Methodologies and Algorithms on Some Non-convex Penalized Models for Ultra High Dimensional Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Bo Peng

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor: Lan Wang

July, 2016

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Lan Wang for the consistent support of my Ph.D. study and research, for her patience, enthusiasm and immense knowledge. Her guidance comes through all the time of my research and composition of this thesis. I could not be more appreciated and fortunate to have her as my advisor and mentor during my Ph.D. career.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Galin Jones, Professor Yuhong Yang and Professor Baolin Wu, for their precious advice, genuine encouragement and insightful comments. Meanwhile, my sincere thanks also goes to my colleagues Yuwen Gu, Fan Yang, Yi Yang, Professor Yichao Wu and Professor Runze Li. I am enormously grateful to their tremendous amount of help from both research and life.

I want to also thank my parents, Guoyi Peng and Junqi He, for giving birth to me at the first place and supporting me spiritually throughout my life. Without them, I would not be able to conquer all the hardness and difficulties in science exploration.

Last but not the least, I would like to express my special thank to my girlfriend, Miki (Yuan Liu), for every single day, single hour and single second during past four years. She is my shield protecting me from distress and frustration and is my sword to disperse darkness and gloom in my Ph.D. career. May our love be eternal and sublime in the rest of our lives.

Dedication

To my parents who nurse me with affections and love and their dedicated partnership for success in my life

To my lover who surrounds me with pure flame and light of love and our promise for lifelong companions

Abstract

In recent years, penalized models have gained considerable importance on dealing with variable selection and estimation problems under high dimensional settings. Of all the candidates, the l_1 penalized, or the LASSO model retains popular application in diverse fields with sophisticated methodology and mature algorithms. However, as a promising alternative of the LASSO, non-convex penalized methods, such as the smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) methods, produce asymptotically unbiased shrinkage estimates and owns attractive advantages over the LASSO. In this thesis, we propose intact methodology and theory for multiple non-convex penalized models. The proposed theoretical framework includes estimator's error bounds, oracle property and variable selection behaviors. Instead of common least square models, we focus on quantile regression and support vector machines (SVMs) for exploration of heterogeneity and binary classification. Though we demonstrate current local linear approximation (LLA) optimization algorithm possesses those nice theoretical properties to achieve the oracle estimator in two iterations, the computation issue is highly challenging when p is large due to the non-smoothness of the loss function and the non-convexity of the penalty function. Hence, we also explore the potential of coordinate descent algorithms for fitting selected models, establishing convergence properties and presenting significant speed increase on current approaches. Simulated and real data analysis are carried out to examine the performance of non-convex penalized models and illustrate the outperformance of our algorithm in computational speed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

High dimensional data are frequently collected in a large variety of research areas such as genomics, functional magnetic resonance imaging, tomography, economics and finance. Analysis of high-dimensional data poses many challenges for statisticians and calls for new statistical methodologies and theories [1, 2]. We consider the ultra-high dimensional regression setting in which the number of covariates p grows at an exponential rate of the sample size n.

When the primary goal is to identify the underlying model structure, a popular approach for analyzing ultra-high dimensional data is to use the regularized regression. For example, [3] proposed the Dantzig selector; [4] proposed weighted l_1-minimization to enhance the sparsity of the Dantzig selector; [5] considered the adaptive lasso when a zero-consistent initial estimator is available; [6] demonstrated the smoothly clipped absolute deviation (SCAD) penalty and investigated non-concave penalized likelihood with ultra-high dimensionality; and [7] proposed a minimax concave penalty (MCP) for penalized regression.

The LASSO penalized regression is computationally attractive and enjoys great performance in prediction. However, it is known that LASSO requires rather stringent conditions on the design matrix to be variable selection consistent [8, 9]. Focusing on identifying the unknown sparsity pattern, non-convex penalized high-dimensional regression has recently received considerable attention. [6] first systematically studied

non-convex penalized likelihood for fixed finite dimension p. In particular, they recommended the SCAD penalty which enjoys the oracle property for variable selection. That is, it can estimate the zero coefficients as exact zero with probability approaching one, and estimate the nonzero coefficients as efficiently as if the true sparsity pattern is known in advance. [10] extended these results by allowing p to grow with n at the rate $p = o(n^{1/5})$ or $p = o(n^{1/3})$. For high dimensional non-convex penalized regression with $p \gg n$, [11] proved that the oracle estimator itself is a local minimum of SCAD penalized least squares regression under very relaxed conditions; [7] devised a novel PLUS algorithm which when used together with MCP can achieve the oracle property under certain regularity conditions. Important insight has also been gained through the recent work on theoretical analysis of the global solution [12, 13].

Although non-convex penalized least square approach is useful, researchers try to extend this methodology to other regression or learning models for specific problem solving. [14] regularized the quantile regression with a non-convex penalty function to deal with ultra-high dimensionality; [6] demonstrated the non-convex penalty can ameliorate the bias problems of LASSO in general linear models; [15] considered the information bound of the oracle estimator by using non-convex regularized Cox's proportional hazard model; [16] also dealt with the Fisher consistency and the oracle property of support vector machines (SVMs) with the SCAD penalty for fixed p case. In my Ph.D. research, I focus on the two models amongst the above all: quantile regression and support vector machine, whose loss functions share similar convexity and non-smoothness properties. Hence it becomes natural to employ the non-convex penalty methodology to these two models and establish theoretical frameworks under ultra-high dimensional settings.

The computation for non-convex penalized methods is much more complicated than the LASSO, because the resulting optimization problem is usually non-convex and will even become non-smooth when the loss function is quantile or hinge loss, respectively in quantile regression and SVMs. Several algorithms have been developed for computing the non-convex penalized estimators. [2] worked out the local quadratic approximation (LQA) algorithm as a unified method for computing the non-convex penalized maximum likelihood. [17] proposed the local linear approximation (LLA) algorithm which turns a concave penalized problem into a series of reweighed l_1 penalized problems. Both LQA and LLA are related to the MM principle [18, 19]. However, the computational speed is

considerably slow when $p \gg n$ even with the algorithm advances aforementioned. For non-convex penalized least squares regression, the coordinate descent algorithm, which are commonly accepted as a much faster alternative for other competing methods, were recently investigated by [20, 21]. [22] proposed a new coordinate descent algorithm for non-convex penalized generalized linear models which enjoys the appealing property of avoiding the computation of a scaling factor in each update of the solutions. These algorithms are very effective in large-scale problems but do not apply to non-convex penalized quantile regression or SVMs.

Here we outline a complete theoretical framework for non-convex penalty methodology in both quantile regression and SVMs models. Variable selection consistency and oracle property are studied under weak regular conditions. Estimation of error bounds of l_1 penalized models offers an effective searching for good initial values in LLA algorithms to solve these non-convex optimization problems. Generally, the LLA algorithm can find the oracle estimator in two iterations even under ultra-high dimensional settings for both two non-convex penalized models. Furthermore, a new improved coordinate descent algorithm is invented to solve the non-convex penalized quantile regression model and owns overwhelming speed advantage in simulated and real data cases. Extension of this algorithm to other models will stay active in my future research works.

1.2 Sparse penalized models

We present a general introduction to sparse penalized models and display necessary notations in this article.

The inputs we have are:

- A random sample $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ from an unknown distribution $P(\mathbf{X}, Y)$. We write $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\intercal \in \mathbb{R}^{p+1}$, where $X_{i0} = 1$ corresponds to the intercept term and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\intercal$ with $Y_i \in \mathbb{R}$, for two-class classification $Y_i \in \{\pm 1\}$.
- Let $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^\intercal$ denote the feature design matrix.
- A convex nonnegