_{j=1}^{p} \frac{a^{|\beta_j^0|}}{\sum_{l=1}^{p} a^{|\beta_l^0|}} (|\beta_j| - |\beta_j^0|). \tag{3.4}$$

Treating the term $\log_a(\sum_j a^{|\beta_j^0|})$ and the term $\sum_{j=1}^{p} \frac{a^{|\beta_j^0|}}{\sum_{l=1}^{p} a^{|\beta_l^0|}} |\beta_j^0|$ as constants because they are not related to the $\boldsymbol{\beta}$, the original optimization problem (3.2) with

SAP can be approximated by:

$$\min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \Big)^2 + \lambda \sum_{j=1}^{p} \frac{a^{|\beta_j^0|}}{\sum_{l=1}^{p} a^{|\beta_l^0|}} |\beta_j|. \tag{3.5}$$

The original problem now becomes a weighted LASSO problem with the weight for each coefficient given by $a^{|\beta_j^0|} / \sum_{l=1}^{p} a^{|\beta_l^0|}$. These weights are automatically specified by the penalty itself and they takes into account the relations between different regression coefficients. By the assumption that $\boldsymbol{\beta}^0$ is close to $\boldsymbol{\beta}$, the weights are large for small coefficients of $\boldsymbol{\beta}$, and the weights are small for large coefficients. This implies SAP automatically penalizes more heavily on small coefficients and gently on large coefficients. Therefore it can reduce the estimation bias on large estimated coefficients.

The SAP method has several connections with the existing methods.

Similar to adaptive LASSO, SAP needs to solve the weighted LASSO problem. However, adaptive LASSO uses a two stages procedure. First a set of appropriate weights need to be specified. These weights are usually obtained by solving the ordinary least squares problem and are based on the individual signal strength of different predictors. Next these weights are plugged into the model and the adaptive LASSO estimators are achieved by solving the weighted LASSO problem. This two stages procedure requires extra computation efforts. Unlike adaptive LASSO, SAP combines these two stages together. The weights are automatically specified by the penalty itself and iteratively updated until model (3.2) is solved. These weights take into account the relation among different coefficients as is implied by the denominators in (3.5).

The format of SAP is similar to the group variable selection method Log-Exp-Sum (LES) penalty which will be presented in Chapter 4. Because SAP is not aimed for group variable selection and no group structure is assumed, the predictors in SAP are treated as if they are all in the same group. This is a special case of the LES penalty when only one single group exists. The natural base "*e*" in the LES penalty is changed to "*a*" in the SAP here, which makes SAP a nonconvex penalty.

It is worth noting that not every group variable selection method can be adjusted and applied to the variable selection problem. We consider two popular group variable selection methods here: the group LASSO and the group bridge. Using the idea of first order approximation, we notice that the weights are proportional to the sizes of the absolute value of coefficient estimators in group LASSO, and the weights are the same for every coefficient estimator in group bridge. Therefore, they are not good candidates for the variable selection problem.

Algorithm

We need to solve the following optimization problem:

$$\min_{\beta_j} Q(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda \log_a \Big(a^{|\beta_1|} + \cdots + a^{|\beta_p|}\Big). \tag{3.6}$$

The SAP is singular at the origin, and it does not have continuous second derivatives. So Newton method and many other sophisticated optimization techniques can not be applied directly. However, the optimization problem can be approximated by a weighted LASSO problem as described in (3.5). Following Friedman et al. (2007) and Wu and Lange (2008), we apply the coordinate descent algorithm to solve the optimization problem (3.5). Numerical examples in section 3.4 demonstrate the efficiency and efficacy of this algorithm. The details of the algorithm are as follows:

Step 1 (Initialization): Set $k = 0$, initialize $\boldsymbol{\beta}^{(0)}$ with some reasonable values.

Step 2 (Update $\boldsymbol{\beta}$): For $j = 1, \ldots, p$, obtain the update for the j-th covariate of $\boldsymbol{\beta}^{(k+1)}$:

$$\beta_j^{(k+1)} = \mathsf{S}\Big(\frac{\sum_{i=1}^{n} x_{ij}\tilde{y}_i}{\sum_{i=1}^{n} x_{ij}^2}\,,\ \frac{n\lambda w_j}{\sum_{i=1}^{n} x_{ij}^2}\Big), \tag{3.7}$$

where $\tilde{y}_i = y_i - \sum_{i \neq j} x_{il}\beta_l^{(k)}$, $w_j = \frac{a^{|\beta_j^{(k)}|}}{\sum_{l=1}^{p} a^{|\beta_l^{(k)}|}}$, and $\mathsf{S}(\cdot,\cdot)$ is a soft threshold function given by $\mathsf{S}(u, v) = sign(u)(|u| - v)_+$.

Step 3 (Termination) Terminate the iteration if $||\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}||$ is small, or the changes of the objective function value $|Q(\boldsymbol{\beta}^{k+1}) - Q(\boldsymbol{\beta}^{k})|$ is small. Otherwise, set $k \leftarrow k+1$, and repeat step 2.

3.3 Theoretical Results

In this section, we study the theoretical properties of the SAP estimators. It is generally difficult to study the global minimizer of a nonconvex function, so as is common in literature, we focus on local minimizer. In the remaining of this section, we first present several conditions which characterize the behavior of a local minimizer. We then show under several assumptions, there exists a local minimizer of SAP such that the nonasymptotic weak oracle properties are satisfied. Throughout this section, we denote the penalized likelihood optimization problem to be:

$$\min_{\beta_i} Q(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\sum_{k=1}^{n}\Big(y_k - \sum_{i=1}^{p} x_{ki}\beta_i\Big)^2 + \lambda J(\boldsymbol{\beta}), \tag{3.8}$$

where $J(\boldsymbol{\beta}) = \log_a\Big(a^{|\beta_1|} + \cdots + a^{|\beta_p|}\Big)$.

Characterization of SAP estimators

In spirit of Theorem 1 in Fan and Lv (2011), we derive the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}$ to be the solution to (3.8).

Theorem 3.1. *A vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a strict local minimizer to $Q(\boldsymbol{\beta})$ if the following conditions are satisfied:*

$$\frac{1}{n}X_1^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1) - \lambda\nabla_1 J(\hat{\boldsymbol{\beta}}_1) = 0; \tag{3.9}$$

$$\frac{1}{n}||X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)||_\infty < \frac{\lambda}{\sum_{i\leqslant s} a^{|\hat{\beta}_i|} + (p-s)}; \tag{3.10}$$

$$\frac{1}{n}X_1^T X_1 + \lambda\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1) > 0. \tag{3.11}$$

On the other hand, if $\hat{\boldsymbol{\beta}}$ is a solution to (3.8), then it is a local minimizer if it satisfies (3.9) - (3.11) with strict inequalities replaced by nonstrict inequalities.

Here we assume $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$, where $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^s$ are the nonzero components of $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}}_2 = \mathbf{0} \in \mathbb{R}^{p-s}$ are the zero components of $\hat{\boldsymbol{\beta}}$. s is the number of nonzero components of $\hat{\boldsymbol{\beta}}$. X_1 and X_2 denote the submatrices of X formed by first s columns and last $p-s$ columns of X, respectively. Denote $T = \sum_{i \leqslant s} a^{|\hat{\beta}_i|} + (p-s)$, $\nabla_1 J(\hat{\boldsymbol{\beta}}_1)$ and $\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1)$ are given by:

$$\nabla_1 J(\hat{\boldsymbol{\beta}}_1) = \begin{bmatrix} \frac{a^{|\hat{\beta}_1|}}{T} sign(\hat{\beta}_1) \\ \vdots \\ \frac{a^{|\hat{\beta}_s|}}{T} sign(\hat{\beta}_s) \end{bmatrix},$$

$$\begin{aligned} & \nabla_1^2 J(\hat{\boldsymbol{\beta}}_1) \\ = \; & -\frac{\log a}{T^2} \begin{bmatrix} -a^{|\hat{\beta}_1|}(T - a^{|\hat{\beta}_1|}) & a^{|\hat{\beta}_1|+|\hat{\beta}_2|} sign(\hat{\beta}_1 \hat{\beta}_2) & \cdots & a^{|\hat{\beta}_1|+|\hat{\beta}_s|} sign(\hat{\beta}_1 \hat{\beta}_s) \\ a^{|\hat{\beta}_1|+|\hat{\beta}_2|} sign(\hat{\beta}_1 \hat{\beta}_2) & -a^{|\hat{\beta}_2|}(T - a^{|\hat{\beta}_2|}) & \cdots & a^{|\hat{\beta}_2|+|\hat{\beta}_s|} sign(\hat{\beta}_2 \hat{\beta}_s) \\ \vdots & \vdots & \ddots & \vdots \\ a^{|\hat{\beta}_1|+|\hat{\beta}_s|} sign(\hat{\beta}_1 \hat{\beta}_s) & a^{|\hat{\beta}_2|+|\hat{\beta}_s|} sign(\hat{\beta}_2 \hat{\beta}_s) & \cdots & -a^{|\hat{\beta}_s|}(T - a^{|\hat{\beta}_s|}) \end{bmatrix}. \end{aligned}$$

The proofs are given in the Appendix for completeness.

It is well known that Karush-Kuhn-Tucker (KKT) conditions characterize the behavior of the minimizer of a convex function. When it comes to nonconvex function, KKT conditions do not apply directly. The theorem we present here plays the role as KKT conditions to characterize the behavior of a local minimizer of the nonconvex SAP method. It is also worth noting that the necessary condition for a local minimizer and sufficient condition for a strict local minimizer differ slightly (nonstrict versus strict inequalities). The strict inequalities are needed in order to ensure the optimizer is indeed a strict local minimizer.

We denote $\lambda_{min}(G)$ to be the minimum eigenvalue of G, where G is any square matrix. The following proposition gives an easy way to check condition (3.11).

Proposition 3.2. *Let s be the number of nonzero components in $\hat{\boldsymbol{\beta}}$. A sufficient condition* *for inequality (3.11) to hold is:*

$$\lambda_{min}(\frac{1}{n}X_1^T X_1) \geqslant -\frac{\lambda}{p-s}\log a. \tag{3.12}$$

Nonasymptotic Weak Oracle Properties

In this section, we study the nonasymptotic property of the SAP estimator. The weak oracle property was introduced by Lv and Fan (2009). It means that when $n \to \infty$, with probably tending to 1, the penalty will identify the sparsity structure of true coefficients, and the L_∞ loss on the nonzero elements of true coefficients is consistent under a rate slower than $\sqrt{n}$. Fan and Lv (2011) showed that there exists a nonconvex SCAD estimator which satisfies the weak oracle property when n is sufficiently large and $log(p) = O(n^{1-2\alpha})$, for some $\alpha \in (0, \frac{1}{2})$. We will extend their arguments and show that a similar SAP estimator exists for our proposed method. It is worth noting that under the rate $log(p) = O(n^{1-2\alpha})$, the number of predictors p is allowed to be much larger than the number of observations n, i.e., $p >> n$. This situation arises in many practical applications and is of great interests to many researchers.

Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be the true coefficients. Assume $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \boldsymbol{\beta}_2^{*T})^T$, where $\boldsymbol{\beta}_1^* \in \mathbb{R}^s$ are the nonzero components of $\boldsymbol{\beta}^*$, and $\boldsymbol{\beta}_2^* = \mathbf{0} \in \mathbb{R}^{p-s}$ are the zero components of $\boldsymbol{\beta}^*$. Denote X_1 and X_2 to be the submatrices of X formed by first s columns and last $p-s$ columns, respectively.

Our main results are stated in the following theorem.

Theorem 3.3. *(Weak Oracle Property) Under the conditions:*

- *(C1)* $||(X_1^T X_1)^{-1}||_\infty = o(n^{-\frac{1}{2}-\alpha}\sqrt{\log n})$*, for some* $\alpha \in (0, \frac{1}{2})$*;*
- *(C2) Let* $d_n = \frac{1}{2}\min\{|\beta_j^*| : |\beta_j^*| \neq 0\} \geqslant n^{-\gamma}\log n$*, for some* $\gamma \in (0, \alpha + t - \frac{1}{2})$*,* $t \in (0, \frac{1}{2})$*;*
- *(C3)* α *and* t *satisfy* $\frac{1}{2} < \alpha + t < 1$*;*
- *(C4)* $||(X_2^T X_1)(X_1^T X_1)^{-1}||_\infty \leqslant \min\{Ca^{-d_n},\ O(n^{\frac{1}{2}-t})\}$*, for some* $C \in (0,1)$*;*

- *(C5)* $s = o(n)$ *and* $\log(p) = O(n^{1-2t})$*;*
- *(C6)* $\lambda_{min}(\frac{1}{n}X_1^T X_1) = O(n^r)$*, for some* $r \in (\alpha - \frac{1}{2} - \gamma, 0)$*;*
- *(C7)* X *is standardized such that the diagonal elements of* $X^T X/n$ *is* 1*.*

If we choose tuning parameter λ *such that* $\lambda = O(n^{\alpha-\frac{1}{2}-\gamma} p\sqrt{\log n})$ *and* $\lambda/(\frac{p\sqrt{\log n}}{n^t}) \to \infty$*. Then there exists an solution* $\hat{\boldsymbol{\beta}}$ *to* $Q(\boldsymbol{\beta})$ *such that for sufficiently large* n*, with probability at least* $1 - (sn^{-1} + (p-s)e^{-n^{1-2t}\log n})$*,* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *satisfies:*

1. *(Sparsity)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$*,*
2. *(L_∞ loss)* $||\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*||_\infty = O(n^{-\gamma}\log n)$*,*

where $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^s$ *and* $\hat{\boldsymbol{\beta}}_2 \in \mathbb{R}^{p-s}$ *are the subvectors of* $\hat{\boldsymbol{\beta}}$ *formed by the first* s *covariates and the last* $p-s$ *covariates of* $\hat{\boldsymbol{\beta}}$*, respectively.*

In Theorem 3.3, condition $(C1)$ essentially requires that matrix $\frac{1}{n}X_1^T X_1$ is not singular and there exists a upper bound for the L_∞-norm of its inverse. Condition $(C2)$ requires that the minimum signal of true coefficients $\boldsymbol{\beta}^*$ can not be too small. Condition $(C3)$ controls the relation between several constants, it is needed for technical proof purposes. Condition $(C4)$ is similar to the irrepresentative condition of Lasso (Zhao and Yu, 2006). It essentially requires the correlation between X_1 and X_2 can not be too strong. Condition $(C5)$ controls the growth rate of the number of important variables and the total number of variables. Condition $(C6)$ explicitly bounds the smallest eigenvalue of $\frac{1}{n}X_1^T X_1$. Condition $(C7)$ assumes the matrix X has been standardized, which is a common procedure in practice.

The proofs are given in Appendix for completeness.

From this theorem, under the given conditions, for sufficiently large n and $p = O(exp(n^{1-2t}))$, with probability tending to 1, there exists a SAP estimator which satisfies the weak oracle property. The diverging rate of p is allowed to grow up to exponentially fast with respect to n. The estimation consistency rate $O(n^{-\gamma}\log n)$ is slower than $O(\sqrt{n})$. This is in line with SCAD penalty as shown in Fan and Lv (2011). Both of the diverging rate of p and the estimation consistency rate depend on the structure of design matrix and the minimum signal of true coefficients. To obtain better estimation consistency rate, we need to make stronger assumption on the design matrix and minimum signals, as well as slower diverging rate of p. And vice versa.

3.4 Simulation Studies

In this section, we conduct simulation studies to evaluate the numerical performance of the SAP method. We compare our penalty with several existing methods, including LASSO, adaptive LASSO, SCAD and MCP. We consider four examples. All examples are based on the following linear regression model:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. For each example, we chose σ to have the signal-to-noise ratio to be 3. The details of each setting are described as follows.

Example 1. There are $n = 60$ observations and $p = 8$ variables in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}_{ii} = 1$. For $i \neq j$, the entry of $\boldsymbol{\Sigma}$ was given by:

$$\Sigma_{ij} = \begin{cases} 0.7 & 1 \leqslant i, j \leqslant 3 \text{ or } 4 \leqslant i, j \leqslant 8 \\ 0.3 & \text{o.w.} \end{cases} \tag{3.13}$$

The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [3, 2, 1.5, 0, 0, 0, 0, 0]^T. \tag{3.14}$$

Example 2. There are $n = 100$ observations and $p = 60$ variables in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}_{ii} = 1$. For $i \neq j$, the entry of $\boldsymbol{\Sigma}$ was given by: $\Sigma_{ij} = 0.5^{|i-j|}$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [0.05, 0.05, 0.05, 0.05, 0.05, \underbrace{0, 0, \ldots, 0, 0}_{55}]^T. \tag{3.15}$$

Example 3. There are $n = 400$ observations and $p = 200$ variables in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}_{ii} = 1$. For $i \neq j$, the entry of $\boldsymbol{\Sigma}$ was given by: $\Sigma_{ij} = 0.5$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [0.5, 1, 1.5, 2, 2.5, 3, -1, -1.5, \underbrace{0, 0, \ldots, 0, 0}_{192}]^T. \tag{3.16}$$

Example 4. There are $n = 60$ observations and $p = 100$ variables in total. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}_{ii} = 1$. For $i \neq j$, the entry of $\boldsymbol{\Sigma}$ was given by: $\Sigma_{ij} = 0.2$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [0.5, 1, 1.5, -0.5, \underbrace{0, 0, \ldots, 0, 0}_{96}]^T. \tag{3.17}$$

We repeated each simulation example for 1,000 times. The SAP was fitted using the algorithm described in Section 3.2. The LASSO was fitted using the R package "glmnet" (Friedman et al., 2010b). The adaptive LASSO was also fitted using the R package "glmnet" with the adaptive weights being the reciprocal of the absolute value of OLS estimators. In cases when $p > n$, the generalized OLS was used. The SCAD and MCP were fitted using the R package "ncvreg". Five-fold cross validation method was used to select the tuning parameters. For the SAP, we used two-dimensional cross-validation and selected the base a from $\{0.01, 0.1, 0.5, e^{-1}, 0.9, 0.99\}$. The parameter "$a$" of the SCAD penalty was fixed at 3.7 as Fan and Li (2001) suggested. The parameter "γ" of the MCP penalty was fixed at 2. The model with the smallest average cross validation error was selected.

To evaluate the variable selection performance of these methods, we considers sensitivity (Sens) and specificity (Spec), which are defined as follows:

$$\text{Sens} = \frac{\#\text{ of selected important variables}}{\#\text{ of important variables}};$$

$$\text{Spec } = \frac{\text{\# of removed unimportant variables}}{\text{\# of unimportant variables}}.$$

For both sensitivity and specificity, a higher value means a better variable selection performance.

To evaluate the prediction performance of these methods, in each simulation, we independently generated a test set data with $5,000$ observations from the same distribution as the training data. We then calculated the mean square error (MSE) on the test set. The MSE is the prediction error, so a smaller value means a better prediction performance.

The simulation results are summarized in Table 3.1.

In Example 1, the LASSO method produces the highest sensitivity and the lowest MSE. This is consistent with Fan and Li (2001) that LASSO has good performance when the noise level is high and the sample size is small. The SAP method has the second lowest MSE, which shows its good prediction performance. Although worse than LASSO and adaptive LASSO, SAP outperforms other non-convex methods in terms of sensitivity. SAP also outperforms other methods except MCP in terms of specificity. The fact that MCP produces the highest specificity is not surprising because Zhang (2010) has shown that MCP has a nice upper bound for the false positive.

In Example 2, LASSO and SAP have similar performance in all three aspects. The difference between these two methods is insignificant. They both produce the highest sensitivity and the lowest MSE which are significantly better than other three methods. Again, MCP produces the highest specificity, while SAP produces the second highest specificity.

In Example 3, the number of observations and the number of noise variables increase dramatically from example 1 and 2. SCAD produces the lowest MSE which is significantly better than other methods. SAP yields the second lowest MSE which differs insignificantly with MCP. LASSO produces the highest sensitivity. MCP again produces the highest specificity and SAP has the second highest specificity.

In Example 4, the number of variables is larger than the number of observations. In this situation, SCAD produces the lowest MSE. SAP gives the third lowest MSE, higher than SCAD and MCP, but lower than LASSO and adaptive LASSO. LASSO gives the highest sensitivity. MCP produces the highest specificity while SAP yields the second highest specificity.

3.5 Real Data Analysis

In this section, we analyzed the publicly available glioblastoma cancer data obtained from The Cancer Genome Atlas (TCGA). The glioblastoma cancer is the the most common and lethal brain cancer in adults (Furnari et al., 2007). It is characterized by its poor responses to all existing therapeutic approaches and the short median survival time for newly diagnosed patients (Furnari et al., 2007). Identification of the relation between survival time and cancer-causing genome will add to the understanding of this deadly disease.

This data set is obtained under the "hg_u133a" platform and it includes 12042 genes taking continuous values. The outcome of interest is the survival time of glioblastoma patients. The data set contains a total of 519 subjects among which 116 subjects (22.4%) are censored in response. We excluded these 116 censored subjects and retained $n = 403$ subjects for our analysis. We took the logarithm transform of the survival time, centered the data and standardized the genes before any analysis was made.

In our analysis, we first applied the Sure Independence Screening (SIS) method proposed by Fan and Lv (2008b) to reduce the total number of predictors down to $n/log(n) \approx 67$, as is suggested in Fan and Lv (2008b). The SIS is a screening method which is widely used for dimension reduction. Combining the screening method SIS with variable selection methods will allow us to narrow down the search for important variables and improve the estimation accuracy (Fan and Lv, 2008b). After the SIS step, we are presented with a subset of the original data which contains $n = 403$ subjects and $p = 67$ predictors.

In this subset, we compared the prediction performance of our proposed SAP method with LASSO, adaptive LASSO, SCAD and MCP. We randomly split the whole subset into a training set with sample size $n_{train} = 300$ and a test set with sample size $n_{test} = 103$ (the ratio of two sample sizes is about $3 : 1$). We fitted models on the training set and evaluated their prediction performances on the test set. We used 5-fold cross validation as our tuning criterion for training. We repeated the whole procedure starting from the random split for 100 times.

Table 3.2 summarizes the average mean square errors (MSE) on test set and the average number of variable selected across 100 random splits. We can see that, among all five methods, LASSO and SAP produced the same MSE up to the third digit, which is the smallest among all five methods. As the cost of this gain in prediction performance, LASSO and SAP selected more variables, on average, than other methods.

Table 3.3 summarizes the selection frequency of each individual variables across 100 random splits. It is worth noting that the selection frequency of SAP is very similar to LASSO. This is because the estimated coefficients in absolute value are small ($\max\hat{\beta}_j < 0.274$). The weights for each predictor in SAP are close to each other which makes the SAP problem becomes similar to the LASSO problem. Among the 67 genes, 11 of them have been selected for more than 90 times by SAP, including: RPS28, TOP1, LAMB4, HEMK1, ZNF208, NPTXR, CLEC5A, CDK3, NELL1, GSN and HOMER1. Among these genes, CDK3 and GSN are found to be closely associated with glioblastoma in literature (Zheng et al., 2008; Kamitani et al., 2002). The connection of other highly selected genes with glioblastoma remains to be found. SAP has narrowed down the collection of genes for biologist to examine.

3.6 Conclusion

In this chapter, we proposed a new nonconvex Self-adaptive penalty (SAP) method for variable selection. We developed an efficient coordinate descent algorithm for solving the corresponding optimization problem. We also studied the nonasymptotic weak oracle property for our proposed method, in which the number of prediction variables is allowed to be much larger than the sample size. Numerical results indicate that the proposed new method works well in terms of both prediction accuracy and variable selection. We applied the SAP method to analyze a TCGA glioblastoma cancer data and obtained clinically meaningful results.

Table 3.1: Summary of simulation results over 1000 replicates. “Sens” means sensitivity of variable selection; “Spec” means specificity of variable selection; “MSE” means mean square error. “adpLASSO” means adaptive LASSO. The numbers in parentheses are the corresponding standard errors.

Method	Sens	Spec	MSE
Simulation 1			
LASSO	0.989 (0.002)	0.672 (0.009)	12.693 (0.030)
adpLASSO	0.919 (0.005)	0.652 (0.010)	13.126 (0.037)
SCAD	0.857 (0.006)	0.727 (0.010)	13.300 (0.042)
MCP	0.820 (0.006)	0.796 (0.010)	13.325 (0.041)
SAP	0.874 (0.006)	0.789 (0.008)	13.044 (0.038)
Simulation 2*			
LASSO	0.999 (0.001)	0.880 (0.003)	1.076 (0.003)
adpLASSO	0.934 (0.004)	0.866 (0.005)	1.193 (0.007)
SCAD	0.958 (0.003)	0.873 (0.002)	1.134 (0.004)
MCP	0.853 (0.005)	0.968 (0.001)	1.145 (0.005)
SAP	0.996 (0.001)	0.891 (0.003)	1.075 (0.003)
Simulation 3			
LASSO	0.939 (0.002)	0.891 (0.002)	16.454 (0.017)
adpLASSO	0.853 (0.004)	0.932 (0.003)	16.591 (0.039)
SCAD	0.891 (0.003)	0.971 (0.001)	15.887 (0.018)
MCP	0.778 (0.004)	0.998 (0.001)	16.000 (0.021)
SAP	0.819 (0.004)	0.985 (0.001)	15.961 (0.018)
Simulation 4			
LASSO	0.864 (0.005)	0.886 (0.003)	2.073 (0.010)
adpLASSO	0.590 (0.009)	0.849 (0.004)	3.281 (0.040)
SCAD	0.854 (0.006)	0.943 (0.001)	1.890 (0.008)
MCP	0.701 (0.007)	0.990 (0.001)	1.932 (0.010)
SAP	0.728 (0.007)	0.963 (0.003)	1.952 (0.010)

* The numbers in MSE column in simulation 2 are increased by 100 times.

Table 3.2: Real Data Analysis Summary: real data analysis results over 100 random splits. "Mean Square Error" column represents average mean square error on the test set over 100 random splits. "Selection Frequency" column represents the average number of variable selected over 100 splits; The numbers in parentheses are the corresponding standard errors.

Mean Square Error and Selection Frequency		
	Mean Square Error	Selection Frequency
LASSO	**0.932** (0.018)	25.75 (0.43)
adpLASSO	0.981 (0.018)	20.02 (0.91)
SCAD	0.988 (0.018)	18.58 (0.30)
MCP	1.010 (0.019)	11.16 (0.50)
SAP	**0.932** (0.018)	25.48 (0.47)

Table 3.3: Real Data Analysis Summary: Individual variable selection frequency across 100 random splits. Variables are sorted according to the selection frequency of SAP.

gene	LASSO	adpLASSO	SCAD	MCP	SAP
RPS28	100	82	94	78	100
TOP1	100	85	97	75	100
LAMB4	100	64	97	87	100
HEMK1	99	81	93	82	99
ZNF208	100	81	92	65	99
NPTXR	99	91	93	80	99
CLEC5A	98	74	95	75	98
CDK3	96	80	79	36	94
NELL1	94	39	66	34	93
GSN	93	75	84	62	93
HOMER1	92	70	76	51	91
PCNXL2	88	30	52	12	87
HIST3H2A	81	40	45	18	80
HOXD11	82	66	72	46	80
FKBP6	69	26	53	18	69
HOXD10	69	25	51	27	68
RBP1	68	12	63	29	68
BDH1	68	31	45	23	67
GRIA1	67	46	41	17	66
ZNF528	66	37	40	12	65
RANBP17	65	42	50	22	65
MAP6D1	64	27	40	19	62
MDK	54	34	29	13	52
GPRASP1	53	48	28	3	52
PARD3	53	29	19	8	51
TIMP1	51	61	24	8	50
MAP3K7IP1	44	46	13	6	44
tcag7.1314	43	23	32	6	41
SND1	35	18	10	2	34
SCG5	32	43	16	8	31
TOLLIP	28	31	11	7	28
RPL10	28	30	18	13	28
UPP1	27	4	11	2	27
CHI3L1	25	41	1	2	24
IQCG	24	26	7	3	24
NOL3	25	22	12	5	24
SPP1	24	12	20	4	24
ABCA1	20	19	8	5	20
SLC25A20	19	17	17	2	19
TMEM112	18	26	4	3	18
CRELD1	16	11	9	4	16
TRAF3IP1	12	12	7	4	12
DIRAS3	10	20	6	4	10
RPL36A	9	23	3	6	9
TCTA	7	13	5	3	7
EIF3H	5	10	3	2	6
CHL1	6	7	1	1	6
C7orf43	6	4	4	0	6
B3GAT3	5	7	0	1	5
LOC201229	5	10	2	4	4
GNG12	4	7	2	2	4
DKFZP564J102	4	3	3	0	4
DKK3	3	3	0	0	3
TMEM22	3	14	1	1	3
MAPK8	3	2	3	2	3
PHYHIP	2	8	0	1	2
HRASLS3	2	9	0	0	2
C13orf18	2	5	2	0	2
FBXO17	2	9	2	2	2
SLC4A3	2	3	0	1	2
EFEMP2	2	19	1	1	2
CAMK1	2	13	2	2	2
TREM2	1	13	1	1	1
KIAA0495	1	17	0	1	1
PGCP	0	6	1	2	0
B3GALNT1	0	15	1	2	0
PDPN	0	5	1	1	0

4 GROUP VARIABLE SELECTION VIA THE LOG-EXP-SUM PENALTY

4.1 Motivation

Breast cancer is the most common cancer in women younger than 45 years of age and is the leading cause of death among females in the United States. However, the survival rate for these young women with breast cancer has continuously improved over the past two decades, primarily because of improved therapies. With this long-term survival, it is important to study the quality of life that may be hampered by this traumatic event and by the long-term side effects from related cancer therapies (Berry et al., 2005).

This chapter is motivated by analyzing a dataset from a study funded by the American Cancer Society (ACS), a large quality of life study of breast cancer survivors diagnosed at a young age. The study included 505 breast cancer survivors (BCS) who were aged 18-45 years old at diagnosis and were surveyed 3-8 years after standard treatments. The study collected many covariates and quality of life outcomes. One outcome that is of particular interest is overall well being (OWB). It is captured by Campbell's index of well being which is measured from seven questionnaire items (Campbell et al., 1976). Studying the OWB status after an adversity is of great interest in an increasing body of research to comprehensively understand the consequences of a traumatic event, for example, cancer at a young age (Zwahlen et al., 2010).

In the present analysis, the covariates include demographic variables and social or behavior construct scores. The constructs are divided into eight non-overlapping groups: personality, physical health, psychological health, spiritual health, active coping, passive coping, social support and self efficacy. The constructs in each group are designed to measure the same aspect of the social or behavioral status of a breast cancer survivor from different angles. In our analysis, we are interested in identifying both important groups and important individual constructs within the selected groups that are related to OWB. These discoveries may help design interventions targeted at these young breast cancer survivors from the perspective of a cancer control program. In statistics, this is a group variable selection problem.

In this chapter, we propose a new Log-Exp-Sum penalty for group variable selection. This new penalty is convex, and it can perform variable selection at both group level and within-group level. We propose an effective algorithm based on a modified coordinate descent algorithm (Friedman et al., 2007; Wu and Lange, 2008) to fit the model. The theoretical properties of our proposed method are thoroughly studied. We establish both the finite sample error bounds and asymptotic group selection consistency of our LES estimator. The proposed method is applied to the ACS breast cancer survivor dataset.

The chapter is organized as follows. In Section 4.2, we propose the Log-Exp-Sum penalty and present the group-level coordinate descent algorithm. In Section 4.3, we develop non-asymptotic inequalities and group selection consistency for our LES estimator in the high-dimensional setting where the number of covariates is allowed to be much larger than the sample size. Simulation results are presented in Section 4.4. In Section 4.5, we apply the proposed method to ACS breast cancer dataset. Finally we conclude the chapter with Section 4.6.

4.2 Method

Log-Exp-Sum penalty

We consider the usual regression setup: we have training data, $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$, where $\mathbf{x}_i$ and y_i are a p-length vector of covariates and response for the ith subject, respectively. We assume the total of p covariates can be divided into K groups. Let the kth group have p_k variables, and we use $\mathbf{x}_{i,(k)} = (x_{i,k1}, \ldots, x_{i,kp_k})^T$ to denote the p_k covariates in the kth group for the ith subject. In most of this section, we assume $\sum_k p_k = p$, i.e., there are no overlap between groups. This is also the situation in ACS breast cancer survivor data. We will discuss the situation that groups are overlapped in Section 4.6.

To model the association between response and covariates, we consider linear regression:

$$y_i = \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj}\beta_{kj} + \epsilon_i, \ i = 1, \ldots, n, \tag{4.1}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms and β_{kj}'s are regression coefficients. We denote $\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kp_k})'$ to be the vector of regression coefficients for covariates in the kth group. Without loss of generality, we assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation, so the intercept term can be removed from the above regression model.

For the purpose of variable selection, we consider the penalized ordinary least square (OLS) estimation:

$$\min_{\beta_{kj}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj}\beta_{kj} \Big)^2 + \lambda \sum_{k=1}^{K} p(\boldsymbol{\beta}_k), \tag{4.2}$$

where $p(\cdot)$ is a sparsity-induced penalty function and λ is a non-negative tuning parameter.

Our LES penalty is motivated by modifying the group LASSO penalty to conduct both group and within-group selection. Note that the group LASSO penalty is a member of a penalty function family: $p(\boldsymbol{\beta}_k) = f^{-1}\Big\{ f(|\beta_{k1}|) + \cdots + f(|\beta_{kp_k}|) \Big\}$ by taking $f(x) = x^2$. Our LES penalty is another member of this family by taking $f(x) = \exp(x)$. To be specific, we propose the following LES penalty:

$$p(\boldsymbol{\beta}_k) = w_k \log \Big\{ \exp(\alpha|\beta_{k1}|) + \cdots + \exp(\alpha|\beta_{kp_k}|) \Big\}, \tag{4.3}$$

where $\alpha > 0$ is a tuning parameters and w_k's are pre-specified weights to adjust for different group sizes, for example, taking $w_k = p_k/p$. The LES penalty is strictly convex, which can be straightforwardly verified by calculating its second derivative. Similar to other group variable selection penalties, the LES penalty utilizes the group structure and is able to perform group selection. Meanwhile, the LES penalty is also singular at any $\beta_{kj} = 0$ point, and hence is able to conduct the within group selection as well.

The LES penalty has three connections to the LASSO penalty. First, when each group contains only one variable ($p_k = 1$), i.e., there is no group structure, the LES penalty reduces to the LASSO penalty. Second, when the design matrix $\mathbf{X}$ is orthonormal, i.e. $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the LES penalty yields the following penalized estimation:

$$\hat{\beta}_{kj} = sign(\hat{\beta}_{kj}^{ols})\left(|\hat{\beta}_{kj}^{ols}| - n\lambda\alpha w_k \frac{\exp\{\alpha|\hat{\beta}_{kj}|\}}{\sum_{l=1}^{p_k}\exp\{\alpha|\hat{\beta}_{kl}|\}}\right)_+, \tag{4.4}$$

where $\hat{\beta}_{kj}^{ols}$'s are the unpenalized ordinary least square estimates, and function $(a)_+ \triangleq max\{0, a\}$ for any real number a.

Then at the group level, we have the following result:

$$\sum_{j=1}^{p_k}|\hat{\beta}_{kj}| = \sum_{j:\hat{\beta}_{kj}\neq 0}|\hat{\beta}_{kj}^{ols}| - n\lambda\alpha w_k \frac{\sum_{j:\hat{\beta}_{kj}\neq 0}\exp\{\alpha|\hat{\beta}_{kj}|\}}{\sum_l \exp\{\alpha|\hat{\beta}_{kl}|\}}. \tag{4.5}$$

Note that when each estimate $\hat{\beta}_{kj}$ in the kth group is non-zero, the following equality holds:

$$\sum_{j=1}^{p_k}|\hat{\beta}_{kj}| = \sum_{j=1}^{p_k}|\hat{\beta}_{kj}^{ols}| - n\lambda\alpha w_k \qquad if \quad \hat{\beta}_{kj} \neq 0, \quad \forall\, j = 1, \ldots p_k, \tag{4.6}$$

which can be viewed as an extension from the thresholding on the individual coefficient by the LASSO penalty to the thresholding on the L_1-norm of the coefficients in each group.

The third connection between the LES penalty and the LASSO penalty is that, given any design matrix $\mathbf{X}$ (not necessarily orthonormal) and an arbitrary grouping structure ($p_k \geqslant 1$), the LASSO penalty can be viewed as a limiting case of the LES penalty. To be specific, we have the following proposition.

Proposition 4.1. *Given the data, for any positive number γ, consider the LASSO estimator and LES estimator as follows:*

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{LASSO} &= \arg\min_{\boldsymbol{\beta}} OLS + \gamma \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}|,\\ \hat{\boldsymbol{\beta}}^{LES} &= \arg\min_{\boldsymbol{\beta}} OLS + \lambda \sum_{k=1}^{K} \frac{p_k}{p} \log\Big(\exp\{\alpha|\beta_{k1}|\} + \cdots + \exp\{\alpha|\beta_{kp_k}|\}\Big),\end{aligned}$$

where $OLS = \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\Big)^2$.

Then we have

$$\hat{\boldsymbol{\beta}}^{LES} - \hat{\boldsymbol{\beta}}^{LASSO} \to \mathbf{0}, \text{ as } \alpha \to 0 \text{ and keeping } \lambda\alpha/p = \gamma.$$

The proof of Proposition 4.1 is given in the Appendix.

Our LES penalty has a property that the estimated coefficients of highly correlated variables within the same group are enforced to be similar to each other. As a consequence of this, when applied to the ACS Breast Cancer Survivor dataset, since the construct scores within the same group can be highly correlated, our LES penalty tends to select or remove the highly correlated constructs within a group together. To be specific, we have the following proposition.

Proposition 4.2. *Let $\hat{\boldsymbol{\beta}}$ be the penalized OLS estimation with the LES penalty. If $\hat{\beta}_{ki}\hat{\beta}_{kj} > 0$, then we have:*

$$|\hat{\beta}_{ki} - \hat{\beta}_{kj}| \leqslant C\sqrt{2(1-\rho_{ki,kj})},$$

where $\rho_{ki,kj} = X_{ki}^T X_{kj}$ is the sample correlation between X_{ki} and X_{kj} and constant C is given by:

$$\frac{1}{n\lambda\alpha^2 w_k}\sqrt{\|\mathbf{y}\|_2^2 + 2n\lambda\sum_{l=1}^{K} w_l \log(p_l)\exp\Big\{\frac{1}{2n\lambda w_k}\|\mathbf{y}\|_2^2 + \sum_{l=1}^{K}\frac{w_l}{w_k}\log(p_l)\Big\}}$$

.

The proof of Proposition 4.2 is given in the Appendix.

Algorithm

We need to solve the following optimization problem:

$$\min_{\beta_{kj}} Q(\beta_{kj}) = \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\Big)^2 + \lambda\sum_{k=1}^{K} w_k \log\Big\{\sum_{l=1}^{p_k}\exp(\alpha|\beta_{kl}|)\Big\}. \tag{4.7}$$

We propose applying the coordinate descent algorithm (Friedman et al., 2007; Wu and Lange, 2008) at the group level. The key idea is to find the minimizer of the original high-dimensional optimization problem (4.7) by solving a sequence of low-dimensional optimization problems, each of which only involves the parameters in one group. To be specific, for fixed $\tilde{\boldsymbol{\beta}}_{-k} = (\tilde{\boldsymbol{\beta}}_1', \ldots, \tilde{\boldsymbol{\beta}}_{k-1}', \tilde{\boldsymbol{\beta}}_{k+1}', \ldots, \tilde{\boldsymbol{\beta}}_K)$, define the function $Q_k = Q(\boldsymbol{\beta}_k; \tilde{\boldsymbol{\beta}}_{-k})$. Our group-level coordinate descent algorithm is implemented by minimizing Q_k with respect to $\boldsymbol{\beta}_k$ for each k at a time, and using the solution to update $\boldsymbol{\beta}$; at the next step, $Q_{k+1} = Q(\boldsymbol{\beta}_{k+1}; \tilde{\boldsymbol{\beta}}_{-(k+1)})$ is minimized and the minimizer is again used to update $\boldsymbol{\beta}$. In this way, we cycle through the indices $k = 1, \ldots, K$. This cycling procedure may repeat for multiple times until some convergence criterion is reached.

To minimize $Q_k(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$, we need to solve the following optimization problem:

$$\min_{\boldsymbol{\beta}_k} \frac{1}{2n}\|\mathbf{c} - \mathbf{A}\boldsymbol{\beta}_k\|_2^2 + \lambda w_k \log\Big(\exp\{\alpha|\beta_{k1}|\} + \cdots + \exp\{\alpha|\beta_{kp_k}|\}\Big), \tag{4.8}$$

where $\mathbf{c}$ is a $n-$length vector with $c_i = y_i - \sum_{l\neq k}\sum_{j=1}^{p_l} x_{i,lj}\tilde{\beta}_{lj}$, and $\mathbf{A}$ is a $n \times p_k$ matrix with $\mathbf{A}_{ij} = x_{i,kj}$. We propose applying the gradient projection method to get the solution. This method was shown in (Figueiredo et al., 2007) to be very computationally efficient.

In order to solve the optimization problem (4.8), we consider a transformation of the original problem. To be specific, let $\mathbf{u}$ and $\mathbf{v}$ be two vectors such that $u_i = (\beta_{ki})_+$

and $v_i = (-\beta_{ki})_+$, where $(a)_+ = max\{0, a\}$. Then the optimization problem (4.8) is equivalent to:

$$\min_{\mathbf{z}} \frac{1}{2n}\mathbf{z}^T\mathbf{B}\mathbf{z} + \mathbf{d}^T\mathbf{z} + f(\mathbf{z}) \equiv F(\mathbf{z}) \quad s.t. \quad \mathbf{z} \geqslant 0, \tag{4.9}$$

$$\text{where} \qquad \mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} -\mathbf{A}^T\mathbf{c} \\ \mathbf{A}^T\mathbf{c} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{A}^T\mathbf{A} & -\mathbf{A}^T\mathbf{A} \\ -\mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{A} \end{bmatrix},$$

$$f(\mathbf{z}) = \lambda w_k \log\Big[\exp\{\alpha(z_1 + z_{p_k+1})\} + \cdots + \exp\{\alpha(z_{p_k} + z_{2p_k})\}\Big].$$

Then we apply the Barzilai-Borwein method in Figueiredo et al. (2007) to get the solution to (4.9):

Step 1 (Initialization): Initialize $\mathbf{z}$ with some reasonable value $\mathbf{z}^{(0)}$. Choose ϕ_{min} and ϕ_{max}, the lower bound and upper bound for the line search step length. Choose $\phi \in [\phi_{min}, \phi_{max}]$, which is an initial guess of the step length. Set the backtracking line search parameter $\rho = 0.5$, and set iteration number $t = 0$.

Step 2 (Backtracking Line Search): Choose $\phi_{(t)}$ to be the first number in the sequence $\phi, \rho\phi, \rho^2\phi, \ldots$, such that the following "Armijo condition" is satisfied:

$$F\Big(\big(\mathbf{z}^{(t)} - \phi_{(t)}\nabla F(\mathbf{z}^{(t)})\big)_+\Big) \leqslant F\Big(\mathbf{z}^{(t)}\Big).$$

Step 3 (Update $\mathbf{z}$) Let $\mathbf{z}^{(t+1)} \leftarrow \big(\mathbf{z}^{(t)} - \phi_{(t)}\nabla F(\mathbf{z}^{(t)})\big)_+$, and compute $\delta^{(t)} = \mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}$.

Step 4 (Update Step Length): Compute $\gamma^{(t)} = \nabla F(\mathbf{z}^{(t+1)}) - \nabla F(\mathbf{z}^{(t)})$, and update step length:

$$\phi = median\{\phi_{min}, \frac{||\delta^{(t)}||_2^2}{(\gamma^{(t)})^T\delta^{(t)}}, \phi_{max}\}$$

Step 5 (Termination) Terminate the iteration if $||\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}||_2$ is small; otherwise, set $t \leftarrow t+1$, and go to step 2. In our numerical experiment, we terminate the iteration with tolerance of 1e-6.

Since our objective function is convex and the LES penalty is separable at the group level, by results in Tseng (2001), our algorithm is guaranteed to converge to the global minimizer. Note that, if we apply the coordinate descent algorithm at the individual coefficient level, the algorithm is not guaranteed to converge, and in our numerical study, we observed that sometimes the updates were trapped in a local area.

Tuning parameter selection

Tuning parameter selection is an important issue in penalized estimation. One often proceeds by finding estimators which correspond to a range of tuning parameter values. The preferred estimator is then identified as the one for the tuning parameter value to optimize some criterion, such as Cross Validation (CV), Generalized Cross Validation (GCV) (Craven and Wahba, 1978), AIC (Akaike, 1973), or BIC (Schwarz, 1978). It is known that CV, GCV and AIC-based methods favor the model with good prediction performance, while BIC-based method tends to identify the correct model (Yang, 2005). To implement GCV, AIC and BIC, one needs to estimate the degrees of freedom (df) of an estimated model. For our LES penalty, the estimate of df does not have an analytic form even when the design matrix is orthonormal. Therefore, we propose using the randomized trace method (Girard, 1987, 1989; Hutchinson, 1989) to estimate df numerically.

We first briefly review the randomized trace method to estimate df of a model which is linear in response $\mathbf{y}$. Let $\hat{\mathbf{y}}$ be the estimation of the response $\mathbf{y}$ based on the model, which is given by:

$$\hat{\mathbf{y}} = \mathbf{A}(\lambda)\mathbf{y} \triangleq f(\mathbf{y}), \tag{4.10}$$

where $\mathbf{A}(\lambda)$ is the so-called "influence matrix", which depends on the design matrix $\mathbf{X}$ and the tuning parameter λ, but does not depend on the response $\mathbf{y}$. Wahba (1983) defined $tr(\mathbf{A}(\lambda))$, the trace of $\mathbf{A}(\lambda)$, to be "equivalent degrees of freedom when λ is used".

The randomized trace method is to estimate $tr(\mathbf{A}(\lambda))$ based on the fact that, for any random variable $\boldsymbol{\delta}$ with mean zero and covariance matrix $\rho^2 I$, we have $tr(\mathbf{A}(\lambda)) = E\boldsymbol{\delta}^T\mathbf{A}(\lambda)\boldsymbol{\delta}/\rho^2$. To be specific, we generate a new set of responses by adding some random perturbations to the original observations: $\mathbf{y}^{new} = \mathbf{y} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \sim N(\mathbf{0}, \rho^2 I)$. Then $tr(\mathbf{A}(\lambda))$ (and hence df) can be estimated by

$$\tilde{df} = \boldsymbol{\delta}^T\mathbf{A}(\lambda)\boldsymbol{\delta}/\rho^2 = \boldsymbol{\delta}^T\mathbf{A}(\mathbf{y}^{new} - \mathbf{y})/\rho^2 \approx \frac{\boldsymbol{\delta}^T\Big(f(\mathbf{y}+\boldsymbol{\delta}) - f(\mathbf{y})\Big)}{\frac{1}{n}\sum_{i=1}^n \delta_i^2}. \tag{4.11}$$

To reduce the variance of the estimator $\tilde{df}$, we can first generate R independent noise vectors: $\boldsymbol{\delta}_{(r)}$, $r = 1, \ldots, R$, and then estimate df by $\tilde{df}$ using each $\boldsymbol{\delta}_{(r)}$. The final estimate of df can be the average of these R estimates:

$$\hat{df} = \frac{1}{R}\sum_{r=1}^{R} \frac{\boldsymbol{\delta}_{(r)}^T\Big(f(\mathbf{y}+\boldsymbol{\delta}_{(r)}) - f(\mathbf{y})\Big)}{\frac{1}{n}\boldsymbol{\delta}_{(r)}^T\boldsymbol{\delta}_{(r)}}, \tag{4.12}$$

which is called an R-fold randomized trace estimate of df.

The estimated model by LES is nonlinear in response $\mathbf{y}$, i.e., in equation (4.10), the influence matrix $\mathbf{A}$ depends on the response $\mathbf{y}$. In general, the estimated models by most penalized estimation methods (like LASSO) are nonlinear in response $\mathbf{y}$. Lin et al. (2000) and Wahba et al. (1995) discussed using the randomized trace method to estimate df of a nonlinear model. Following their discussions, when an estimated model is linear, i.e., equation (4.10) is satisfied, we can see that $tr\{\mathbf{A}(\lambda)\} = \sum_{i=1}^n \frac{\partial f(\mathbf{y})_i}{\partial y_i}$. So the randomized trace estimate of df can be viewed as an estimation of $\sum_i \frac{\partial f(\mathbf{y})_i}{\partial y_i}$ using divided differences. When the model is nonlinear, a response-independent influence matrix $\mathbf{A}(\lambda)$ does not exist. However, the divided differences $\frac{f(\mathbf{y}+\boldsymbol{\delta})_i - f(\mathbf{y})_i}{\delta_i}$ generally exist, so we can still estimate df by the routine R-fold randomized trace estimator defined in (4.12).

In our numerical experiments, we found that the 5-fold randomized trace estimator worked well. Its computation load is no heavier than the 5-fold cross-validation. In addition, as a proof of concept, we conducted several simulation studies to estimate df of LASSO. Our simulation results (not presented in this chapter) showed that the 5-fold randomized trace estimates of df for LASSO were very close to the number of non-zero estimated regression coefficients, which is given in Zou and Hastie (2005) as an estimator of df for LASSO.

4.3 Theoretical Results

In this section, we present the theoretical properties of our LES estimator. We are interested in the situation when the number of covariates is much larger than the number of observations, i.e., $p >> n$. We first establish the non-asymptotic error bounds for the LES estimator. Then we study the asymptotic group selection consistency for the LES estimator. Throughout the whole section, we consider the following LES penalized OLS estimation:

$$\min_{\beta_{kj}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj} \beta_{kj} \Big)^2 + \lambda \sum_{k=1}^{K} p_k \log\Big(\exp\{\alpha|\beta_{k1}|\} + \cdots + \exp\{\alpha|\beta_{kp_k}|\} \Big). \tag{4.13}$$

Non-asymptotic error bounds

In this subsection, we extend the argument in Bickel et al. (2009) to establish finite-sample bounds for our LES estimator. We make the following Restricted Eigenvalue assumption with group structure (REgroup), which is similar to the Restricted Eigenvalue (RE) assumption in Bickel et al. (2009).

REgroup assumption: Assume group structure is prespecified and p covariates can be divided into K groups with p_k covariates in each group. For a positive integer s and any $\boldsymbol{\Delta} \in \mathbb{R}^p$, the following condition holds:

$$\kappa(s) \triangleq \min_{\substack{G \subseteq \{1,\ldots,K\},\\ |G| \leqslant s}} \; \min_{\substack{\forall \boldsymbol{\Delta} \neq 0,\\ \sum_{k \notin G} \|\boldsymbol{\Delta}_k\|_1 \leqslant \sum_{k \in G}(1+2p_k)\|\boldsymbol{\Delta}_k\|_1}} \frac{2\|\mathbf{X}\boldsymbol{\Delta}\|_2}{\sqrt{n}\sqrt{\sum_{k \in G} p_k(1+p_k)^2 \|\boldsymbol{\Delta}_k\|_2^2}} > 0,$$

where G is a subset of $\{1, \ldots, K\}$, and $|G|$ is the cardinality of set G. $\boldsymbol{\Delta}_k \in \mathbb{R}^{p_k}$ is a subvector of $\boldsymbol{\Delta}$ for the k-th group, i.e. $\boldsymbol{\Delta}_k = (\Delta_{k1}, \ldots, \Delta_{kp_k})^T$. We denote $\| \cdot \|_2$ and $\| \cdot \|_1$ to be Euclidean norm and L_1-norm, respectively.

Theorem 4.3. *Consider linear regression model (4.1). Let $\boldsymbol{\beta}^*$ be the vector of true regression coefficients. Assume the random error terms $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. from the normal distribution with mean zero and variance σ^2. Suppose the diagonal elements of matrix $\mathbf{X}^T\mathbf{X}/n$ are equal to 1.*

Let $G(\boldsymbol{\beta})$ be the set of indices of groups that contain at least one nonzero element for a vector $\boldsymbol{\beta}$, i.e. $G(\boldsymbol{\beta}) = \{ \, k \mid \exists \, j, \; 1 \leqslant j \leqslant p_k, \; s.t : \beta_{kj} \neq 0; 1 \leqslant k \leqslant K\}$. Assume the REgroup assumption holds with $\kappa = \kappa(s) > 0$, where $s = |G(\boldsymbol{\beta}^*)|$. Let A be a real number bigger than $2\sqrt{2}$ and $\gamma = A\sigma\sqrt{\frac{\log p}{n}}$. Let two tuning parameters satisfy $\lambda\alpha = \gamma$. Denote $\hat{\boldsymbol{\beta}}$ to be the solution to optimization problem (4.13). Then with probability at least $1 - p^{1-A^2/8}$, the following inequalities hold:

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \;\leqslant\; \frac{16A^2\sigma^2 s}{\kappa^2} * \frac{\log p}{n}; \tag{4.14}$$

$$\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_1 \leqslant \frac{16A\sigma s}{\kappa^2} * \sqrt{\frac{\log p}{n}}; \tag{4.15}$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \;\leqslant\; (2\sqrt{s}+1)\frac{8A\sigma\sqrt{s}}{\kappa^2} * \sqrt{\frac{\log p}{n}}. \tag{4.16}$$

The proof of Theorem 4.3 is given in the Appendix. From Theorem 4.3, under certain conditions, for any n, p and any design matrix $\mathbf{X}$, with certain probability, we obtained the upper bounds on the estimation errors with prediction loss, L_1-norm loss and Euclidian-norm loss.

We can generalize our non-asymptotic results to asymptotic results if we further assume that $\kappa(s)$ is bounded from zero, i.e., there exists a constant $u > 0$, such that $\kappa(s) \geqslant u > 0, \forall\, n,\ p$ and $\mathbf{X}$. Then as $n \to \infty$, $p \to \infty$ and $\log p/n \to 0$, we have

$$\frac{1}{n}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \to 0,\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \to 0,\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \to 0, \tag{4.17}$$

which implies the consistency in estimation when the tuning parameters are properly selected.

In addition, the LASSO estimator is a special case of our LES estimator when each group has only one variable. If in Theorem 4.3, $p_k = 1\ \forall\, k$, our REgroup assumption is exactly the same as RE assumption in Bickel et al. (2009). And we obtain exactly the same bounds in (11) and (12) as presented in Theorem 7.2 in Bickel et al. (2009). Therefore, our results can be viewed as an extension of results in Bickel et al. (2009) from the setting without group structure to the setting with group structure.

Group selection consistency

In this subsection, we study the asymptotic group selection consistency for the LES estimator when both p and n tend to infinity. We would like to show that, with probability tending to 1, the LES estimator will select all important groups (groups that contain at least one important variable) while removing all unimportant groups.

Let $\mathscr{O}$ be the event that there exists a solution $\hat{\boldsymbol{\beta}}$ to optimization problem (4.13) such that $\|\hat{\boldsymbol{\beta}}_k\|_\infty > 0$ for all $k \in G(\boldsymbol{\beta}^*)$ and $\|\hat{\boldsymbol{\beta}}_k\|_\infty = 0$ for all $k \notin G(\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ is the vector of true regression coefficients for model (4.1) and $G(\boldsymbol{\beta}^*)$ is the set of indices of groups that contain at least one nonzero element for a vector $\boldsymbol{\beta}^*$. We would like to show the group selection consistency as the following:

$$P(\mathscr{O}) \to 1, \qquad n \to \infty. \tag{4.18}$$

Following Nardi and Rinaldo (2008), we make the following assumptions. For notation simplicity, in the remaining of this subsection, we use G to stand for $G(\boldsymbol{\beta}^*)$.

(C1) The diagonal elements of matrix $\mathbf{X}^T\mathbf{X}/n$ are equal to 1;

(C2) Let $p_0 = \sum_{k \in G} p_k$ and $d_n = \min_{k \in G} \|\boldsymbol{\beta}_k^*\|_\infty$, where $\|\cdot\|_\infty$ is the L_∞-norm. Denote c to be the smallest eigenvalue of $\frac{1}{n}\mathbf{X}_G^T\mathbf{X}_G$ and assume $c > 0$, where $\mathbf{X}_G$ is the submatrix of $\mathbf{X}$ formed by the columns whose group index is in G. Assume

$$\frac{1}{d_n}\Big(\sqrt{\frac{\log p_0}{nc}} + \frac{\lambda\alpha\sqrt{p_0}}{c}\max_{k \in G} p_k\Big) \to 0; \tag{4.19}$$

(C3) For some $0 < \tau < 1$,

$$\|\mathbf{X}_{G^c}^T\mathbf{X}_G^T(\mathbf{X}_G^T\mathbf{X}_G)^{-1}\|_\infty \leqslant \frac{1-\tau}{\max_{k \in G} p_k}, \tag{4.20}$$

where $\mathbf{X}_{G^c}$ is the submatrix of $\mathbf{X}$ formed by the columns whose group index is not in G;

(C4)

$$\frac{1}{\lambda\alpha}\sqrt{\frac{\log(p - p_0)}{n}} \to 0. \tag{4.21}$$

Condition $(C1)$ assumes the matrix $\mathbf{X}$ is standardized, which is a common procedure in practice. Condition $(C2)$ essentially requires the minimum among the strongest signal $\|\boldsymbol{\beta}_k^*\|_\infty$ of important groups can not be too small; and the eigenvalue of $\frac{1}{n}\mathbf{X}_G^T\mathbf{X}_G$ does not vanish too fast. Notice that the dimension of $\mathbf{X}_G$ is $n \times p_0$, this implicitly requires p_0 can not be larger than n. Condition $(C3)$ controls the multiple regression coefficients of unimportant group covariates $\mathbf{X}_{G^c}$ on the important group covariates $\mathbf{X}_G$. This condition mimics the irrepresentable condition as in Zhao and Yu (2006) which assumes no group structure. The bound of condition $(C3)$ actually depends on the choice of weights w_i's. By choosing a different set of weights, we could obtain a more relaxed bound in $(C3)$. Finally, condition $(C4)$ controls the growth rate of p and n where p can grow at exponential rate of n.

Theorem 4.4. *Consider linear regression model (4.1), under the assumptions $(C1) - (C4)$, the sparsity property (4.18) holds for our LES estimator.*

The proofs follow the spirit in Nardi and Rinaldo (2008) and the details are given in the Appendix. It is in general difficult to study the individual variable selection consistency of a group variable selection method. Because it is hard to obtain consistency estimators for the unimportant variables in the important groups. So we only focus our discussions on the group selection consistency rather than extending to individual variable selection consistency.

4.4 Simulation Studies

In this section, we perform simulation studies to evaluate the finite sample performance of the LES method, and compare the results with several existing methods, including LASSO, group LASSO (gLASSO), group bridge (gBrdige) and sparse group LASSO (sgLASSO). We consider four examples. All examples are based on the linear regression model: $y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i,\ i = 1, \ldots, n,$ where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. We chose σ to control the signal-to-noise ratio to be 3. The details of the settings are described as follows.

> **Example 1 ("All-In-All-Out")** There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables in each group. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I})$. The true $\boldsymbol{\beta}^*$ was specified as:
>
> $$\boldsymbol{\beta}^* = [\underbrace{2, 2, 2, -2, -2}_{group1}, \underbrace{0, 0, 0, 0, 0}_{group2}, \underbrace{0, 0, 0, 0, 0}_{group3}, \underbrace{0, 0, 0, 0, 0}_{group4}, \underbrace{0, 0, 0, 0, 0}_{group5}]^T. \tag{4.22}$$
>
> **Example 2 ("Not-All-In-All-Out")** There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables in each group. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$,

where $\mathbf{\Sigma}$ was given by:

$$\Sigma = \begin{pmatrix} P & & & & \\ & P & & & \\ & & Q & & \\ & & & Q & \\ & & & & Q \end{pmatrix} \quad with\ P = \begin{pmatrix} 1 & .7 & .7 & .1 & .1 \\ .7 & 1 & .7 & .1 & .1 \\ .7 & .7 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .7 \\ .1 & .1 & .1 & .7 & 1 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & .7 & .7 & .7 & .7 \\ .7 & 1 & .7 & .7 & .7 \\ .7 & .7 & 1 & .7 & .7 \\ .7 & .7 & .7 & 1 & .7 \\ .7 & .7 & .7 & .7 & 1 \end{pmatrix}.$$

The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{2,2,2,0,0}_{group1},\underbrace{2,2,2,0,0}_{group2},\underbrace{0,0,0,0,0}_{group3},\underbrace{0,0,0,0,0}_{group4},\underbrace{0,0,0,0,0}_{group5}]^T. \qquad (4.23)$$

Example 3 (mixture) There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables in each group. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ was the same as in simulation setting 2. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{0,0,0,2,2}_{group1},\underbrace{0,0,0,2,2}_{group2},\underbrace{1,1,1,1,1}_{group3},\underbrace{1,1,1,1,1}_{group4},\underbrace{0,0,0,0,0}_{group5}]^T. \qquad (4.24)$$

Example 4 (mixture) There are $K = 6$ groups and $p = 50$ variables in total. For group 1, 2, 4 and 5, each contains 10 variables; for group 3 and 6, each contains 5 variables. We generated $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Phi})$, where $\mathbf{\Phi}$ is a block diagonal matrix given by diag$(\mathbf{\Sigma}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ here was the same as in simulation setting 2 and 3. The true $\boldsymbol{\beta}^*$ was specified as:

$$\begin{aligned}\boldsymbol{\beta}^* = [&\underbrace{0,0,0,2,2,0,0,0,2,2}_{group1},\underbrace{1,1,1,1,1,0,0,0,0,0}_{group2},\underbrace{1,1,1,1,1}_{group3},\\ &\underbrace{0,0,0,0,0,0,0,0,0,0}_{group4},\underbrace{0,0,0,0,0,0,0,0,0,0}_{group5},\underbrace{0,0,0,0,0}_{group6}]^T. \qquad (4.25)\end{aligned}$$

For each setup, the sample size is $n = 100$. We repeated simulations 1,000 times. The LES was fitted using the algorithm described in Section 2. The LASSO was fitted using the R package "glmnet" (Friedman et al., 2010b). The group LASSO was fitted using the R package "grplasso". The group bridge was fitted using the R package "grpreg". The sparse group LASSO was fitted using the R package "SGL".

To select the tuning parameters in each of the five methods, we consider two approaches. The first approach is based on data validation. To be specific, in each simulation, besides the training data, we also independently generated a set of tuning data with the same distribution and with a same sample size as the training data. Then for each tuning parameter, we fitted the model on the training data and used the fitted model to predict the response on the tuning set and calculated the corresponding mean square error (prediction error). The model with the smallest tuning error was selected.

Our second approach for tuning parameter selection is based on BIC, which is defined to be:

$$BIC = \log(\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_2^2/n) + \log n \cdot df/n,$$

where df is the degrees of freedom of an estimated model. This format of BIC is based on the profile likelihood to get rid of σ^2, the variance of the errors. It is used in Wang et al. (2007a) and was shown to have a good performance. For the LES method, the df was estimated using the randomized trace method described in Section 4.2. For the LASSO method, the df was estimated by the number of non-zero estimated coefficients (Zou and Hastie, 2005). For the group LASSO method, the df was estimated as suggested in Yuan and Lin (2006). For the group bridge method, the df was estimated as suggested in Huang et al. (2009). For the sparse group LASSO method, the corresponding papers did not consider estimation of df, and we used the number of non-zero estimated coefficients as the estimator for its df.

To evaluate the variable selection performance of methods, we consider sensitivity (Sens) and specificity (Spec), which are defined as:

$$\text{Sens} = \frac{\text{\# of selected important variables}}{\text{\# of important variables}},$$

$$\text{Spec} = \frac{\text{\# of removed unimportant variables}}{\text{\# of unimportant variables}}.$$

For both sensitivity and specificity, higher value means a better variable selection performance. For each of five methods considered in our simulation, we further obtain the sensitivities and specificities of models along its full solution paths of (by fitting models with many tuning parameter values), create the ROC curve with respect to these sensitivities and specificity, and calculate the corresponding area under curve (AUC). For all five methods, it is possible that several models have the same specificity but different sensitivity. When this happens, we use the highest sensitivity to construct the ROC curve, representing the best variable selection performance of the method.

To evaluate the prediction performance of methods, following Tibshirani (1996), we consider the model error (ME) which is defined as:

$$\text{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector, $\boldsymbol{\beta}^*$ is the true coefficient vector, and $\boldsymbol{\Sigma}$ is the covariance matrix of the design matrix $\mathbf{X}$. We would like to acknowledge that the model error is closely related to the predictive mean square error proposed in Wahba (1985) and Leng et al. (2006). We also calculate the bias of estimator defined as Bias $= \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$.

The simulation results are summarized in Table 4.1.

In Example 1, the group bridge method has the lowest model error as well as the highest specificity. This is not surprising, because Example 1 is a relatively simple "All-In-All-Out" case, i.e., all covariates in a group are either all important or all unimportant. Under this situation, the non-convex group bridge penalty has an advantage over other methods in terms of removing unimportant groups. Although slightly worse than the group bridge method, the LES method outperformed the other three methods in terms of model error. Note that, because of the diagonal covariance matrix of X in this example, the bias is exactly the same as the model error. All five methods had identical AUC values.

In Example 2, the LES method produced the smallest model error when the tuning set approach was used, and produced the smallest bias when the BIC tuning was used. No method dominated in specificity. All five methods had almost identical sensitivities. Except group LASSO, the other four methods had almost identical AUC values as well. Note that the under-performance of LES when BIC tuning criteria is used, may be due to the inefficiency of randomized trace method to capture the true degrees of freedom of the model.

In Example 3, the LES method produced the smallest model errors no matter which tuning criterion was used. It has the smallest bias when BIC tuning was used as well. The group LASSO method had the highest sensitivity, but its specificity was very low. This means that the group LASSO method tended to include a large amount of variables in the model. The LES method had the highest AUC value among five methods.

Example 4 is similar to Example 3, but has more covariates and more complex group structure. The conclusion about comparisons is similar to that in Example 3. One difference is that, both the LES method and the sparse group LASSO method had the highest AUC values among five methods.

4.5 American Cancer Society Breast Cancer Survivor Data Analysis

In this section, we analyze the data from ACS breast cancer study which was conducted at the Indiana University School of Nursing. The participants of the study were survivors of the breast cancer aged $18-45$ years old at diagnosis and were surveyed between $3-8$ years from completion of chemotherapy, surgery, with or without radiation therapy. The purpose of the present analysis is to find out what factors in the psychological, social and behavior domains are important for the OWB of these survivors. Identification of these factors and establishment of their association with OWB may help develop intervention programs to improve the quality of life of breast cancer survivors.

The variables included in our current analysis are 54 social and behavior construct scores and three demographic variables. The 54 scores are divided into eight non-overlapping groups: personality, physical health, psychological health, spiritual health, active coping, passive coping, social support and self efficacy. Each group contains up to 15 different scores. The three demographic variables are: "age at diagnosis" (Agediag), "years of education" (Yrseduc) and "How many months were you in initial treatment for breast cancer" (Bcmths). We treated each demographic variable as an individual group. There are 6 subjects with missing values in either covariates or response, and we removed them from our analysis. In summary, we have 499 subjects and 57 covariates in 11 groups in our analysis.

We applied five methods in the data analysis: LASSO, group LASSO, group bridge sparse group LASSO and our LES method. We randomly split the whole dataset into a training set with sample size $n = 332$ and a test set with sample size $n = 167$ (the ratio of two sample sizes is about $2 : 1$). We fitted models on the training set, using two tuning strategies: one used 10-fold CV, the other used BIC. The BIC tuning procedure for all of the five methods is the same as what we described in the simulation studies. We then evaluated the prediction performances on the test set. We repeated the whole procedure beginning with a new random split 100 times.

The upper part of Table 4.2 summarizes, over 100 replications, the average number of selected groups, the average number of selected individual variables, and the average mean square errors (MSE) on the test sets, for the five methods. We can see that, for all five methods, the models selected by the 10-fold CV tuning had smaller MSEs (better prediction performance) than the models selected by the BIC tuning. As the cost of this gain in prediction performance, the models selected by 10-fold CV tuning included more groups and more individual variables than the models selected by BIC tuning. We can also see that, our LES methods had the smallest MSE among five methods no matter which tuning strategy was used.

The lower part of Table 4.2 summarizes the selection frequency of each group across 100 replicates. A group is considered to be selected if at least one variable within the group is selected. Since there are some theoretical works showing that

BIC tuning tends to identify the true model (Wang et al., 2007a), we focus on the selection results with BIC tuning. We can see that the psychological health group is always selected by all of five methods. For our LES methods, three other groups have very high selection frequency: spiritual health (91 out of 100), active coping (89 out of 100) and self efficacy (99 out of 100). These three groups are considered to be importantly associated with OWB in literature. Spirituality is a resource regularly used by patients with cancer coping with diagnosis and treatment (Gall et al., 2005). Purnell and Andersen (2009) reported that spiritual well-being was significantly associated with quality of life and traumatic stress after controlling for disease and demographic variables. Self-efficacy is the measure of one's own ability to complete tasks and reach goals, which is considered by psychologists to be important for one to build a happy and productive life (Parle et al., 1997). Rottmann et al. (2010) assessed the effect of self-efficacy and reported a strong positive correlation between self-efficacy and quality of life and between self-efficacy and mood. They also suggested that self-efficacy is a valuable target of rehabilitation programs. Coping refers to "cognitive and behavioral efforts made to master, tolerate, or reduce external and internal demands and conflicts" (Folkman and Lazarus, 1980). The coping strategies are usually categorized into two aspects: active coping and passive coping (Carrico et al., 2006). Active coping efforts are aimed at facing a problem directly and determining possible viable solutions to reduce the effect of a given stressor. Meanwhile, passive coping refers to behaviors that seek to escape the source of distress without confronting it (Folkman and Lazarus, 1985). Setting aside the nature of individual patients or specific external conditions, there have been consistent findings that the use of active coping strategies produce more favorable outcomes compared to passive coping strategies, such as less pain as well as depression, and better quality of life (Holmes and Stevenson, 1990). Another interesting observation is that, compared to other methods, our LES method identified much more frequently the importance of Social Support (including communication with health care team both at diagnosis and at follow up, and support from health care providers). There seems to be more awareness for the importance of this construct both scientifically and publicly. In the New York

Times Science Section of 10-Feb-2014, Dr. Arnold S. Relman, a prominent Medical Professor, Writer and Editor, discussed his experience as a hospital patient, where he found out how very important his interactions with nurses were.

In addition, the within group selection results from our LES method provide insights about which aspects/items within selected constructs are most important. The details of within group selection results of five methods are presented in the figures below. For example, positive reframing/thinking and religious coping are two most frequently picked items from the Active coping group. Other items such as emotional support, planning, acceptance are not frequently picked. When designing interventions to boost Active coping for patients, focus may be directed towards positive reframing and religious coping.

4.6 Conclusion and Discussion

In this chapter, we propose a new convex Log-Exp-Sum penalty for group variable selection. The new method keeps the advantage of group LASSO in terms of effectively removing unimportant groups, and at the same time enjoys the flexibility of removing unimportant variables within identified important groups. We have developed an effective group-level coordinate descent algorithm to fit the model. The theoretical properties of our proposed method have been thoroughly studied. We have established non-asymptotic error bounds and asymptotic group selection consistency for our proposed method, in which the number of variables is allowed to be much larger than the sample size. Numerical results indicate that the proposed method works well in both prediction and variable selection. We also applied our method to the American Cancer Society breast cancer survivor dataset. The analysis results are clinically meaningful and have potential impact on interventions to improve the quality of life of breast cancer survivors.

The grouping structure we have considered in this section does not have overlaps. However, it is not unusual for a variable to be a member of several groups. For example, given some biologically defined gene sets, say pathways, not surprisingly, there will be considerable overlaps among these sets. Our LES penalty can be modified for variable selection when the groups have overlaps. With a little change of notation, the p variables are denoted by $X_1, \ldots, X_p$ and their corresponding regression coefficients are $\beta_1, \ldots, \beta_p$. Let $V_k \subseteq \{1, 2, \ldots, p\}$ be the set of indices of variables in the kth group. We consider the following optimization problem:

$$\frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j\Big)^2 + \lambda \sum_{k=1}^{K} w_k \log\Big\{\sum_{j\in V_k}\exp(\alpha m_j|\beta_j|)\Big\}, \tag{4.26}$$

where $w_k,\ k = 1, \ldots, K$ are weights to adjust for possible different size of each group, say, taking $w_k = p_k/p$, and $m_j,\ j = 1, \ldots, p$ are weights to adjust for possible different frequency of each variable included in the penalty term, say taking $m_j = 1/n_j$, where n_j is the number of groups which include the variable X_j. It is easy to see that the objective function (4.26) reduces to the objective function with the original LES penalty (8) when there is no overlap among the K groups.

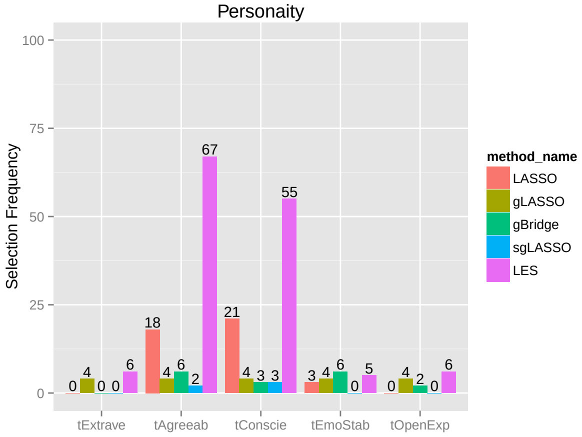

Figure 4.1: Within Group Selection Results for Personality Group

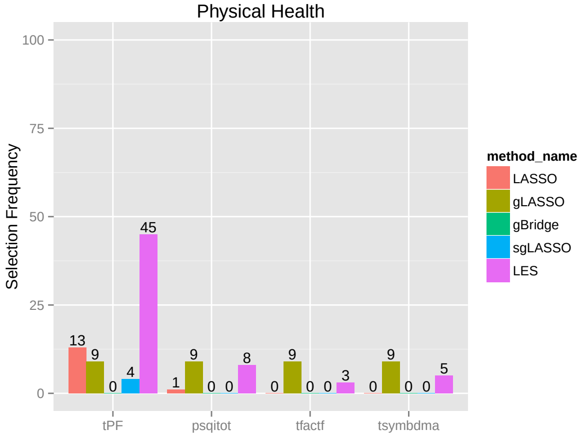

Figure 4.2: Within Group Selection Results for Physical Health Group

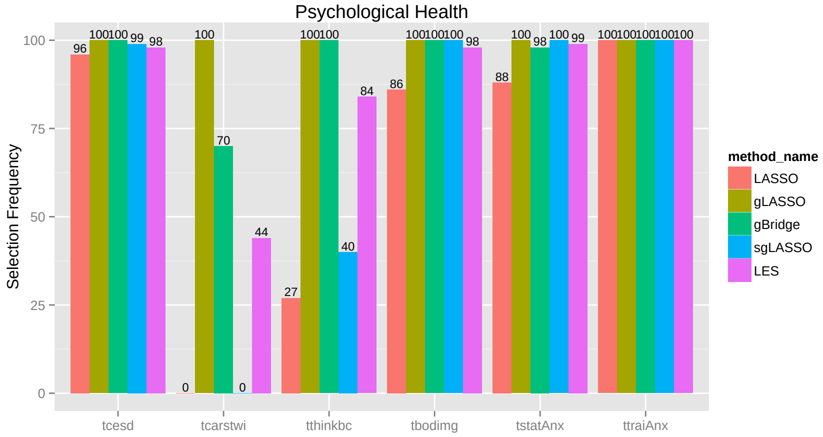

Figure 4.3: Within Group Selection Results for Psychological Health Group

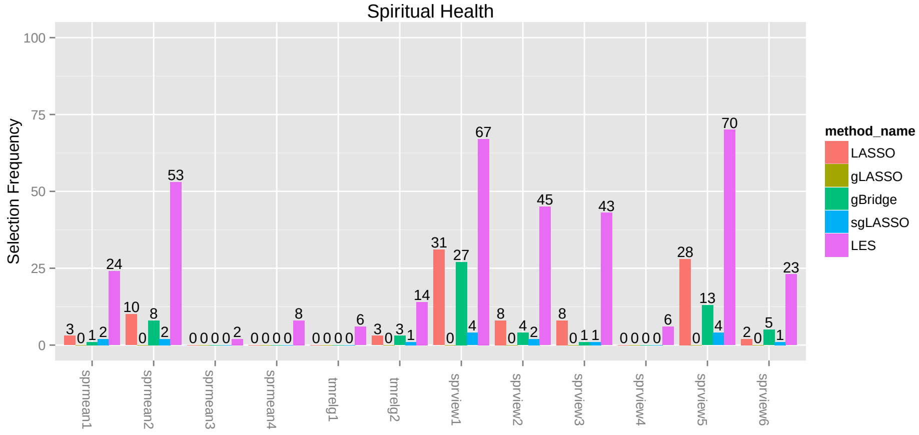

Figure 4.4: Within Group Selection Results for Spiritual Health Group

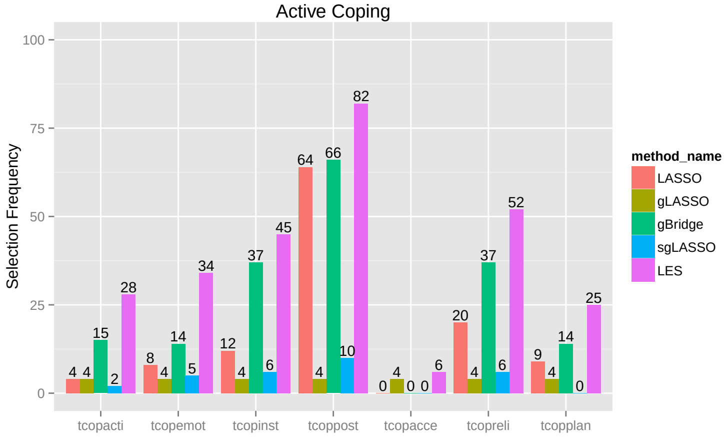

Figure 4.5: Within Group Selection Results for Active Coping Group

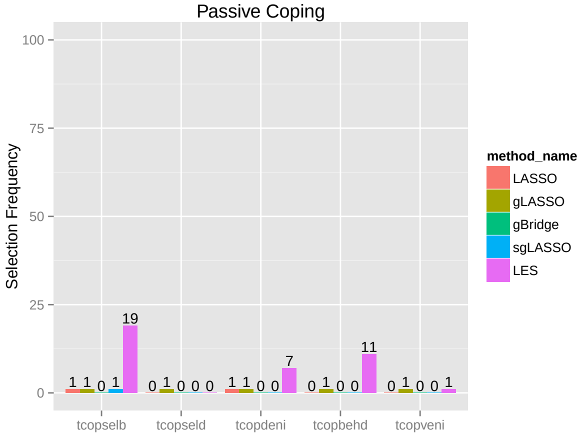

Figure 4.6: Within Group Selection Results for Passive Coping Group

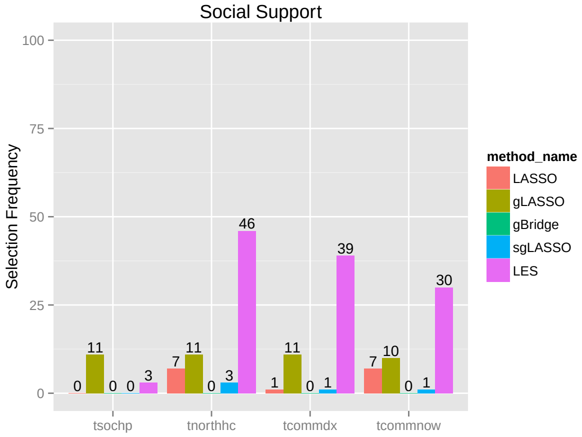

Figure 4.7: Within Group Selection Results for Social Support Group

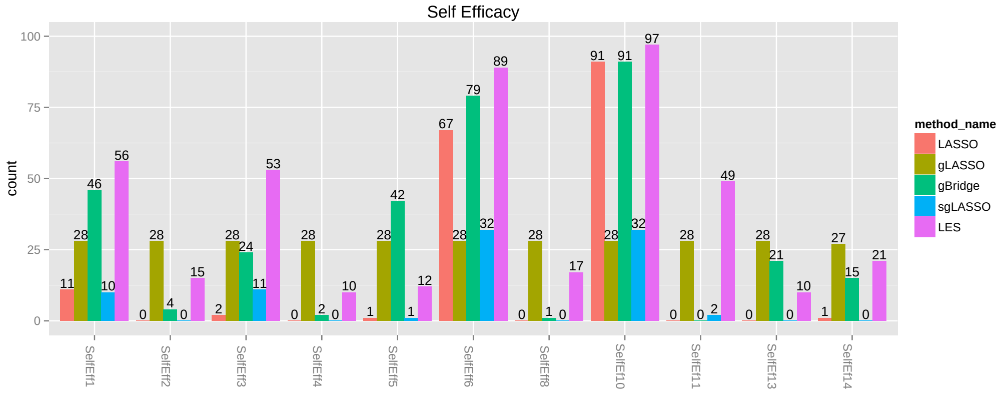

Figure 4.8: Within Group Selection Results for Self Efficacy Group

Table 4.1: Summary of simulation results over 1,000 replicates. “1-Sens” means one minus the sensitivity of variable selection; “1-Spec” means one minus the specificity of variable selection; “ME” means the model error; “Bias” means the bias of the estimator, which is defined as $||\hat{\beta} - \beta^*||^2$. “AUC” means the area under ROC curve of sensitivities and specificities of variable selection across different tuning parameter values. The numbers in parentheses are the corresponding standard errors. The bold numbers are significantly better than others at a significance level of 0.05.

	Tuning Set Tuning				BIC Tuning				AUC
Method	1-Sens	1-Spec	ME	Bias	1-Sens	1-Spec	ME	Bias	
Simulation 1									
LASSO	0.000	0.418	1.022	1.022	0.000	0.136	1.318	1.318	1.000
	(0.000)	(0.006)	(0.016)	(0.016)	(0.000)	(0.004)	(0.021)	(0.021)	(0.000)
gLASSO	0.000	0.608	0.717	0.717	0.000	0.061	0.856	0.856	1.000
	(0.000)	(0.009)	(0.012)	(0.012)	(0.000)	(0.004)	(0.015)	(0.015)	(0.000)
gBrdige	0.000	**0.057**	**0.543**	**0.543**	0.000	0.092	**0.641**	**0.641**	1.000
	(0.000)	(0.004)	(0.010)	(0.010)	(0.000)	(0.003)	(0.011)	(0.011)	(0.000)
sgLASSO	0.000	0.612	0.811	0.811	0.000	**0.028**	1.050	1.050	1.000
	(0.000)	(0.009)	(0.013)	(0.013)	(0.000)	(0.003)	(0.018)	(0.018)	(0.000)
LES	0.000	0.537	**0.544**	**0.544**	0.000	0.282	0.770	0.770	1.000
	(0.000)	(0.006)	(0.010)	(0.010)	(0.000)	(0.006)	(0.016)	(0.016)	(0.000)
Simulation 2									
LASSO	0.006	0.379	2.289	3.902	0.013	0.065	2.682	3.746	0.995
	(0.001)	(0.006)	(0.035)	(0.072)	(0.001)	(0.003)	(0.045)	(0.069)	(0.001)
gLASSO	0.000	0.846	3.463	6.117	**0.000**	0.349	4.262	4.845	0.895
	(0.000)	(0.007)	(0.046)	(0.096)	(0.000)	(0.006)	(0.062)	(0.070)	(0.000)
gBrdige	0.004	**0.146**	2.061	3.712	0.005	0.062	**2.321**	3.600	0.997
	(0.001)	(0.004)	(0.032)	(0.072)	(0.001)	(0.002)	(0.040)	(0.072)	(0.001)
sgLASSO	0.000	0.470	2.030	2.333	0.003	0.058	2.608	2.865	1.000
	(0.000)	(0.008)	(0.031)	(0.043)	(0.001)	(0.003)	(0.045)	(0.057)	(0.000)
LES	0.001	0.466	**1.931**	2.344	0.006	0.169	2.629	**2.426**	1.000
	(0.000)	(0.009)	(0.031)	(0.045)	(0.001)	(0.006)	(0.047)	(0.054)	(0.000)
Simulation 3									
LASSO	0.101	0.410	4.158	8.303	0.145	0.146	4.886	8.254	0.914
	(0.002)	(0.007)	(0.046)	(0.094)	(0.003)	(0.005)	(0.066)	(0.096)	(0.002)
gLASSO	**0.000**	0.975	6.018	13.040	**0.000**	0.761	7.011	11.585	0.839
	(0.000)	(0.003)	(0.063)	(0.149)	(0.000)	(0.007)	(0.082)	(0.129)	(0.003)
gBrdige	0.100	**0.399**	4.337	8.407	0.135	**0.089**	5.559	8.347	0.932
	(0.002)	(0.006)	(0.048)	(0.097)	(0.002)	(0.003)	(0.068)	(0.092)	(0.002)
sgLASSO	0.030	0.673	3.563	4.759	0.105	0.191	4.738	7.282	0.994
	(0.001)	(0.007)	(0.044)	(0.062)	(0.002)	(0.005)	(0.060)	(0.095)	(0.000)
LES	0.028	0.642	**3.295**	4.933	0.051	0.365	**4.459**	**5.000**	**0.999**
	(0.002)	(0.008)	(0.041)	(0.067)	(0.003)	(0.009)	(0.063)	(0.076)	(0.000)
Simulation 4									
LASSO	0.127	0.243	5.539	9.081	0.174	0.080	6.830	9.067	0.899
	(0.003)	(0.004)	(0.060)	(0.097)	(0.003)	(0.003)	(0.091)	(0.092)	(0.002)
gLASSO	**0.000**	0.910	9.460	16.780	**0.002**	0.515	11.573	13.795	0.881
	(0.000)	(0.005)	(0.089)	(0.170)	(0.001)	(0.006)	(0.150)	(0.140)	(0.001)
gBrdige	0.118	**0.138**	5.058	8.998	0.147	**0.059**	6.151	8.777	0.913
	(0.003)	(0.003)	(0.059)	(0.112)	(0.003)	(0.002)	(0.074)	(0.109)	(0.002)
sgLASSO	0.023	0.445	4.830	**4.517**	0.107	0.076	6.996	7.028	0.991
	(0.001)	(0.007)	(0.045)	(0.047)	(0.003)	(0.002)	(0.084)	(0.074)	(0.000)
LES	0.028	0.472	**4.638**	5.559	0.034	0.245	6.284	**5.273**	0.992
	(0.002)	(0.008)	(0.054)	0.069	(0.002)	(0.006)	(0.086)	(0.071)	(0.000)

Table 4.2: Summary of ACS breast cancer survivor data analysis results. Results are based on 100 random splits. "Variable selection" reports the average number of selected individual variables; "Group selection" reports the average number of selected groups and "MSE" reports the average mean square errors on test sets. The numbers in parentheses are the corresponding standard errors.

Selection Frequency and Mean Square Error						
	10-fold CV Tuning			BIC Tuning		
	Variable selection	Group selection	MSE	Variable selection	Group selection	MSE
LASSO	23.18	8.76	2.6288	8.59	3.97	2.7949
	(0.53)	(0.14)	(0.0286)	(0.33)	(0.15)	(0.0309)
gLASSO	51.32	8.80	2.6484	10.46	1.64	2.8620
	(0.42)	(0.12)	(0.0286)	(0.64)	(0.09)	(0.0307)
gBrdige	16.24	3.34	2.6239	11.56	2.94	2.7548
	(0.64)	(0.13)	(0.0293)	(0.26)	(0.07)	(0.0356)
sgLASSO	25.83	8.58	2.6221	5.89	1.59	2.8765
	(0.60)	(0.18)	(0.0265)	(0.31)	(0.11)	(0.0280)
LES	33.50	9.58	**2.6072**	19.86	6.37	**2.7026**
	(0.74)	(0.11)	(0.0283)	(0.91)	(0.20)	(0.0298)

Individual Group Selection Frequency						
10-fold CV Tuning						
Group Name	Agediag	Bcmths	Yrseduc	Personaity	Physical Health	Psychological Health
LASSO	73	55	19	96	75	100
gLASSO	82	66	19	94	82	100
gBrdige	4	2	0	30	2	100
sgLASSO	80	56	18	91	71	100
LES	91	71	21	97	89	100
Group Name	Spiritual Health	Active Coping	Passive Coping	Social Support	Self Efficacy	
LASSO	100	100	68	90	100	
gLASSO	98	100	46	93	100	
gBrdige	35	70	0	3	88	
sgLASSO	97	98	60	89	98	
LES	100	100	86	97	100	
BIC Tuning						
Group Name	Agediag	Bcmths	Yrseduc	Personaity	Physical Health	Psychological Health
LASSO	5	1	0	32	14	100
gLASSO	2	1	4	4	9	100
gBrdige	1	0	0	6	0	100
sgLASSO	1	0	0	3	4	100
LES	28	14	3	75	48	100
Group Name	Spiritual Health	Active Coping	Passive Coping	Social Support	Self Efficacy	
LASSO	63	69	2	14	97	
gLASSO	0	4	1	11	28	
gBrdige	29	67	0	0	91	
sgLASSO	5	10	1	3	32	
LES	91	89	31	59	99	

5 CONCLUDING REMARKS

In this thesis, we study the variable selection problem under the penalized least squares regression setting. Two novel penalties, K-Smallest Items (KSI) penalty and Self-adaptive penalty (SAP), are proposed for variable selection; and one penalty, Log-Exp-Sum (LES) penalty, is proposed for group variable selection.

Both the KSI penalty and SAP are motivated to solve the estimation bias reduction problem. They are nonconvex penalties, and they place relatively small or even no penalty on important predictors with large estimated coefficients. We fully investigate the theoretical properties of these two penalties, and show that under desirable conditions, they both possess the weak oracle property. If stronger assumptions are assumed, we show that KSI penalty possesses the oracle property. Simulation examples and real data analysis demonstrates the goodness of these two penalties. Moreover, the idea of KSI penalty can be applied to many existing penalties, including LASSO, SCAD and MCP. Therefore, we have a large collection of penalties which belong to the KSI penalties family. By using the idea of KSI penalty, we introduce flexibility into those existing penalties. We could further extend the idea of KSI to handle group variable selection problem. For example, in the case of group lasso penalty, we may choose to penalize on the K-smallest L_2-norm of the group coefficients.

The LES penalty is motivated to solve the group variable selection problem when group structures exist among covariates. It is a convex penalty and it can select the important group as well as the important variables within the selected groups. We have established the non-asymptotic error bounds and asymptotic group selection consistency for LES penalty. Simulation examples and real data analysis indicate that the proposed method works well in terms of both variable selection and prediction accuracy. Although the LES penalty is proposed when there are no overlaps among different group, it can be easily modified to handle the case when the groups have overlaps. It is also worth noting that, we apply the randomized trace method to numerically estimated the degrees of freedom of an estimated model. In many variable selection and group variable selection problems, the model degrees of freedom are not very clear. Sometimes people simply use the number of nonzero estimated coefficients as a proxy for the estimated degrees of freedom. The randomized trace method provide an alternative to estimate the degrees of freedom numerically.

Proof of Proposition 2.1

Proof. We first assume $p = 2$, $K = 1$, and $0 \leqslant u_2 \leqslant u_1$. Then (2.6) can be simplify as:

$$\min_{\beta_1,\beta_2} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_{[2]}|). \tag{A.1}$$

We want to show, the $\beta_{[2]}$ in (A.1) is indeed β_2, i.e., the following inequality is satisfied:

$$\begin{aligned} & \min_{\beta_1 \geqslant \beta_2} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_2|) && \text{(A.2)} \\ \leqslant \ & \min_{\beta_2 \geqslant \beta_1} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_1|). && \text{(A.3)} \end{aligned}$$

Because $p_\lambda(\cdot)$ satisfies condition $(C1)$, $p'_\lambda(t) \geqslant 0$ for $t \in [0, \infty)$. The solution $\hat{\beta}_2$ to the following optimization problem:

$$\hat{\beta}_2 = \arg\min_{\beta_2} \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_2|), \tag{A.4}$$

satisfies $0 \leqslant \hat{\beta}_2 \leqslant u_2 \leqslant u_1$. Therefore, $\beta_1 = u_1$ and $\beta_2 = \hat{\beta}_2$ is the unique minimizers to the optimization problem (A.2) and we have:

$$\min_{\beta_1 \geqslant \beta_2} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_2|) = \min_{\beta_2} \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_2|). \tag{A.5}$$

Here $\beta_1 = u_1$ and $\beta_2 = \hat{\beta}_2$ is exactly what step 3 and step 4 in the algorithm will produce. Next we show the validity of (A.2) being less than (A.3).

Because $p'_\lambda(t) \geqslant 0$ for $t \in [0, \infty)$, and $0 \leqslant u_2 \leqslant u_1$, it is easy to verify that:

$$\min_{\beta_2} \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_2|) \leqslant \min_{\beta_1} \frac{1}{2}(\beta_1 - u_1)^2 + \lambda p_\lambda(|\beta_1|). \tag{A.6}$$

So we have:

$$\begin{aligned}
& \min_{\beta_1} \frac{1}{2}(\beta_1 - u_1)^2 + \lambda p_\lambda(|\beta_1|) \\
\leqslant\ & \min_{\beta_1,\beta_2} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_1|) \\
\leqslant\ & \min_{\beta_2 \geqslant \beta_1} \frac{1}{2}(\beta_1 - u_1)^2 + \frac{1}{2}(\beta_2 - u_2)^2 + \lambda p_\lambda(|\beta_1|), \qquad \text{(A.7)}
\end{aligned}$$

where the first inequality is because the term $\frac{1}{2}(\beta_2 - u_2)^2$ is positive and the second inequality is because unconstrained minimizer is smaller than the constrained minimizer.

Combining (A.5), (A.6) and (A.7), we show (A.2) is indeed less than (A.3).

In the more general case, we denote set $\mathscr{A} = \{1, 2, \ldots, p-K\}$ and set $\mathscr{B} = \{p-K+1, \ldots, p\}$. We assume $\mathbf{u}$ satisfies $|u_{j_1}| \geqslant |u_{j_2}|, \forall j_1 \in \mathscr{A}$, and $j_2 \in \mathscr{B}$. We want to show,

$$\begin{aligned}
& \min_{\beta_j} \frac{1}{2}\sum_{j=1}^{p}(\beta_j - u_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|) \qquad \text{(A.8)} \\
=\ & \min_{\Omega} \frac{1}{2}\sum_{j=1}^{p}(\beta_j - u_j)^2 + \lambda \sum_{j=p-K+1}^{p} p_\lambda(|\beta_j|). \qquad \text{(A.9)}
\end{aligned}$$

Here Ω is a constraint such that $\Omega = \{\beta_j \,\big|\, |\beta_{j_1}| \geqslant |\beta_{j_2}|, \forall j_1, j_2, \text{ s.t. } j_1 \in \mathscr{A}, \text{ and } j_2 \in \mathscr{B}\}$. Under the constraint Ω, we have $\sum_{j=p-K+1}^{p} p_\lambda(|\beta_{[j]}|) = \sum_{j=p-K+1}^{p} p_\lambda(|\beta_j|)$. Therefore, there is no order constraint in (A.9). It is easy to see that if we follow Step 3 and Step 4 in the algorithm, we will solve the minimization problem (A.9).

Now let's assume constraint Ω is violated. Then there must be indices $\tilde{j_1} \in \mathscr{A}$ and $\tilde{j_2} \in \mathscr{B}$, such that the order constraint is not valid, i.e. $|\beta_{\tilde{j}1}| < |\beta_{\tilde{j}2}|$. By what we showed earlier, we have:

$$\min_{|\beta_{\tilde{j}_1}|\geqslant|\beta_{\tilde{j}_2}|}\frac{1}{2}(\beta_{\tilde{j}_1}-u_{\tilde{j}_1})^2+\frac{1}{2}(\beta_{\tilde{j}_2}-u_{\tilde{j}_2})^2+\lambda p_\lambda(|\beta_{\tilde{j}_2}|) \tag{A.10}$$

$$\leqslant \min_{|\beta_{\tilde{j}_2}|>|\beta_{\tilde{j}_1}|}\frac{1}{2}(\beta_{\tilde{j}_1}-u_{\tilde{j}_1})^2+\frac{1}{2}(\beta_{\tilde{j}_2}-u_{\tilde{j}_2})^2+\lambda p_\lambda(|\beta_{\tilde{j}_1}|). \tag{A.11}$$

The above inequality implies if constrain Ω is violated, we will always obtain a larger objective function value. Therefore, the equality between (A.8) and (A.9) is valid. This completes the proof. □

Proof of Proposition 2.2

Proof. We first prove part (a).

By the conditions in part (a), we have $Q(\hat{\boldsymbol{\beta}})=C(\hat{\boldsymbol{\beta}})$ for a small area centered at $\hat{\boldsymbol{\beta}}$. For $\forall\,\boldsymbol{\beta}$ within this area, we have:

$$\begin{aligned}
Q(\hat{\boldsymbol{\beta}}) &= \frac{1}{2n}\|\mathbf{y}-X\hat{\boldsymbol{\beta}}\|_2^2+\lambda\sum_{j=p-K+1}^{p}p_\lambda(|\hat{\beta}_{[j]}|)\\
&\leqslant \frac{1}{2n}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=p-K+1}^{p}p_\lambda(|\beta_{[j]}|)\\
&\leqslant \frac{1}{2n}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=p-K+1}^{p}p_\lambda(|\beta_{j}|)\\
&= C(\boldsymbol{\beta}).
\end{aligned} \tag{A.12}$$

So $\hat{\boldsymbol{\beta}}$ is a local minimizer of $C(\boldsymbol{\beta})$.

We now prove part (b).

Consider a small open ball with center $\hat{\boldsymbol{\beta}}$ and radius $r=\min\frac{1}{2}\{|\hat{\beta}_h|-|\hat{\beta}_i|\Big|1\leqslant h\leqslant p-K, p-K+1\leqslant i\leqslant p\}$. For $\forall\,\boldsymbol{\beta}$ within this open ball, the first K covariates of $\boldsymbol{\beta}$ in absolute value are strictly greater than the last $p-K$ covariates of $\boldsymbol{\beta}$ in absolute value. So $\sum_{j=p-K+1}^{K}p_\lambda(|\beta_{[j]}|)=\sum_{j=p-K+1}^{K}p_\lambda(|\beta_j|)$, which implies within this open ball, $Q(\boldsymbol{\beta})$ and $C(\boldsymbol{\beta})$ have the same form. Therefore, the local minimizer of $C(\boldsymbol{\beta})$ is also the local minimizer of $Q(\boldsymbol{\beta})$. □

Proof of Theorem 2.3

Proof. We will first derive the necessary condition.

Suppose $\hat{\boldsymbol{\beta}}$ is a local minimizer of $C(\hat{\boldsymbol{\beta}})$, then $\exists$ $\mathbf{v} = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$, such that:

$$\nabla C(\hat{\boldsymbol{\beta}}) = -\frac{1}{n} X^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + \lambda \mathbf{v} = 0, \tag{A.13}$$

where $v_j = 0$ for $1 \leqslant j \leqslant p - K$; $v_j = p_\lambda'(|\hat{\beta}_j|)sgn(\hat{\beta}_j)$ for $p - K + 1 \leqslant j \leqslant s$; and $v_j \in [-p_\lambda'(0+),\ p_\lambda'(0+)]$ for $s + 1 \leqslant j \leqslant p$.

It is easy to see equation (A.13) can be equivalently written as:

$$\frac{1}{n} X_1^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1) - \nabla J_\lambda(\hat{\boldsymbol{\beta}}_1) = 0; \tag{A.14}$$

$$\frac{1}{n} ||X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)||_\infty \leqslant \lambda p_\lambda'(0+). \tag{A.15}$$

Note that $\hat{\boldsymbol{\beta}}$ is also a local minimizer of $C(\boldsymbol{\beta})$ constrained on the s-dimension subspace $\mathbb{S} = \{\boldsymbol{\beta} \in \mathbb{R}^p \big| \beta_j = 0,\ s+1 \leqslant j \leqslant p\}$. It follows from the second order condition that $\frac{1}{n} X_1^T X_1 + \nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_1)$ is positive semi-definite. So,

$$\frac{1}{n} X_1^T X_1 + \nabla^2 J_\lambda(\hat{\boldsymbol{\beta}}_1) \geqslant 0. \tag{A.16}$$

Next, we prove the sufficient condition.

We first constrain $C(\boldsymbol{\beta})$ on the s-dimension subspace $\mathbb{S} = \{\boldsymbol{\beta} \in \mathbb{R}^p \big| \beta_j = 0,\ s + 1 \leqslant j \leqslant p\}$ of $\mathbb{R}^p$. Consider a neighborhood $\mathbb{B}_0$ in the subspace $\mathbb{S}$ centered at $\hat{\boldsymbol{\beta}}$, conditions (2.9) and (2.11) entail that $\hat{\boldsymbol{\beta}}$ is the unique local minimizer of $C(\boldsymbol{\beta})$ in $\mathbb{B}_0 \subset \mathbb{R}^s$.

It remains to show that $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $C(\boldsymbol{\beta})$ in $\mathbb{R}^p$. Consider a sufficiently small ball $\mathbb{B}_1 \subset \mathbb{R}^p$ centered at $\hat{\boldsymbol{\beta}}$, such that $\mathbb{B}_1 \cap \mathbb{S} \subset \mathbb{B}_0$. For any

$\boldsymbol{\eta}^1 \in \mathbb{B}_1 \setminus \mathbb{B}_0$, let $\boldsymbol{\eta}^2$ be the projection of $\boldsymbol{\eta}^1$ onto the subspace $\mathbb{S}$. Then we have $\boldsymbol{\eta}^2 \in \mathbb{S}$, which entails that $C(\boldsymbol{\eta}^2) > C(\hat{\boldsymbol{\beta}})$ if $\boldsymbol{\eta}^2 \neq \hat{\boldsymbol{\beta}}$. Thus, it suffices to show that $C(\boldsymbol{\eta}^1) > C(\boldsymbol{\eta}^2)$.

By the mean value theorem, we have:

$$C(\boldsymbol{\eta}^1) - C(\boldsymbol{\eta}^2) = \nabla C(\boldsymbol{\eta}^0)^T(\boldsymbol{\eta}^1 - \boldsymbol{\eta}^2), \tag{A.17}$$

where $\boldsymbol{\eta}^0 = \boldsymbol{\eta}^2 + c(\boldsymbol{\eta}^1 - \boldsymbol{\eta}^2)$ for some $c \in (0, 1)$.

Let η_j^1 and η_j^2 be the j-th component of $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$, respectively. Then we have:

$$\eta_j^1 = \eta_j^2 \quad \text{for } 1 \leqslant j \leqslant s; \tag{A.18}$$

$$\eta_j^2 = 0 \quad \text{for } s+1 \leqslant j \leqslant p. \tag{A.19}$$

Therefore, $\eta_j^0 = \eta_j^1$, for $1 \leqslant j \leqslant s$; and $\eta_j^0 = c\eta_j^1$, for $s+1 \leqslant j \leqslant p$.

The right hand side of (A.17) can then be expressed as:

$$-\frac{1}{n}[X_2^T(\mathbf{y} - X\boldsymbol{\eta}^0)]^T\boldsymbol{\eta}_2^1 + \lambda \sum_{j=s+1}^{p} p'_\lambda(|\eta_j^0|)|\eta_j^1|, \tag{A.20}$$

where $\boldsymbol{\eta}_2^1$ is a subvector of $\boldsymbol{\eta}^1$ formed by the last $p - s$ components of $\boldsymbol{\eta}^1$. By $\boldsymbol{\eta}^1 \in \mathbb{B}_1 \setminus \mathbb{B}_0$, we have $\boldsymbol{\eta}_2^1 \neq \mathbf{0}$.

It follows from concavity of $p_\lambda(\cdot)$ in condition $(C1)$ that $p'_\lambda(t)$ is decreasing in $t \in [0, \infty)$. By condition (2.10) and the continuity of $p'_\lambda(t)$, there exists some $\delta > 0$ such that for any $\boldsymbol{\beta}$ in a open ball in $\mathbb{R}^p$ centered at $\hat{\boldsymbol{\beta}}$ with radius δ, the following inequality holds:

$$||\frac{1}{n}X_2^T(\mathbf{y} - X\boldsymbol{\beta})||_\infty < \lambda p'_\lambda(\delta). \tag{A.21}$$

We further shrink the radius of the ball $\mathbb{B}_1$ to less than δ so that $|\eta_j^1| < \delta$ for $s+1 \leqslant j \leqslant p$ and (A.21) holds for any $\boldsymbol{\beta} \in \mathbb{B}_1$. Since $\boldsymbol{\eta}^0 \in \mathbb{B}_1$, it follows from (A.21) that the term (A.20) is strictly greater than:

$$-\lambda p'_\lambda(\delta)|\boldsymbol{\eta}_2^1| + \lambda p'_\lambda(\delta)|\boldsymbol{\eta}_2^1| = 0, \tag{A.22}$$

where the monotonicity of $p'_\lambda(\cdot)$ was used in the second term. Thus we conclude $C(\boldsymbol{\eta}^1) > C(\boldsymbol{\eta}^2)$. This completes the proof. □

Proof of Proposition 2.4

Proof. Conditions (2.9) - (2.11) are sufficient to ensure that $\hat{\boldsymbol{\beta}}$ is a strict local minimizer to $C(\boldsymbol{\beta})$. By part (b) in Proposition 2.2, we know $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q(\boldsymbol{\beta})$. On the other hand, the local minimizer of $Q(\boldsymbol{\beta})$ is also a local minimizer of its un-ordered counterpart $C(\boldsymbol{\beta})$, therefore it must satisfy the necessary conditions of $C(\boldsymbol{\beta})$. □

Proof of Theorem 2.5

Proof. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T = X^T(\mathbf{y} - X\boldsymbol{\beta}^*) = X^T\boldsymbol{\epsilon}$, consider event sets

$$\mathscr{A}_1 = \{||\boldsymbol{\xi}_S||_\infty \leqslant \sqrt{2\sigma^2 n \log n}\} \quad \text{and} \quad \mathscr{A}_2 = \{||\boldsymbol{\xi}_{S^c}||_\infty \leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}\}, \tag{A.23}$$

where index set $S = \{j\,|\,1 \leqslant j \leqslant s\}$ and index set $S^c = \{j\,|\,s+1 \leqslant j \leqslant p\}$. Notice that the diagonal elements of $\frac{X^TX}{n}$ is 1, so $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n x_{ij}\epsilon_i$ satisfies standard normal distribution. By normal distribution tail inequality, for $n \geqslant 2$, we have:

$$\begin{aligned} P(\mathscr{A}_1^c) &\leqslant \sum_{j=1}^s P(|\xi_j| > \sqrt{2\sigma^2 n \log n}) = \sum_{j=1}^s P(\frac{1}{\sigma\sqrt{n}}|\xi_j| > \sqrt{2\log n}) \\ &\leqslant \sum_{j=1}^s \frac{1}{n\sqrt{\pi \log n}} \leqslant \frac{s}{n}, \end{aligned} \tag{A.24}$$

$$\begin{aligned} P(\mathscr{A}_2^c) &\leqslant \sum_{j=s+1}^{p} P(|\xi_j| > n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}) = \sum_{j=s+1}^{p} P(\frac{1}{\sigma\sqrt{n}}|\xi_j| > n^{\frac{1}{2}-t}\sqrt{2\log n}) \\ &\leqslant \sum_{j=s+1}^{p} \frac{e^{-n^{1-2t}\log n}}{n^{\frac{1}{2}-t}\sqrt{\pi \log n}} \leqslant (p-s)e^{-n^{1-2t}\log n}. \end{aligned} \tag{A.25}$$

By (A.24) and (A.25), we know $P(\mathscr{A}_1 \cap \mathscr{A}_2) \geqslant 1 - (sn^{-1} + (p-s)e^{-n^{1-2t}\log n})$. Under the intersection of $\mathscr{A}_1 \cap \mathscr{A}_2$, we will show that there exists an solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ to (2.9)-(2.11) such that $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^*)$ and $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_\infty = O(n^{-\gamma}\log n)$, where the function *sign* applies componentwise.

Step 1: Existence of a solution to equation (2.9). We first show for sufficiently large n, equation (2.9) has a solution $\hat{\boldsymbol{\beta}}_1$ inside the hyper cube:

$$\mathscr{N} = \{\boldsymbol{\delta} \in \mathbb{R}^s : ||\boldsymbol{\delta} - \boldsymbol{\beta}_1^*||_\infty = n^{-\gamma}\log n\}. \tag{A.26}$$

For $\forall\, \boldsymbol{\delta} \in \mathscr{N}$, since $d_n > n^{-\gamma}\log n$, we have:

$$\min_{j\in S} |\delta_j| \geqslant \min_{j\in S} |\beta_j^*| - n^{-\gamma}\log n \geqslant d_n, \tag{A.27}$$

and $sign(\boldsymbol{\delta}) = sign(\boldsymbol{\beta}_2^*)$.

By the monotonicity condition of $p'_{\lambda_n}(t)$ and (A.27), we have:

$$||\nabla J_{\lambda_n}(\boldsymbol{\delta})||_\infty \leqslant \lambda_n p'_{\lambda_n}(d_n). \tag{A.28}$$

By inequality (A.28) and the definition of $\boldsymbol{\xi}_S$, we have:

$$||\boldsymbol{\xi}_S - n\nabla J_{\lambda_n}(\boldsymbol{\delta})||_\infty \leqslant \sqrt{2\sigma^2 n \log n} + n\lambda_n p'_{\lambda_n}(d_n). \tag{A.29}$$

Define $\Phi(\boldsymbol{\delta})$ which corresponds to n times the left hand side of in equality (2.9):

$$\begin{aligned} \Phi(\boldsymbol{\delta}) &\triangleq X_1^T \mathbf{y} - X_1^T X_1 \boldsymbol{\delta} - n \nabla J_{\lambda_n}(\boldsymbol{\delta}) \\ &= X_1^T(\mathbf{y} - X_1 \boldsymbol{\beta}_1^*) - X_1^T X_1(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*) - n \nabla J_{\lambda_n}(\boldsymbol{\delta}) \\ &= (\boldsymbol{\xi}_S - n \nabla J_{\lambda_n}(\boldsymbol{\delta})) - X_1^T X_1(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*). \end{aligned} \tag{A.30}$$

Let

$$\widetilde{\Phi}(\boldsymbol{\delta}) \triangleq (X_1^T X_1)^{-1} \Phi(\boldsymbol{\delta}) = \mathbf{u} - (\boldsymbol{\delta} - \boldsymbol{\beta}_1^*), \tag{A.31}$$

where $\mathbf{u} = (X_1^T X_1)^{-1}(\boldsymbol{\xi}_S - n \nabla J_{\lambda_n}(\boldsymbol{\delta}))$. By assumptions $(C4)$ and $(C5)$:

$$\begin{aligned} ||\mathbf{u}||_\infty &\leqslant ||(X_1^T X_1)^{-1}||_\infty (||\boldsymbol{\xi}_S||_\infty + n ||\nabla J_{\lambda_n}(\boldsymbol{\delta})||_\infty) \\ &= o(n^{-\gamma} \log n). \end{aligned} \tag{A.32}$$

So for sufficiently large n, if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*)_j = n^{-\gamma} \log n$, we have:

$$\widetilde{\Phi}_j(\boldsymbol{\delta}) \leqslant ||\mathbf{u}||_\infty - n^{-\gamma} \log n \leqslant 0, \tag{A.33}$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*)_j = -n^{-\gamma} \log n$, we have:

$$\widetilde{\Phi}_j(\boldsymbol{\delta}) \geqslant -||\mathbf{u}||_\infty + n^{-\gamma} \log n \geqslant 0, \tag{A.34}$$

where $\widetilde{\Phi}_j(\boldsymbol{\delta})$ is the jth covariate of $\widetilde{\Phi}(\boldsymbol{\delta})$. By the continuity of $\widetilde{\Phi}(\boldsymbol{\delta})$, (A.33) and (A.34), Miranda's existence theorem guarantees that equation $\widetilde{\Phi}(\boldsymbol{\delta}) = 0$ has a solution in $\mathscr{N}$, which we denote as $\hat{\boldsymbol{\beta}}_1$. It is easy to see, $\hat{\boldsymbol{\beta}}_1$ also solves $\Phi(\boldsymbol{\delta}) = 0$. Thus, equation (2.9) does have a solution inside the hypercube $\mathscr{N}$.

Step 2: Verification of Strict Inequality (2.10).

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T \in \mathbb{R}^p$, with $\hat{\boldsymbol{\beta}}_1 \in \mathscr{N}$ being a solution to equation (2.9). We want to show that $\hat{\boldsymbol{\beta}}$ satisfies inequality (2.10), i.e.:

$$\frac{1}{n} ||X_2^T(\mathbf{y} - X_1 \hat{\boldsymbol{\beta}}_1)||_\infty < \lambda_n p'_{\lambda_n}(0+). \tag{A.35}$$

Notice that

$$\begin{aligned} & \frac{1}{n}||X_2^T(\mathbf{y}-X_1\hat{\boldsymbol{\beta}}_1)||_\infty \\ = \; & \frac{1}{n}\Big|\Big|X_2^T(\mathbf{y}-X_1\boldsymbol{\beta}_1^*)-X_2^TX_1(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*)\Big|\Big|_\infty \\ \leqslant \; & \frac{1}{n}||\boldsymbol{\xi}_{S^c}||_\infty+\frac{1}{n}||X_2^TX_1(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*)||_\infty \\ \triangleq \; & (I)+(II). \end{aligned} \tag{A.36}$$

Because $||\boldsymbol{\xi}_{S^c}||_\infty \leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n\log n}$ and $n^{-t}\sqrt{\log n}/\lambda_n = o(1)$, we have:

$$(I) = O(n^{-t}\sqrt{\log n}) = o(\lambda_n). \tag{A.37}$$

Because $\hat{\boldsymbol{\beta}}_1$ is a solution to $\Phi(\boldsymbol{\delta}) = 0$, we have:

$$\hat{\boldsymbol{\beta}_1}-\boldsymbol{\beta}_1^* = (X_1^TX_1)^{-1}(\boldsymbol{\xi}_S - n\nabla J_{\lambda_n}(\hat{\boldsymbol{\beta}}_1)). \tag{A.38}$$

Therefore,

$$\begin{aligned} (II) &= \frac{1}{n}||X_2^TX_1(X_1^TX_1)^{-1}(\boldsymbol{\xi}_S-n\nabla J_{\lambda_n}(\hat{\boldsymbol{\beta}}_1))||_\infty \\ &\leqslant \frac{1}{n}||X_2^TX_1(X_1^TX_1)^{-1}||_\infty||(\boldsymbol{\xi}_S-n\nabla J_{\lambda_n}(\hat{\boldsymbol{\beta}}_1))||_\infty \\ &\leqslant \frac{1}{n}||X_2^TX_1(X_1^TX_1)^{-1}||_\infty\Big(||\boldsymbol{\xi}_S||_\infty+||n\nabla J_{\lambda_n}(\hat{\boldsymbol{\beta}}_1)||_\infty\Big) \\ &= (III)+(IV). \end{aligned} \tag{A.39}$$

By conditions $C(3)$ and $C(5)$,

$$\begin{aligned} (III) &= \frac{1}{n}||X_2^TX_1(X_1^TX_1)^{-1}||_\infty||\boldsymbol{\xi}_S||_\infty = O(n^{\frac{1}{2}-t}\sqrt{n\log n}/n) \\ &= O(n^{-t}\sqrt{\log n}) = o(\lambda_n). \end{aligned} \tag{A.40}$$

Again by condition $C(3)$,

$$\begin{aligned}
(IV) &= \|X_2^T X_1 (X_1^T X_1)^{-1}\|_\infty \|\nabla_1 J(\hat{\boldsymbol{\beta}}_1))\|_\infty \\
&\leqslant C\frac{p'_{\lambda_n}(0+)}{p'_{\lambda_n}(d_n)} * \lambda_n * p'_{\lambda_n}(d_n) \\
&= \lambda_n C p'_{\lambda_n}(0+) \\
&< \lambda_n p'_{\lambda_n}(0+). \qquad \text{(A.41)}
\end{aligned}$$

Combining (I), (II), (III), (IV) and the fact that $p'_{\lambda_n}(0+)$ does not depend on λ_n, we have proved $\hat{\boldsymbol{\beta}}$ satisfies inequality (2.10).

Finally, by condition $(C6)$, we know $\frac{1}{n}X_1^T X_1 + \nabla^2 J_{\lambda_n}(\hat{\boldsymbol{\beta}}_1)$ is positive definite. Inequality (2.11) is satisfied for sufficiently large n. Therefore, $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $C(\boldsymbol{\beta})$.

By condition $(C7)$ and $\hat{\boldsymbol{\beta}}_1 \in \mathscr{N}$, we have:

$$|\hat{\beta}_{[h]}| \geqslant |\beta^*_{[h]}| - n^{-\gamma}\log n > |\beta^*_{[i]}| + n^{-\gamma}\log n \geqslant |\hat{\beta}_{[i]}| \qquad \forall h,\ i, \qquad \text{(A.42)}$$

where $1 \leqslant h \leqslant p - K_n$, and $p - K_n + 1 \leqslant i \leqslant s$. And $\hat{\beta}_{[i]} = 0$, for $i \geqslant s+1$. Following the ideas of Proposition 2.4, $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q(\boldsymbol{\beta})$.

This completes the proof. □

s

Proof of Theorem 2.6

Proof. STEP 1: CONSISTENCY IN THE S-DIMENSIONAL SUBSPACE.

We first constrain $C(\boldsymbol{\beta})$ on the s-dimension subspace: $\mathbb{S} = \{\boldsymbol{\beta} \in \mathbb{R}^p \big| \beta_j = 0,\ s+1 \leqslant j \leqslant p\}$ of $\mathbb{R}^p$. Then the minimizer of $\min_{\beta\in\mathbb{S}} C(\boldsymbol{\beta})$ is the ordinary least square estimators given by

$$\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y} = \boldsymbol{\beta}_1^* + (X_1'X_1)^{-1}X_1'\boldsymbol{\epsilon}. \qquad \text{(A.43)}$$

By standard arguments, we have:

$$\begin{aligned} nc\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2^2 &\leqslant \|(X_1'X_1)^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*)\|_2^2 \\ &= \boldsymbol{\epsilon}'X_1(X_1'X_1)^{-1}X_1'\boldsymbol{\epsilon} \\ &= O_p\Big[\mathbb{E}\{\boldsymbol{\epsilon}'X_1(X_1'X_1)^{-1}X_1'\boldsymbol{\epsilon}\}\Big] \\ &= \sigma^2 tr\{X_1(X_1'X_1)^{-1}X_1\}O_p(1) \\ &= O_p(s). \end{aligned} \tag{A.44}$$

Therefore, $\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2 = O_p(\sqrt{s/n})$.

Step 2: Sparsity.

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T \in \mathbb{R}^p$, we want to show that $\hat{\boldsymbol{\beta}}$ satisfies inequality (2.10), i.e.:

$$\frac{1}{n}\|X_2^T(\mathbf{y}-X_1\hat{\boldsymbol{\beta}}_1)\|_\infty < \lambda_n p'_{\lambda_n}(0+) \tag{A.45}$$

Following the same idea as in the proof of Theorem 2.5, we let $\boldsymbol{\xi} = (\xi_1,\dots,\xi_p)^T = X^T(\mathbf{y}-X\boldsymbol{\beta}^*) = X^T\boldsymbol{\epsilon}$, and consider the event set

$$\mathscr{A}_2 = \{\|\boldsymbol{\xi}_{S^c}\|_\infty \leqslant \lambda_n\sqrt{2\sigma^2 n}\}, \tag{A.46}$$

where index set $S^c = \{j \mid s+1 \leqslant j \leqslant p\}$. Notice that the diagonal elements of $\frac{X^TX}{n}$ is 1, so $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n x_{ij}\epsilon_i$ satisfies standard normal distribution. By normal distribution tail inequality, for $n \geqslant 2$, we have:

$$\begin{aligned} P(\mathscr{A}_2^c) &\leqslant \sum_{j=s+1}^p P(|\xi_j| > \lambda_n\sqrt{2\sigma^2 n}) = \sum_{j=s+1}^p P(\frac{1}{\sigma\sqrt{n}}|\xi_j| > \sqrt{2}\lambda_n) \\ &\leqslant \sum_{j=s+1}^p \frac{e^{-\lambda_n^2}}{\lambda_n\sqrt{\pi}} \leqslant (p-s)\frac{e^{-\lambda_n^2}}{\lambda_n\sqrt{\pi}} \to 0, \quad \text{as } n\to\infty. \end{aligned} \tag{A.47}$$

Notice that

$$\begin{aligned}
& \frac{1}{n}||X_2^T(\mathbf{y}-X_1\hat{\boldsymbol{\beta}}_1)||_\infty \\
= \; & \frac{1}{n}\left|\left|X_2^T(\mathbf{y}-X_1\boldsymbol{\beta}_1^*)-X_2^TX_1(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*)\right|\right|_\infty \\
\leqslant \; & \frac{1}{n}||\boldsymbol{\xi}_{S^c}||_\infty+\frac{1}{n}||X_2^TX_1(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*)||_\infty \\
\leqslant \; & \frac{1}{n}||\boldsymbol{\xi}_{S^c}||_\infty+\frac{1}{n}||X_2^TX_1||_{2,\infty}||\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*||_2 \\
\leqslant \; & \sqrt{\frac{2\sigma^2}{n}}\lambda_n+\frac{\lambda_n}{\sqrt{n}}=o(\lambda_n). \qquad \text{(A.48)}
\end{aligned}$$

Therefore, for sufficient large n, $\hat{\boldsymbol{\beta}}$ satisfies inequality (2.10).

By step 1 and step 2 together with condition $(C10)$, we know $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $C(\boldsymbol{\beta})$.

By condition $(C11)$ and $||\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*||_2=O_p(\sqrt{s/n})$, we have $\min\{|\hat{\beta}_j|\,\big|1\leqslant j\leqslant s\}$ is bounded away from 0. By Proposition 2.4, $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q(\boldsymbol{\beta})$.

This completes the proof. □

Proof of Theorem 2.9

Proof. Let $\boldsymbol{\xi}=(\xi_{11},\ldots,\xi_{1p_1},\ldots,\xi_{G,p_G})^T=X^T(\mathbf{y}-X\boldsymbol{\beta}^*)=X^T\boldsymbol{\epsilon}$, consider event sets

$$\mathscr{A}_1=\{||\boldsymbol{\xi}_{H_0}||_\infty\leqslant\sqrt{2\sigma^2 n\log n}\} \quad \text{and} \quad \mathscr{A}_2=\{||\boldsymbol{\xi}_{H_0^c}||_\infty\leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n\log n}\}. \qquad \text{(A.49)}$$

Notice that the diagonal elements of $\frac{X^TX}{n}$ is 1, so $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n x_{ij}\epsilon_i$ satisfies standard normal distribution. By normal distribution tail inequality, for $n\geqslant 2$, we have:

$$P(\mathscr{A}_1^c)\leqslant\frac{s_0}{n}; \qquad \text{(A.50)}$$

$$P(\mathscr{A}_2^c)\leqslant(p-s_0)e^{-n^{1-2t}\log n}. \qquad \text{(A.51)}$$

By (A.50) and (A.51), we know $P(\mathscr{A}_1 \cap \mathscr{A}_2) \geqslant 1 - (s_0 n^{-1} + (p - s_0) e^{-n^{1-2t} \log n})$. Under the intersection of $\mathscr{A}_1 \cap \mathscr{A}_2$, we will show that there exists a solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ of (2.21)-(2.24).

Step 1: Existence of a solution for equations (2.21) and (2.22). We first show for sufficiently large n, equations (2.21) and (2.22) have a solution $\hat{\boldsymbol{\beta}}_{H_0}$ inside the hyper cube:

$$\mathscr{N} = \{\boldsymbol{\delta} \in \mathbb{R}^{H_0} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^*_{H_0}\|_\infty = n^{-\gamma} \log n\}. \tag{A.52}$$

For each g such that $1 \leqslant g \leqslant G_0$, let $|\beta^*_{gj}| = \max_{1 \leqslant l \leqslant p_g} |\beta^*_{gl}| \neq 0$, by $d_n > n^{-\gamma} \log n$, we have:

$$|\delta_{gj}| \geqslant |\beta^*_{gj}| - n^{-\gamma} \log n \geqslant d_n \tag{A.53}$$

and $sign(\delta_{gj}) = sign(\beta^*_{gj})$. Therefore, in the the hyper cube $\mathscr{N}$, $\|\boldsymbol{\delta}_g\|_2 \neq 0$, for $1 \leqslant g \leqslant G_0$. This implies that if group g is an important group of $\boldsymbol{\beta}^*$, group g will also be an important group of $\boldsymbol{\delta}$.

By the monotonicity condition of $p'_{\lambda_n}(t)$ and (A.53), we have:

$$\|\lambda_n p'_{\lambda_n}(\|\boldsymbol{\delta}_g\|_2) \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\|_\infty \leqslant \lambda_n p'_{\lambda_n}(f_n). \tag{A.54}$$

By inequality (A.54) and the definition of ξ_{H_0}, we have:

$$\|\xi_{H_0} - n \nabla J_{\lambda_n}(\boldsymbol{\delta})\|_\infty \leqslant \sqrt{2\sigma^2 n \log n} + n \lambda_n p'_{\lambda_n}(f_n) \tag{A.55}$$

Define $\Phi(\boldsymbol{\delta})$ which corresponds to n times the left hand sides of equalities (2.21) and (2.22):

$$\begin{aligned}
\Phi(\boldsymbol{\delta}) &\triangleq X^T_{H_0}\mathbf{y} - X^T_{H_0} X_{H_0} \boldsymbol{\delta} - n \nabla J_{\lambda_n}(\boldsymbol{\delta}) \\
&= X^T_{H_0}(\mathbf{y} - X_{H_0}\boldsymbol{\beta}^*_{H_0}) - X^T_{H_0} X_{H_0}(\boldsymbol{\delta} - \boldsymbol{\beta}^*_{H_0}) - n \nabla J_{\lambda_n}(\boldsymbol{\delta}) \\
&= (\xi_{H_0} - n \nabla J_{\lambda_n}(\boldsymbol{\delta})) - X^T_{H_0} X_{H_0}(\boldsymbol{\delta} - \boldsymbol{\beta}^*_{H_0}).
\end{aligned} \tag{A.56}$$

Let

$$\widetilde{\Phi}(\boldsymbol{\delta}) \triangleq (X^T_{H_0} X_{H_0})^{-1} \Phi(\boldsymbol{\delta}) = \mathbf{u} - (\boldsymbol{\delta} - \boldsymbol{\beta}^*_{H_0}), \tag{A.57}$$

where $\mathbf{u} = (X_{H_0}^T X_{H_0})^{-1}(\boldsymbol{\xi}_{H_0} - n\nabla J_{\lambda_n}(\boldsymbol{\delta}))$. By assumptions $(C15)$ and $(C16)$:

$$\begin{aligned} \|\mathbf{u}\|_\infty &\leqslant \|(X_{H_0}^T X_{H_0})^{-1}\|_\infty(\|\boldsymbol{\xi}_{H_0}\|_\infty + n\|\nabla J_{\lambda_n}(\boldsymbol{\delta})\|_\infty) \\ &= o(n^{-\gamma}\log n). \end{aligned} \tag{A.58}$$

So for sufficiently large n, if $(\boldsymbol{\delta} - \boldsymbol{\beta}_{H_0}^*)_{gj} = n^{-\gamma}\log n$, we have:

$$\widetilde{\Phi}_{gj}(\boldsymbol{\delta}) \leqslant \|\mathbf{u}\|_\infty - n^{-\gamma}\log n \leqslant 0; \tag{A.59}$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_{H_0}^*)_{gj} = -n^{-\gamma}\log n$, we have:

$$\widetilde{\Phi}_{gj}(\boldsymbol{\delta}) \geqslant -\|\mathbf{u}\|_\infty + n^{-\gamma}\log n \geqslant 0, \tag{A.60}$$

where $\widetilde{\Phi}_{gj}(\boldsymbol{\delta})$ is the gjth covariate of $\widetilde{\Phi}(\boldsymbol{\delta})$. By the continuity of $\widetilde{\Phi}(\boldsymbol{\delta})$, (A.59) and (A.60), Miranda's existence theorem guarantees that equation $\widetilde{\Phi}(\boldsymbol{\delta}) = \mathbf{0}$ has a solution in $\mathscr{N}$, which we denote as $\hat{\boldsymbol{\beta}}_{H_0}$. It is easy to see, $\hat{\boldsymbol{\beta}}_{H_0}$ also solves $\Phi(\boldsymbol{\delta}) = \mathbf{0}$. Thus, equations (2.21) and (2.22) do have a solution inside the hypercube $\mathscr{N}$.

Step 2: Verification of Strict Inequality (2.23).

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_{G_0}^T, \mathbf{0}^T)^T \in \mathbb{R}^p$, with $(\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_{G_0}^T)^T \in \mathscr{N}$ being a solution to equations (2.21) and (2.22). We want to show that $\hat{\boldsymbol{\beta}}$ satisfies inequality (2.23), i.e.:

$$\frac{1}{n}\|X_g^T(\mathbf{y} - X_{H_0}\hat{\boldsymbol{\beta}}_{H_0})\|_2 < \lambda_n p_{\lambda_n}'(0+), \quad G_0 + 1 \leqslant g \leqslant G. \tag{A.61}$$

Notice that

$$\begin{aligned} & \frac{1}{n}\|X_g^T(\mathbf{y} - X_{H_0}\hat{\boldsymbol{\beta}}_{H_0})\|_2 \\ = & \frac{1}{n}\left\|X_g^T(\mathbf{y} - X_{H_0}\boldsymbol{\beta}_{H_0}^*) - X_g^T X_{H_0}(\hat{\boldsymbol{\beta}}_{H_0} - \boldsymbol{\beta}_{H_0}^*)\right\|_2 \\ \leqslant & \frac{1}{n}\sqrt{p_g}\|\boldsymbol{\xi}_{H_0^c}\|_\infty + \frac{1}{n}\|X_g^T X_{H_0}(\hat{\boldsymbol{\beta}}_{H_0} - \boldsymbol{\beta}_{H_0}^*)\|_2 \\ \triangleq & (I) + (II). \end{aligned} \tag{A.62}$$

Because $||\xi_{H_0^c}||_\infty \leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}$ and $n^{-t}\sqrt{L\log n}/\lambda_n = o(1)$, we have:

$$(I) = O(n^{-t}\sqrt{\log n}) = o(\lambda_n). \tag{A.63}$$

Because $\hat{\beta}_{H_0}$ is a solution to $\Phi(\delta) = \mathbf{0}$, we have:

$$\hat{\beta}_{H_0} - \beta^*_{H_0} = (X_{H_0}^T X_{H_0})^{-1}(\xi_{H_0} - n\nabla J_{\lambda_n}(\hat{\beta}_{H_0})). \tag{A.64}$$

Therefore,

$$\begin{aligned}
(II) &= \frac{1}{n}||X_g^T X_{H_0}(X_{H_0}^T X_{H_0})^{-1}(\xi_{H_0} - n\nabla J_{\lambda_n}(\hat{\beta}_{H_0}))||_2 \\
&\leqslant \frac{1}{n}||X_g^T X_{H_0}(X_{H_0}^T X_{H_0})^{-1}||_{\infty,2}||(\xi_{H_0} - n\nabla J_{\lambda_n}(\hat{\beta}_{H_0}))||_\infty \\
&\leqslant \frac{1}{n}||X_g^T X_{H_0}(X_{H_0}^T X_{H_0})^{-1}||_{\infty,2}\Big(||\xi_{H_0}||_\infty + ||n\nabla J_{\lambda_n}(\hat{\beta}_{H_0})||_\infty\Big) \\
&= (III) + (IV).
\end{aligned} \tag{A.65}$$

By conditions $C(14)$ and $C(16)$,

$$\begin{aligned}
(III) &= \frac{1}{n}||X_g^T X_{H_0}(X_{H_0}^T X_{H_0})^{-1}||_{\infty,2}||\xi_{H_0}||_\infty = O(n^{\frac{1}{2}-t}\sqrt{n\log n}/n) \\
&= O(n^{-t}\sqrt{\log n}) = o(\lambda_n).
\end{aligned} \tag{A.66}$$

Again by condition $C(14)$,

$$\begin{aligned}
(IV) &= ||X_g^T X_{H_0}(X_{H_0}^T X_{H_0})^{-1}||_{\infty,2}||\nabla J(\hat{\beta}_{H_0}))||_\infty \\
&\leqslant C\frac{p'_{\lambda_n}(0+)}{p'_{\lambda_n}(f_n)} * \lambda_n * p'_{\lambda_n}(f_n) \\
&= \lambda_n C p'_{\lambda_n}(0+) \\
&< \lambda_n p'_{\lambda_n}(0+).
\end{aligned} \tag{A.67}$$

Combining (I), (II), (III), (IV) and the fact that $p'_{\lambda_n}(0+)$ does not depend on λ_n, we have proved $\hat{\beta}$ satisfies inequality (2.23).

Finally, by condition $(C17)$, we know $\frac{1}{n}X_{H_0}^T X_{H_0} + \nabla^2 J_{\lambda_n}(\hat{\boldsymbol{\beta}}_{H_0})$ is positive definite and inequality (2.24) is satisfied for sufficiently large n.

By condition $(C18)$ and $\hat{\boldsymbol{\beta}}_{H_0} \in \mathscr{N}$, we have:

$$||\hat{\beta}_{[g_1]}||_2 \geqslant ||\beta^*_{[g_1]}||_2 - \sqrt{L}n^{-\gamma}\log n > ||\beta^*_{[g_2]}||_2 + \sqrt{L}n^{-\gamma}\log n \geqslant ||\hat{\beta}_{[g_2]}||_2, \qquad \text{(A.68)}$$

where $1 \leqslant g_1 \leqslant G - K_n$, and $G - K_n + 1 \leqslant g_2 \leqslant G_0$. And $\hat{\beta}_{[g]} = \mathbf{0}$, for $g \geqslant G_0 + 1$. By Theorem 2.8, we know $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q_g(\boldsymbol{\beta})$.

This completes the proof. □

Proof of Theorem 3.1

Proof. We will first derive the necessary condition.

Suppose $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\hat{\boldsymbol{\beta}})$, then $\exists\ \mathbf{v} = (v_1, \dots, v_p)^T \in \mathbb{R}^p$, such that:

$$\nabla Q(\hat{\boldsymbol{\beta}}) = -\frac{1}{n} X^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}) + \lambda\mathbf{v} = 0. \tag{A.69}$$

Here:

$$v_i = \frac{a^{|\hat{\beta}_i|}}{T} sign(\hat{\beta}_i) \quad \text{if } \hat{\beta}_i \neq 0; \tag{A.70}$$

$$v_i \in [-\frac{1}{T}, \frac{1}{T}] \qquad \text{if } \hat{\beta}_i = 0. \tag{A.71}$$

By (A.69), (A.70) and (A.71), it is easy to verify the following is true:

$$\frac{1}{n} X_1^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1) - \lambda\nabla_1 J(\hat{\boldsymbol{\beta}}_1) = 0, \tag{A.72}$$

$$\frac{1}{n} ||X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)||_\infty \leqslant \frac{\lambda}{T}. \tag{A.73}$$

Note that $\hat{\boldsymbol{\beta}}$ is also a local minimizer of (3.8) constrained on the s-dimension subspace $\mathbb{S} = \{\boldsymbol{\beta} \in \mathbb{R}^p \Big| \beta_i = 0,\ s+1 \leqslant i \leqslant p\}$. It follows from the second order condition that $\frac{1}{n} X_1^T X_1 + \nabla_1^2 J(\hat{\boldsymbol{\beta}}_1)$ is positive semi-definite. So,

$$\frac{1}{n} X_1^T X_1 + \lambda\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1) \geqslant 0. \tag{A.74}$$

Next, we prove the sufficient condition.

We first constrain the penalized likelihood $Q(\boldsymbol{\beta})$ on the s-dimension subspace $\mathbb{S}$ of $\mathbb{R}^p$. Consider a neighborhood $\mathbb{B}_0$ in the subspace $\mathbb{S}$ centered around $\hat{\boldsymbol{\beta}}$, conditions (3.9) and (3.11) entail that $\hat{\boldsymbol{\beta}}$ is the unique local minimizer of $Q(\boldsymbol{\beta})$ in $\mathbb{B}_0 \subset \mathbb{R}^s$.

It remains to show that $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q(\boldsymbol{\beta})$ in $\mathbb{R}^p$. Consider a sufficiently small ball $\mathbb{B}_1 \subset \mathbb{R}^p$ centered around $\hat{\boldsymbol{\beta}}$, such that $\mathbb{B}_1 \cap \mathbb{S} \subset \mathbb{B}_0$. For any $\boldsymbol{\eta}^1 \in \mathbb{B}_1 \setminus \mathbb{B}_0$, let $\boldsymbol{\eta}^2$ be the projection of $\boldsymbol{\eta}^1$ onto the subspace $\mathbb{S}$. Then we have $\boldsymbol{\eta}^2 \in \mathbb{S}$, which entails that $Q(\hat{\boldsymbol{\beta}}) < Q(\boldsymbol{\eta}^2)$ if $\boldsymbol{\eta}^2 \neq \hat{\boldsymbol{\beta}}$. Thus, it suffices to show that

$Q(\boldsymbol{\eta}^2) < Q(\boldsymbol{\eta}^1)$.

By the mean value theorem, we have:

$$Q(\boldsymbol{\eta}^1) - Q(\boldsymbol{\eta}^2) = \nabla Q(\boldsymbol{\eta}^0)^T(\boldsymbol{\eta}^1 - \boldsymbol{\eta}^2), \tag{A.75}$$

where $\boldsymbol{\eta}^0 = \boldsymbol{\eta}^2 + t(\boldsymbol{\eta}^1 - \boldsymbol{\eta}^2)$ for some $t \in [0, 1]$.

Let η_i^1 and η_i^2 be the i-th component of $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$, respectively. Then we have:

$$\eta_i^1 - \eta_i^2 = 0 \quad \text{for } 1 \leqslant i \leqslant s, \tag{A.76}$$

$$\eta_i^2 = 0 \quad \text{for } s+1 \leqslant i \leqslant p. \tag{A.77}$$

Therefore, $\eta_i^0 = \eta_i^1$, for $1 \leqslant i \leqslant s$; and $\eta_i^0 = t\eta_i^1$, for $s+1 \leqslant i \leqslant p$.

The right hand side of (A.75) can be expressed as:

$$-\frac{1}{n}X_2^T(\mathbf{y} - X\boldsymbol{\eta}^0)\boldsymbol{\eta}_2^1 + \lambda \sum_{i=s+1}^{p} \frac{\partial J(\boldsymbol{\eta}^0)}{\partial \eta_i^0}\eta_i^1, \tag{A.78}$$

where $\boldsymbol{\eta}_2^1$ is a subvector of $\boldsymbol{\eta}^1$ formed by the last $p - s$ components of $\boldsymbol{\eta}^1$. By $\boldsymbol{\eta}^1 \in \mathbb{B}_1 \setminus \mathbb{B}_0$, we have $\boldsymbol{\eta}_2^1 \neq \mathbf{0}$.

Because $\eta_i^0 = t\eta_i^1$, for $s+1 \leqslant i \leqslant p$. So if $\eta_i^1 \neq 0$,

$$\frac{\partial J(\boldsymbol{\eta}^0)}{\partial \eta_i^0}\eta_i^1 = \frac{a^{|\eta_i^0|}}{\sum_{j=1}^{p} a^{|\eta_j^0|}} sign(\eta_i^0)\eta_i^1 = \frac{a^{|\eta_i^0|}}{\sum_{j=1}^{p} a^{|\eta_j^0|}}|\eta_i^1|. \tag{A.79}$$

For given $\hat{\beta}_1, \ldots, \hat{\beta}_s$, consider function $f(x)$

$$f(x) \triangleq \frac{\lambda a^x}{\sum_{i=1}^{s} a^{|\hat{\beta}_i| - x} + (p - s - 1) + a^x}, \qquad x \geqslant 0. \tag{A.80}$$

Then inequality (3.10) is equivalent to:

$$\frac{1}{n}||X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)||_\infty < f(0). \tag{A.81}$$

By the continuity of $f(x)$ and (A.81), there exists some $\delta > 0$ with $\delta < \min\{|\hat{\beta}_i| \mid \hat{\beta}_i \neq 0\}$, such that for any $\boldsymbol{\beta}$ in a ball in $\mathbb{R}^p$ centered at $\hat{\boldsymbol{\beta}}$ with radius δ, we have:

$$\frac{1}{n}\|X_2^T(\mathbf{y} - X\boldsymbol{\beta})\|_\infty < \frac{\lambda a^\delta}{\sum_{i\leqslant s} a^{|\hat{\beta}_i|-\delta} + (p-s-1) + a^\delta}. \tag{A.82}$$

By inequalities (A.82), we further shrink the ball $\mathbb{B}_1$ such that: (1). $|\eta_i^1 - \hat{\beta}_i| < \delta$, for $i \leqslant s$; and (2). $|\eta_i^1| < \delta$ for $i \geqslant s+1$. Therefore, for $i \geqslant s+1$, we have:

$$\begin{aligned}
\left|\frac{\partial J(\boldsymbol{\eta}^0)}{\partial \eta_i^0}\right| &= \frac{a^{|\eta_i^0|}}{\sum_{j=1}^s a^{|\eta_j^0|} + a^{|\eta_i^0|} + \sum_{j>s, j\neq i} a^{|\eta_j^0|}} \\
&> \frac{a^{|\eta_i^0|}}{\sum_{j=1}^s a^{|\hat{\beta}_j|-\delta} + a^{|\eta_i^0|} + (p-s-1)} \\
&> \frac{a^\delta}{\sum_{j\leqslant s} a^{|\hat{\beta}_j|-\delta} + a^\delta + (p-s-1)}.
\end{aligned} \tag{A.83}$$

Combining (A.75), (A.78), (A.79) and (A.83), we have

$$\begin{aligned}
& Q(\boldsymbol{\eta}^1) - Q(\boldsymbol{\eta}^2) \\
> & -\frac{\lambda a^\delta}{\sum_{j\leqslant s} a^{|\hat{\beta}_j|-\delta} + a^\delta + (p-s-1)}|\boldsymbol{\eta}_2^1| + \frac{\lambda a^\delta}{\sum_{j\leqslant s} a^{|\hat{\beta}_j|-\delta} + a^\delta + (p-s-1)}|\boldsymbol{\eta}_2^1| \\
= & \ 0.
\end{aligned}$$

By the arbitrariness of $\boldsymbol{\eta}_1$, we complete our proof. □

Proof of Proposition 3.2

Proof. Notice that (3.12) can be further decomposed as:

$$\begin{aligned}
& \nabla_1^2 J(\hat{\boldsymbol{\beta}}_1) \\
= \ & \frac{\log a}{T^2}\left(- \begin{bmatrix} a^{|\hat{\beta}_1|+|\hat{\beta}_1|} sign(\hat{\beta}_1\hat{\beta}_1) & \cdots & a^{|\hat{\beta}_1|+|\hat{\beta}_s|} sign(\hat{\beta}_1\hat{\beta}_s) \\ \vdots & \ddots & \vdots \\ a^{|\hat{\beta}_1|+|\hat{\beta}_s|} sign(\hat{\beta}_1\hat{\beta}_s) & \cdots & a^{|\hat{\beta}_s|+|\hat{\beta}_s|} sign(\hat{\beta}_s\hat{\beta}_s) \end{bmatrix} + T \begin{bmatrix} a^{|\hat{\beta}_1|} & & \\ & \ddots & \\ & & a^{|\hat{\beta}_s|} \end{bmatrix} \right) \\
= \ & -\frac{\log a}{T^2}\left(\begin{bmatrix} a^{|\hat{\beta}_1|} sign(\hat{\beta}_1) \\ \vdots \\ a^{|\hat{\beta}_s|} sign(\hat{\beta}_s) \end{bmatrix} \begin{bmatrix} a^{|\hat{\beta}_1|} sign(\hat{\beta}_1) & \cdots & a^{|\hat{\beta}_s|} sign(\hat{\beta}_s) \end{bmatrix} + T \begin{bmatrix} -a^{|\hat{\beta}_1|} & & \\ & \ddots & \\ & & -a^{|\hat{\beta}_s|} \end{bmatrix} \right) \\
\triangleq \ & -\frac{\log a}{T^2}(A + TB). \qquad \text{(A.84)}
\end{aligned}$$

then we have:

$$\begin{aligned}
\lambda_{min}(\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1)) \ & \geqslant \ -\frac{\log a}{T^2}(\lambda_{min}(A) + T\lambda_{min}(B)) \\
& > \ -\frac{\log a}{T^2}(0 - T) > \frac{\log a}{p - s}. \qquad \text{(A.85)}
\end{aligned}$$

The second inequality is because the diagonal element of B matrix is strictly greater than -1. The third inequality is because $T > p - s$. Therefore,

$$\lambda_{min}(\frac{1}{n}X_1^T X_1 + \lambda\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1)) \geqslant \lambda_{min}(\frac{1}{n}X_1^T X_1) + \lambda_{min}(\lambda\nabla_1^2 J(\hat{\boldsymbol{\beta}}_1)) > 0, \qquad \text{(A.86)}$$

as long as (3.12) holds. □

Proof of Theorem 3.3

Proof. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T = X^T(\mathbf{y} - X\boldsymbol{\beta}) = X^T\boldsymbol{\epsilon}$, consider event sets

$$\mathscr{A}_1 = \{||\boldsymbol{\xi}_S||_\infty \leqslant \sqrt{2\sigma^2 n \log n}\} \quad \text{and} \quad \mathscr{A}_2 = \{||\boldsymbol{\xi}_{S^c}||_\infty \leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}\}, \qquad \text{(A.87)}$$

where index set $S = \{j \,|\, 1 \leqslant j \leqslant s\}$ and index set $S^c = \{j \,|\, s+1 \leqslant j \leqslant p\}$. Notice that the diagonal elements of $\frac{X^TX}{n}$ is 1, so $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n x_{ij}\epsilon_i$ satisfies standard normal distribution. By normal distribution tail inequality, for $n \geqslant 2$, we have:

$$\begin{aligned} P(\mathscr{A}_1^c) &\leqslant \sum_{j=1}^{s} P(|\xi_j| > \sqrt{2\sigma^2 n \log n}) = \sum_{j=1}^{s} P(\frac{1}{\sigma\sqrt{n}}|\xi_j| > \sqrt{2\log n}) \\ &\leqslant \sum_{j=1}^{s} \frac{1}{n\sqrt{\pi \log n}} \leqslant \frac{s}{n}; \end{aligned} \tag{A.88}$$

$$\begin{aligned} P(\mathscr{A}_2^c) &\leqslant \sum_{j=s+1}^{p} P(|\xi_j| > n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}) = \sum_{j=s+1}^{p} P(\frac{1}{\sigma\sqrt{n}}|\xi_j| > n^{\frac{1}{2}-t}\sqrt{2\log n}) \\ &\leqslant \sum_{j=s+1}^{p} \frac{e^{-n^{1-2t}\log n}}{n^{\frac{1}{2}-t}\sqrt{\pi \log n}} \leqslant (p-s)e^{-n^{1-2t}\log n}. \end{aligned} \tag{A.89}$$

By (A.88) and (A.89), we know $P(\mathscr{A}_1 \cap \mathscr{A}_2) \geqslant 1 - (sn^{-1} + (p-s)e^{-n^{1-2t}\log n})$. Under the intersection of $\mathscr{A}_1 \cap \mathscr{A}_2$, we will show that there exists an solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ to (3.9)-(3.11) such that $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^*)$ and $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_\infty = O(n^{-\gamma}\log n)$, where the function *sign* applies componentwise.

Step 1: Existence of a solution to equation (3.9). We first show for sufficiently large n, equation (3.9) has a solution $\hat{\boldsymbol{\beta}}_1$ inside the hyper cube:

$$\mathscr{N} = \{\boldsymbol{\delta} \in \mathbb{R}^s : ||\boldsymbol{\delta} - \boldsymbol{\beta}_1^*||_\infty = n^{-\gamma}\log n\}. \tag{A.90}$$

For $\forall\, \boldsymbol{\delta} \in \mathscr{N}$, since $d_n \geqslant n^{-\gamma}\log n$, we have:

$$\min_{1\leqslant j\leqslant s} |\delta_j| \geqslant \min_{1\leqslant j\leqslant s} |\beta_j^*| - n^{-\gamma}\log n = d_n, \tag{A.91}$$

and $sign(\boldsymbol{\delta}) = sign(\boldsymbol{\beta}_1^*)$.

For any given i, $1 \leqslant i \leqslant s$, it is easy to see:

$$\begin{aligned} & \frac{a^{|\delta_i|}}{\sum_{j=1}^{s} a^{|\delta_j|} + (p-s)} \\ < \; & \frac{1}{\sum_{j=1}^{s} a^{|\delta_j|} + (p-s)} \\ < \; & \frac{1}{p-s}. \end{aligned} \tag{A.92}$$

Therefore:

$$\|\nabla_1 J(\boldsymbol{\delta})\|_\infty = \max\{\frac{a^{|\delta_i|}}{\sum_{j=1}^{s} a^{|\delta_j|} + (p-s)}, 1 \leqslant i \leqslant s\} < \frac{1}{p-s}. \tag{A.93}$$

Notice that equation (3.11) can be written as:

$$\begin{aligned} & X_1^T \mathbf{y} - X_1^T X_1 \boldsymbol{\delta} - n\lambda \nabla_1 J(\boldsymbol{\delta}) \\ = \; & X_1^T(\mathbf{y} - X_1 \boldsymbol{\beta}_1^*) - X_1^T X_1(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*) - n\lambda \nabla_1 J(\boldsymbol{\delta}) \\ = \; & (\boldsymbol{\xi}_S - n\lambda \nabla_1 J(\boldsymbol{\delta})) - X_1^T X_1(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*) \\ \triangleq \; & \Phi(\boldsymbol{\delta}). \end{aligned} \tag{A.94}$$

Let

$$\widetilde{\Phi}(\boldsymbol{\delta}) \triangleq (X_1^T X_1)^{-1} \Phi(\boldsymbol{\delta}) = \mathbf{u} - (\boldsymbol{\delta} - \boldsymbol{\beta}_1^*), \tag{A.95}$$

where $\mathbf{u} = (X_1^T X_1)^{-1}(\boldsymbol{\xi}_S - n\lambda \nabla_1 J(\boldsymbol{\delta}))$. By assumption $(C1)$, the choice of λ, and the fact that $\gamma < \alpha$, we have:

$$\begin{aligned} \|\mathbf{u}\|_\infty \; & \leqslant \; \|(X_1^T X_1)^{-1}\|_\infty (\|\boldsymbol{\xi}_S\|_\infty + n\lambda \|\nabla_1 J(\boldsymbol{\delta}))\|_\infty) \\ & = \; o(n^{-\frac{1}{2}-\alpha}\sqrt{\log n} * n^{\frac{1}{2}}\sqrt{\log n}) + o(\frac{n^{-\gamma} p \log n}{p-s}) \\ & = \; o(n^{-\gamma} \log n). \end{aligned} \tag{A.96}$$

So for sufficiently large n, if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*)_j = n^{-\gamma} \log n$, we have:

$$\widetilde{\Phi}_j(\boldsymbol{\delta}) \leqslant \|\mathbf{u}\|_\infty - n^{-\gamma} \log n \leqslant 0; \tag{A.97}$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1^*)_j = -n^{-\gamma} \log n$, we have:

$$\widetilde{\Phi}_j(\boldsymbol{\delta}) \geqslant -\|\mathbf{u}\|_\infty + n^{-\gamma} \log n \geqslant 0, \tag{A.98}$$

where $\widetilde{\Phi}(\boldsymbol{\delta}) = (\widetilde{\Phi}_1(\boldsymbol{\delta}), \ldots, \widetilde{\Phi}_s(\boldsymbol{\delta}))^T$. By the continuity of $\widetilde{\Phi}(\boldsymbol{\delta})$, (A.97) and (A.98), Miranda's existence theorem guarantees that equation $\widetilde{\Phi}(\boldsymbol{\delta}) = 0$ has a solution in $\mathscr{N}$, which we denote as $\hat{\boldsymbol{\beta}}_1$. It is easy to see, $\hat{\boldsymbol{\beta}}_1$ also solve $\Phi(\boldsymbol{\delta}) = 0$. Thus, equation (3.9) does have a solution inside the hypercube $\mathscr{N}$.

Step 2: Verification of Strict Inequality (3.10).

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T \in \mathbb{R}^n$, with $\hat{\boldsymbol{\beta}}_1 \in \mathscr{N}$ being a solution to equation (3.9). We want to show that $\hat{\boldsymbol{\beta}}$ satisfies inequality (3.10), i.e.:

$$\frac{1}{n}\|X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)\|_\infty < \frac{\lambda}{\sum_{i \leqslant s} a^{|\hat{\beta}_i|} + (p-s)}. \tag{A.99}$$

Notice that:

$$\begin{aligned} & \frac{1}{n\lambda}\|X_2^T(\mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1)\|_\infty \\ = \; & \frac{1}{n\lambda}\left\|\left(X_2^T(\mathbf{y} - X_1\boldsymbol{\beta}_1^*) - X_2^T X_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\right)\right\|_\infty \\ \leqslant \; & \frac{1}{n\lambda}\|\boldsymbol{\xi}_{S^c}\|_\infty + \frac{1}{n\lambda}\|X_2^T X_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\|_\infty \\ \triangleq \; & (I) + (II). \end{aligned} \tag{A.100}$$

Because $\|\boldsymbol{\xi}_{S^c}\|_\infty \leqslant n^{\frac{1}{2}-t}\sqrt{2\sigma^2 n \log n}$ and $\lambda/(\frac{p\sqrt{\log n}}{n^t}) \to \infty$

$$(I) = o(\frac{n^t}{np\sqrt{\log n}} n^{\frac{1}{2}-t}\sqrt{n \log n}) = o(\frac{1}{p}). \tag{A.101}$$

Because $\hat{\beta}_1$ is a solution to $\Phi(\delta) = 0$, we have:

$$\hat{\beta}_1 - \beta_1^* = (X_1^T X_1)^{-1}(\xi_S - n\lambda\nabla_1 J(\hat{\beta}_1)). \qquad (A.102)$$

Therefore,

$$\begin{aligned} (II) &= \frac{1}{n\lambda}||X_2^T X_1(X_1^T X_1)^{-1}(\xi_S - n\lambda\nabla_1 J(\hat{\beta}_1))||_\infty \\ &\leqslant \frac{1}{n\lambda}||X_2^T X_1(X_1^T X_1)^{-1}||_\infty||(\xi_S - n\lambda\nabla_1 J(\hat{\beta}_1))||_\infty \\ &\leqslant \frac{1}{n\lambda}||X_2^T X_1(X_1^T X_1)^{-1}||_\infty\Big(||\xi_S||_\infty + ||n\lambda\nabla_1 J(\hat{\beta}_1))||_\infty\Big) \\ &= (III) + (IV). \end{aligned} \qquad (A.103)$$

$$\begin{aligned} (III) &= \frac{1}{n\lambda}||X_2^T X_1(X_1^T X_1)^{-1}||_\infty||\xi_S||_\infty \\ &= o(\frac{1}{p}). \end{aligned} \qquad (A.104)$$

$$\begin{aligned} (IV) &= ||X_2^T X_1(X_1^T X_1)^{-1}||_\infty||\nabla_1 J(\hat{\beta}_1))||_\infty \\ &< \frac{1}{a^{d_n}} * \frac{a^{d_n}}{\sum_{j=1}^s a^{|\hat{\beta}_j|} + (p-s)} \\ &= \frac{1}{\sum_{j=1}^s a^{|\hat{\beta}_j|} + (p-s)}. \end{aligned} \qquad (A.105)$$

Combining (I), (II), (III) and (IV), we have shown that $\hat{\beta}$ satisfies inequality (3.10).

Finally, because

$$\frac{\lambda}{p-s} = O(\frac{n^{\alpha-\frac{1}{2}-\gamma}p\sqrt{\log n}}{p-s}) = O(n^{\alpha-\frac{1}{2}-\gamma}\sqrt{\log n}). \qquad (A.106)$$

By proposition 3.2, and condition $(C6)$, we know $\hat{\beta}$ satisfies the inequality (3.11) for sufficiently large n. This completes the proof. □

Proof of Proposition 4.1

Here we present the proof of Proposition 4.1.

Proof. : By KKT condition, $\hat{\boldsymbol{\beta}}^{LASSO}$, the solution of LASSO satisfies:

$$\frac{1}{n}X_{kj}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{LASSO}) = \gamma * sign(\hat{\beta}_{kj}^{LASSO}) \quad if \quad \hat{\beta}_{kj}^{LASSO} \neq 0, \tag{A.107}$$

$$\frac{1}{n}|X_{kj}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{LASSO})| \leqslant \gamma \quad if \quad \hat{\beta}_{kj}^{LASSO} = 0. \tag{A.108}$$

Similarly, $\hat{\boldsymbol{\beta}}^{LES}$, the solution of LES satisfies:

$$\frac{1}{n}X_{kj}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{LES}) = \frac{\lambda\alpha p_k}{p} * \frac{\exp(\alpha|\hat{\beta}_{kj}^{LES}|)}{\sum_{l=1}^{p_k}\exp(\alpha|\hat{\beta}_{kl}^{LES}|)} * sign(\hat{\beta}_{kj}^{LES}) \quad if \quad \hat{\beta}_{kj}^{LES} \neq 0, \tag{A.109}$$

$$\frac{1}{n}|X_{kj}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{LES})| \leqslant \frac{\lambda\alpha p_k}{p} * \frac{1}{1+\sum_{l\neq j}\exp(\alpha|\hat{\beta}_{kl}^{LES}|)} \quad if \quad \hat{\beta}_{kj}^{LES} = 0. \tag{A.110}$$

It is easy to see that if we let tuning parameter $\alpha \to 0$, each exponential term in the right hand side of the equations (A.109) and (A.110) will be close to 1. If we choose the tuning parameter λ such that $\frac{\lambda\alpha}{p} = \gamma$, then KKT condition for LES is approximately the same as KKT condition for LASSO. Therefore we have $\hat{\boldsymbol{\beta}}^{LES} - \hat{\boldsymbol{\beta}}^{LASSO} \to \mathbf{0}$. This completes the proof. □

Proof of Proposition 4.2

Here we present the proof of Proposition 4.2.

Proof. : Let $\hat{\boldsymbol{\beta}}$ be the solution of LES. By plugging $\boldsymbol{\beta} = 0$ into the objective function, we have:

$$\frac{1}{2n}\sum_{i=1}^{n}\Big(y_i-\sum_{k=1}^{K}\sum_{j=1}^{p_k}x_{i,kj}\hat{\beta}_{kj}\Big)^2+\lambda\sum_{k=1}^{K}w_k\log\Big\{\exp(\alpha|\hat{\beta}_{k1}|)+\cdots+\exp(\alpha|\hat{\beta}_{kp_k}|)\Big\}$$
$$\leqslant\frac{1}{2n}||\mathbf{y}||_2^2+\lambda\sum_{k=1}^{K}w_k\log(p_k). \tag{A.111}$$

From (A.111), we have the following two inequalities:

$$\sum_{i=1}^{n}\Big(y_i-\sum_{k=1}^{K}\sum_{j=1}^{p_k}x_{i,kj}\hat{\beta}_{kj}\Big)^2=||\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}||_2^2\leqslant||\mathbf{y}||_2^2+2n\lambda\sum_{k=1}^{K}w_k\log(p_k), \tag{A.112}$$

and

$$\exp(\alpha|\hat{\beta}_{k1}|)+\cdots+\exp(\alpha|\hat{\beta}_{kp_k}|)\leqslant\exp\Big\{\frac{1}{2n\lambda w_k}||\mathbf{y}||_2^2+\sum_{j=1}^{K}\frac{w_j}{w_k}\log(p_j)\Big\}. \tag{A.113}$$

So, if $\hat{\beta}_{ki}\hat{\beta}_{kj}>0$, by KKT condition:

$$X_{ki}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=n\lambda\alpha w_k\frac{\exp(\alpha|\hat{\beta}_{ki}|)}{\sum_l\exp(\alpha|\hat{\beta}_{kl}|)}sign(\hat{\beta}_{ki}); \tag{A.114}$$

$$X_{kj}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=n\lambda\alpha w_k\frac{\exp(\alpha|\hat{\beta}_{kj}|)}{\sum_l\exp(\alpha|\hat{\beta}_{kl}|)}sign(\hat{\beta}_{kj}). \tag{A.115}$$

Without loss of generality, we assume $\hat{\beta}_{ki}\geqslant\hat{\beta}_{kj}>0$, and $sign(\hat{\beta}_{ki})=sign(\hat{\beta}_{kj})=1$, then, by taking the difference between (A.114) and (A.115), we have:

$$(X_{ki}-X_{kj})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=n\lambda\alpha w_k\frac{\exp(\alpha\hat{\beta}_{ki})-\exp(\alpha\hat{\beta}_{kj})}{\sum_l\exp(\alpha|\hat{\beta}_{kl}|)}. \tag{A.116}$$

By the convexity of exponential function, we have:

$$\exp(\alpha\hat{\beta}_{ki}) - \exp(\alpha\hat{\beta}_{kj}) \geqslant \exp(\alpha\hat{\beta}_{kj}) * \alpha(\hat{\beta}_{ki} - \hat{\beta}_{kj}) \geqslant \alpha|\hat{\beta}_{ki} - \hat{\beta}_{kj}|. \quad \text{(A.117)}$$

By equality (A.116), we have:

$$|\exp(\alpha\hat{\beta}_{ki}) - \exp(\alpha\hat{\beta}_{kj})| \leqslant \frac{\sum_l \exp(\alpha|\hat{\beta}_{kl}|)}{n\lambda\alpha w_k}||X_{ki} - X_{kj}||_2||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||_2. \quad \text{(A.118)}$$

Combining (A.112),(A.113), (A.117) and (A.118), we can have the result we want. This completes the proof. □

Proof of Theorem 4.3

To prove Theorem 4.3, we first prove a useful lemma.

Lemma A.1. *Consider the model (4.1). Let $\boldsymbol{\beta}^*$ be the true coefficients of the linear model. Assume the random noise $\epsilon_1, \ldots, \epsilon_n$ are iid from normal distribution with mean zero and variance σ^2. Suppose the diagonal elements of matrix $\mathbf{X}^T\mathbf{X}/n$ are equal to 1. Let A be a real number bigger than $2\sqrt{2}$ and $\gamma = A\sigma\sqrt{\frac{\log p}{n}}$. Let two tuning parameters λ and α satisfy $\lambda\alpha = \gamma$. For any solution $\hat{\boldsymbol{\beta}}$ to the minimization problem (4.13), and any $\boldsymbol{\beta} \in \mathbb{R}^p$, with probability at least $1 - p^{1-A^2/8}$, the following inequality holds:*

$$\begin{aligned} & \frac{1}{n}||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2^2 + \gamma||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 \\ \leqslant \; & \frac{1}{n}||\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)||_2^2 + 2\gamma \sum_{k \in G(\boldsymbol{\beta})} (1 + p_k)||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k||_1. \end{aligned} \quad \text{(A.119)}$$

Proof. : Let $\hat{\boldsymbol{\beta}}$ be the solution to (4.13), then, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have:

$$\frac{1}{n}||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||_2^2 + J_{\lambda,\alpha}(\hat{\boldsymbol{\beta}}) \leqslant \frac{1}{n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + J_{\lambda,\alpha}(\boldsymbol{\beta}). \quad \text{(A.120)}$$

Here for notation simplicity, we denote $J_{\lambda,\alpha}(\boldsymbol{\beta}) \triangleq 2\lambda \sum_{k=1}^K p_k \log\{\exp(\alpha|\beta_{k1}|) +$

$\cdots + \exp(\alpha|\beta_{kp_k}|)\}$.

Because $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, the above inequality is equivalent to:

$$\frac{1}{n}||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2^2 \leqslant \frac{1}{n}||\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)||_2^2 + \frac{2}{n}\boldsymbol{\epsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + J_{\lambda,\alpha}(\boldsymbol{\beta}) - J_{\lambda,\alpha}(\hat{\boldsymbol{\beta}}). \quad \text{(A.121)}$$

Consider the random event set $\mathscr{A} = \{\frac{2}{n}||\mathbf{X}^T\boldsymbol{\epsilon}||_\infty \leqslant \gamma\}$, where $||\mathbf{X}^T\boldsymbol{\epsilon}||_\infty = \max_{kj}|\sum_{i=1}^n x_{i,kj}\epsilon_i|$. Because the diagonal elements of matrix $\mathbf{X}^T\mathbf{X}/n$ are equal to 1, the random variable $z_{kj} \triangleq \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n x_{i,kj}\epsilon_i$ follows the standard normal distribution, even though the z_{kj} may not be independent from each other. Let Z be another random variable from the standard normal distribution. For any kj, we have: $Pr(|\sum_{i=1}^n x_{i,kj}\epsilon_i| \geqslant \frac{\gamma n}{2}) = Pr(|Z| \geqslant \frac{\gamma\sqrt{n}}{2\sigma})$. Then by Gaussian tail inequality, we have:

$$Pr(\mathscr{A}^c) \leqslant Pr\{\cup_{k=1}^K \cup_{j=1}^{p_k} (|z_{kj}| \geqslant \frac{\gamma\sqrt{n}}{2\sigma})\} \leqslant p * Pr(|Z| \geqslant \frac{\gamma\sqrt{n}}{2\sigma}) \leqslant p^{1-A^2/8}. \quad \text{(A.122)}$$

Therefore, on the event set $\mathscr{A}$, with probability at least $1 - p^{1-A^2/8}$, we have:

$$\begin{aligned} & \frac{1}{n}||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2^2 \\ \leqslant\ & \frac{1}{n}||\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)||_2^2 + \frac{2}{n}||\mathbf{X}^T\epsilon||_\infty||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 + J_{\lambda,\alpha}(\boldsymbol{\beta}) - J_{\lambda,\alpha}(\hat{\boldsymbol{\beta}}) \\ \leqslant\ & \frac{1}{n}||\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)||_2^2 + \gamma||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 + J_{\lambda,\alpha}(\boldsymbol{\beta}) - J_{\lambda,\alpha}(\hat{\boldsymbol{\beta}}). \quad \text{(A.123)} \end{aligned}$$

Adding $\gamma\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ to both sides of (A.123), we have:

$$\begin{aligned}
& \frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \gamma\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\
\leqslant \ & \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + 2\gamma\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + J_{\lambda,\alpha}(\boldsymbol{\beta}) - J_{\lambda,\alpha}(\hat{\boldsymbol{\beta}}) \\
= \ & \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + 2\gamma \sum_{k \in G(\boldsymbol{\beta})} \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \\
& + 2\lambda \sum_{k \in G(\boldsymbol{\beta})} p_k \Big\{ \log(\sum_j e^{\alpha|\beta_{kj}|}) - \log(\sum_j e^{\alpha|\hat{\beta}_{kj}|}) \Big\} \\
& + 2\gamma \sum_{k \notin G(\boldsymbol{\beta})} \|\hat{\boldsymbol{\beta}}_k\|_1 + 2\lambda \sum_{k \notin G(\boldsymbol{\beta})} p_k \Big\{ \log(p_k) - \log(\sum_j e^{\alpha|\hat{\beta}_{kj}|}) \Big\}. \quad \text{(A.124)}
\end{aligned}$$

The last equality uses the fact that $\boldsymbol{\beta}_k = \mathbf{0}$, for $k \notin G(\boldsymbol{\beta})$.

We next show two simple inequalities. Suppose $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$ are $2m$ arbitrary real numbers, then the following inequalities hold:

$$\sum_{i=1}^{m} |a_i| \leqslant m \log\Big\{ (e^{|a_1|} + \ldots + e^{|a_m|})/m \Big\}, \quad \text{(A.125)}$$

$$\log(e^{|a_1|} + \ldots + e^{|a_m|}) - \log(e^{|b_1|} + \ldots + e^{|b_m|}) \leqslant \sum_{i=1}^{m} |a_i - b_i|. \quad \text{(A.126)}$$

The first inequality can be shown by using the arithmetic inequality and the geometric mean inequality; the second inequality follows by Log-Sum inequality.

By inequality (A.125), since $\lambda\alpha = \gamma$, we have:

$$2\gamma \sum_{k \notin G(\boldsymbol{\beta})} \|\hat{\boldsymbol{\beta}}_k\|_1 + 2\lambda \sum_{k \notin G(\boldsymbol{\beta})} p_k \Big\{ \log(p_k) - \log(\sum_j e^{\alpha|\hat{\beta}_{kj}|}) \Big\} \leqslant 0. \quad \text{(A.127)}$$

By inequality (A.126), we have:

$$2\lambda \sum_{k\in G(\boldsymbol{\beta})} p_k \Big\{ \log(\sum_j e^{\alpha|\beta_{kj}|}) - \log(\sum_j e^{\alpha|\hat{\beta}_{kj}|}) \Big\} \leqslant 2\lambda\alpha \sum_{k\in G(\boldsymbol{\beta})} p_k \|\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k\|_1 \tag{A.128}$$

Simplify inequality (A.124) using (A.127) and (A.128), then we have proved the lemma. □

Now we prove Theorem 3.

Proof. : Let $s = |G(\boldsymbol{\beta}^*)|$. In the event set $\mathscr{A} \triangleq \{\frac{2}{n}\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty \leqslant \gamma\}$, let $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in (A.119), then:

$$\begin{aligned} \frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 &\leqslant 2\gamma \sum_{k\in G(\boldsymbol{\beta}^*)} (1+p_k)\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_1 \\ &\leqslant 2\gamma\sqrt{s}\sqrt{\sum_{k\in G(\boldsymbol{\beta}^*)} p_k(1+p_k)^2\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_2^2}, \end{aligned} \tag{A.129}$$

where the second inequality follows by Cauchy inequality.

Using a similar argument, we have:

$$\sum_{k\notin G(\boldsymbol{\beta}^*)} \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_1 \leqslant \sum_{k\in G(\boldsymbol{\beta}^*)} (1+2p_k)\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_1. \tag{A.130}$$

If REgroup assumption holds with $\kappa = \kappa(s)$, then we have:

$$\sqrt{\sum_{k\in G(\boldsymbol{\beta}^*)} p_k(1+p_k)^2\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_2^2} \leqslant \frac{2\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2}{\kappa\sqrt{n}}. \tag{A.131}$$

By (A.129), (A.130) and (A.131), we have:

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leqslant \frac{16s\gamma^2}{\kappa^2}. \tag{A.132}$$

Notice that inequality (A.119) implies:

$$\begin{aligned} ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 &\leqslant 2 \sum_{k \in G(\boldsymbol{\beta}^*)} (1 + p_k) ||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*||_1 \leqslant 2\sqrt{s} \sqrt{\sum_{k \in G(\boldsymbol{\beta}^*)} p_k (1 + p_k)^2 ||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*||_2^2} \\ &\leqslant \frac{4\sqrt{s}||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2}{\kappa\sqrt{n}} \leqslant \frac{16 s \gamma}{\kappa^2}. \end{aligned} \tag{A.133}$$

Finally, we work on the bound for $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2$. For notation simplicity, we denote $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, and let $G = G(\boldsymbol{\beta}^*)$. When the REgroup assumption holds, we have:

$$\begin{aligned} ||\boldsymbol{\delta}_{G^c}||_2 &\leqslant ||\boldsymbol{\delta}_{G^c}||_1 \leqslant \sum_{k \in G(\boldsymbol{\beta}^*)} (1 + 2p_k) ||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*||_1 \leqslant 2\sqrt{s} \sqrt{\sum_{k \in G(\boldsymbol{\beta}^*)} p_k (1 + p_k)^2 ||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*||_2^2} \\ &\leqslant \frac{4\sqrt{s}||\mathbf{X}\boldsymbol{\delta}||_2}{\kappa\sqrt{n}} \leqslant \frac{16 s \gamma}{\kappa^2}. \end{aligned} \tag{A.134}$$

Because $p_k \geqslant 1$, then, by the REgroup assumption,

$$||\boldsymbol{\delta}_G||_2 \leqslant \sqrt{\sum_{k \in G(\boldsymbol{\beta}^*)} p_k (1 + p_k)^2 ||\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*||_2^2} \leqslant \frac{2||\mathbf{X}\boldsymbol{\delta}||_2}{\kappa\sqrt{n}} \leqslant \frac{8\sqrt{s}\gamma}{\kappa^2}. \tag{A.135}$$

Therefore, we have:

$$||\boldsymbol{\delta}||_2 \leqslant ||\boldsymbol{\delta}_G||_2 + ||\boldsymbol{\delta}_{G^c}||_2 \leqslant (2\sqrt{s} + 1) \frac{8\sqrt{s}\gamma}{\kappa^2}. \tag{A.136}$$

The inequalities (A.132), (A.133) and (A.136) complete proof of the theorem. □

Proof of Theorem 4.4

Here we present the proof of Theorem 4.4.

Proof. Let g_{kj} to be the subgradient for LES penalty with respect to β_{kj}, then:

$$g_{kj} = \lambda\alpha p_k \frac{\exp(\alpha|\hat{\beta}_{kj}|)}{\sum_{l=1}^{p_k}\exp(\alpha|\hat{\beta}_{kl}|)} * \partial|\hat{\beta}_{kj}|, \tag{A.137}$$

where $\partial|\hat{\beta}_{kj}| = sign(\hat{\beta}_{kj})$ if $\hat{\beta}_{kj} \neq 0$; and $\partial|\hat{\beta}_{kj}| \in [-1, 1]$ if $\hat{\beta}_{kj} = 0$.

Because the LES penalized OLS estimation (4.13) is convex, by KKT condition, event $\mathscr{O}$ holds if and only if the following two equations are satisfied:

$$\hat{\boldsymbol{\beta}}_G = \boldsymbol{\beta}^*_G + (\frac{1}{n}\mathbf{X}_G^T\mathbf{X}_G)^{-1}(\frac{1}{n}\mathbf{X}_G^T\boldsymbol{\epsilon} - \mathbf{g}_G), \tag{A.138}$$

$$\mathbf{g}_{G^c} = \frac{1}{n}\mathbf{X}_{G^c}^T\boldsymbol{\epsilon} + \frac{1}{n}\mathbf{X}_{G^c}^T\mathbf{X}_G(\frac{1}{n}\mathbf{X}_G^T\mathbf{X}_G)^{-1}(\mathbf{g}_G - \frac{1}{n}\mathbf{X}_G^T\boldsymbol{\epsilon}), \tag{A.139}$$

where $\mathbf{g}_G$ is a vector of subgradient g_{kj}'s with $k \in G$ and $\mathbf{g}_{G^c}$ is a vector of subgradient g_{kj}'s with $k \notin G$.

In order to prove the theorem, we only need to show the following two limits:

$$P(||\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G||_\infty < d_n) \to 1, \qquad n \to \infty; \tag{A.140}$$

and

$$P(||\mathbf{g}_{G^c}||_\infty < \lambda\alpha) \to 1, \qquad n \to \infty. \tag{A.141}$$

Recall $d_n = \min_{k \in G}||\boldsymbol{\beta}^*_k||_\infty$, therefore (A.140) implies $||\hat{\boldsymbol{\beta}}_k||_\infty > 0$, for all $k \in G$. When $||\hat{\boldsymbol{\beta}}_k||_\infty > 0$, if we let $|\hat{\beta}_{kj}| = ||\hat{\boldsymbol{\beta}}_k||_\infty > 0$, then:

$$|g_{kj}| = \frac{\lambda\alpha p_k \exp(\alpha|\hat{\beta}_{kj}|)}{\exp(\alpha|\hat{\beta}_{k1}|) + \ldots + \exp(\alpha|\hat{\beta}_{kp_k}|)} = \frac{\lambda\alpha p_k}{1 + \sum_{l \neq j}\exp\{\alpha(|\hat{\beta}_{kl}| - |\hat{\beta}_{kj}|)\}} \geqslant \lambda\alpha. \tag{A.142}$$

By inequality (A.142), we know (A.141) implies $||\hat{\boldsymbol{\beta}}_k||_\infty = 0$, for all $k \notin G$. In turn, (A.140) and (A.141) imply

$$P(\mathscr{O}) \to 1, \qquad n \to \infty,$$

as claimed. We will use equations (A.138) and (A.139) to show (A.140) and (A.141).

We will first show (A.140). For simplicity, we denote $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}_G^T\mathbf{X}_G$. Consider the p_0-dimensional vector $\mathbf{Z} = \frac{1}{n}\boldsymbol{\Sigma}^{-1}\mathbf{X}_G^T\boldsymbol{\epsilon}$. Then the expectation and covariance of $\mathbf{Z}$ are $E(\mathbf{Z}) = \mathbf{0}$ and $Var(\mathbf{Z}) = \frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}$. By Ledoux and Talagrand (2011), the maximum of a Gaussian vector is bounded by:

$$E(||\mathbf{Z}||_\infty) \leqslant 3\sigma\sqrt{\frac{\log p_0}{nc}}. \tag{A.143}$$

Notice that:

$$||\boldsymbol{\Sigma}^{-1}\mathbf{g}_G||_\infty \leqslant ||\boldsymbol{\Sigma}^{-1}||_\infty||\mathbf{g}_G||_\infty \leqslant \sqrt{p_0}||\boldsymbol{\Sigma}^{-1}||_2 * \lambda\alpha p_k \leqslant \frac{\sqrt{p_0}}{c}\lambda\alpha p_k. \tag{A.144}$$

By (A.143), (A.144) and Markov inequality, we have:

$$\begin{aligned} P(||\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_G^*||_\infty < d_n) &\leqslant \frac{E(||\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_G^*||_\infty)}{d_n} \\ &\leqslant \frac{1}{d_n}\Big\{E(||\mathbf{Z}||_\infty) + ||\boldsymbol{\Sigma}^{-1}\mathbf{g}_G||_\infty\Big\} \\ &\leqslant \frac{1}{d_n}\Big(3\sigma\sqrt{\frac{\log p_0}{nc}} + \frac{\sqrt{p_0}}{c}\lambda\alpha \max_{k\in G} p_k\Big), \end{aligned} \tag{A.145}$$

which goes to zero when $n \to \infty$ under assumption $(C2)$. Thus (A.140) is established.

Next we show (A.141). Let

$$\begin{aligned} \mathbf{g}_{G^c} &= \frac{1}{n}\mathbf{X}_{G^c}^T\mathbf{X}_G\boldsymbol{\Sigma}^{-1}\mathbf{g}_G + \frac{1}{n}\mathbf{X}_{G^c}^T(I - \mathbf{X}_G\boldsymbol{\Sigma}^{-1}\mathbf{X}_G^T)\boldsymbol{\epsilon} \\ &\triangleq \mathbf{u} + \mathbf{v}. \end{aligned} \tag{A.146}$$

Then for any $k \in G^c$, $\|\mathbf{g}_{G^c}\|_\infty \leqslant \|\mathbf{u}\|_\infty + \|\mathbf{v}\|_\infty$. By assumption $(C3)$, we have:

$$\begin{aligned} \|\mathbf{u}\|_\infty &\leqslant \|\mathbf{X}_{G^c}^T \mathbf{X}_G (\mathbf{X}_G^T \mathbf{X}_G)^{-1}\|_\infty \|\mathbf{g}_G\|_\infty \\ &\leqslant \lambda\alpha \max_{k \in G} p_k * \|\mathbf{X}_{G^c}^T \mathbf{X}_G (\mathbf{X}_G^T \mathbf{X}_G)^{-1}\|_\infty \\ &\leqslant \lambda\alpha(1-\tau). \end{aligned} \tag{A.147}$$

Notice that for any element v_i of $\mathbf{v}$, $v_i = \frac{1}{n} X_{kl}^T (I - \mathbf{X}_G \mathbf{\Sigma}^{-1} \mathbf{X}_G^T)\boldsymbol{\epsilon}$ for some column kl of matrix $\mathbf{X}$, $k \notin G$. Then we have $E(v_i) = 0$ and

$$Var(v_i) = \frac{\sigma^2}{n^2} X_{kl}^T \Big(I - \mathbf{X}_G \mathbf{\Sigma}^{-1} \mathbf{X}_G^T\Big) X_{kl} \leqslant \frac{\sigma^2}{n^2} \|X_{kl}\|_2^2 = \frac{\sigma^2}{n}. \tag{A.148}$$

By a similar argument we used in (A.143), we have:

$$E(\|\mathbf{v}\|_\infty) \leqslant 3\sigma \sqrt{\frac{\log(p-p_0)}{n}}. \tag{A.149}$$

By Markov's inequality,

$$P(\|\mathbf{v}\|_\infty > \frac{\tau}{2}\lambda\alpha) \leqslant \frac{6\sigma}{\tau\lambda\alpha} \sqrt{\frac{\log(p-p_0)}{n}}, \tag{A.150}$$

which goes to zero by assumption $(C4)$.

Combining (A.147) and (A.150), we have:

$$P(\|\mathbf{g}_{G^c}\|_\infty \geqslant (1 - \frac{\tau}{2})\lambda\alpha) \to 0, \quad n \to \infty, \tag{A.151}$$

which gives (A.141). This completes the proof. □

REFERENCES

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *International symposium on information theory, 2 nd, tsahkadsor, armenian ssr*, 267–281.

Berry, D.A., K.A. Cronin, S.K. Plevritis, D.G. Fryback, L. Clarke, M. Zelen, J.S. Mandelblatt, A.Y. Yakovlev, J.D.F. Habbema, and E.J. Feuer. 2005. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine* 353(17):1784–1792.

Bickel, P.J., Y. Ritov, and A.B. Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 37(4):1705–1732.

Breheny, Patrick, and Jian Huang. 2009. Penalized methods for bi-level variable selection. *Statistics and its interface* 2(3):369.

Campbell, A., P.E. Converse, and W.L. Rodgers. 1976. *The quality of american life: Perceptions, evaluations, and satisfactions*. Russell Sage Foundation.

Carrico, A.W., M.H. Antoni, R.E. Durán, G. Ironson, F. Penedo, M.A. Fletcher, N. Klimas, and N. Schneiderman. 2006. Reductions in depressed mood and denial coping during cognitive behavioral stress management with hiv-positive gay men treated with haart. *Annals of Behavioral Medicine* 31(2):155–164.

Craven, P., and G. Wahba. 1978. Smoothing noisy data with spline functions. *Numerische Mathematik* 31(4):377–403.

Dicker, L., B. Huang, and X. Lin. 2011. Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica*. To appear.

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

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

———. 2008b. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5): 849–911.

———. 2011. Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* 57(8):5467–5484.

Fan, J., and H. Peng. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32(3):928–961.

Figueiredo, M.A.T., R.D. Nowak, and S.J. Wright. 2007. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *Selected Topics in Signal Processing, IEEE Journal of* 1(4):586–597.

Folkman, S., and R.S. Lazarus. 1980. An analysis of coping in a middle-aged community sample. *Journal of health and social behavior* 219–239.

———. 1985. If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of personality and social psychology* 48(1):150.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2):302–332.

Friedman, J., T. Hastie, and R. Tibshirani. 2010a. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1): 1.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010b. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Furnari, Frank B, Tim Fenton, Robert M Bachoo, Akitake Mukasa, Jayne M Stommel, Alexander Stegh, William C Hahn, Keith L Ligon, David N Louis, Cameron Brennan, et al. 2007. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes & development* 21(21):2683–2710.

Gall, T.L., C. Charbonneau, N.H. Clarke, K. Grant, A. Joseph, and L. Shouldice. 2005. Understanding the nature and role of spirituality in relation to coping and health: A conceptual framework. *Canadian Psychology/Psychologie Canadienne* 46(2): 88.

Girard, A. 1989. A fast âĽ˜monte-carlo cross-validationâĽ™procedure for large least squares problems with noisy data. *Numerische Mathematik* 56(1):1–23.

Girard, D. 1987. Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille. *Rapport de Recherche RR*.

Goldstein, Alan A. 1964. Convex programming in hilbert space. *Bulletin of the American Mathematical Society* 70(5):709–710.

Harrison Jr, David, and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management* 5(1): 81–102.

Holmes, J.A., and C.A. Stevenson. 1990. Differential effects of avoidant and attentional coping strategies on adaptation to chronic and recent-onset pain. *Health Psychology* 9(5):577.

Huang, J., S. Ma, H. Xie, and C.H. Zhang. 2009. A group bridge approach for variable selection. *Biometrika* 96(2):339.

Huang, Jian, Shuangge Ma, and Cun-Hui Zhang. 2008. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 18(4):1603.

Hutchinson, MF. 1989. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* 18(3):1059–1076.

Kamitani, Hideki, Seijiro Taniura, Kenji Watanabe, Makoto Sakamoto, Takashi Watanabe, and Thomas Eling. 2002. Histone acetylation may suppress human glioma cell proliferation when p21waf/cip1 and gelsolin are induced. *Neuro-oncology* 4(2):95–101.

Ledoux, Michel, and Michel Talagrand. 2011. *Probability in banach spaces: isoperimetry and processes*, vol. 23. Springer.

Leng, C., Y. Lin, and G. Wahba. 2006. A note on the lasso and related procedures in model selection. *Statistica Sinica* 16(4):1273.

Levitin, Evgeny S, and Boris T Polyak. 1966. Constrained minimization methods. *USSR Computational mathematics and mathematical physics* 6(5):1–50.

Lin, X., G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. 2000. Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *Annals of Statistics* 1570–1600.

Lin, Y., and H.H. Zhang. 2006. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics* 34(5):2272–2297.

Lv, J., and Y. Fan. 2009. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* 37(6A):3498–3528.

Meinshausen, N., and P. Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3):1436–1462.

Nardi, Y., and A. Rinaldo. 2008. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* 2:605–633.

Parle, M., P. Maguire, and C. Heaven. 1997. The development of a training model to improve health professionals' skills, self-efficacy and outcome expectancies when communicating with cancer patients. *Social Science & Medicine* 44(2):231–240.

Purnell, J.Q., and B.L. Andersen. 2009. Religious practice and spirituality in the psychological adjustment of survivors of breast cancer. *Counseling and values* 53(3): 165–182.

Rottmann, N., S.O. Dalton, J. Christensen, K. Frederiksen, and C. Johansen. 2010. Self-efficacy, adjustment style and well-being in breast cancer patients: a longitudinal study. *Quality of Life Research* 19(6):827–836.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics* 6(2): 461–464.

Shen, X., W. Pan, and Y. Zhu. 2011. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*. To appear.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani. 2012. A sparse-group lasso. *Journal of Computational and Graphical Statistics, DOI* 10(10618600.2012):681250.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tseng, P. 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications* 109(3):475–494.

Wahba, G. 1983. Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)* 133–150.

———. 1985. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics* 1378–1402.

Wahba, G., DR. Johnson, F. Gao, and J. Gong. 1995. Adaptive tuning of numerical weather prediction models: randomized gcv in three-and four-dimensional data assimilation. *Monthly Weather Review* 123(11):3358–3369.

Wang, H., R. Li, and C.L. Tsai. 2007a. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3):553–568.

Wang, Lifeng, Guang Chen, and Hongzhe Li. 2007b. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics* 23(12): 1486–1494.

Wu, T.T., and K. Lange. 2008. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* 2(1):224–244.

Yang, Y. 2005. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika* 92(4):937–950.

Yuan, M., and Y. Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhang, C.H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2):894–942.

Zhang, C.H., and J. Huang. 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36(4):1567–1594.

Zhao, P., G. Rocha, and B. Yu. 2006. Grouped and hierarchical model selection through composite absolute penalties. *Department of Statistics, UC Berkeley, Tech. Rep* 703.

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

Zheng, Duo, Yong-Yeon Cho, Andy TY Lau, Jishuai Zhang, Wei-Ya Ma, Ann M Bode, and Zigang Dong. 2008. Cyclin-dependent kinase 3–mediated activating transcription factor 1 phosphorylation enhances cell transformation. *Cancer research* 68(18):7650–7660.

Zhou, N., and J. Zhu. 2010. Group variable selection via a hierarchical lasso and its oracle property. *Arxiv preprint arXiv:1006.2871*.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301–320.

Zwahlen, D., N. Hagenbuch, M.I. Carley, J. Jenewein, and S. Buchi. 2010. Posttraumatic growth in cancer patients and partnersï¿½effects of role, gender and the dyad on couples' posttraumatic growth experience. *Psycho-Oncology* 19(1):12–20.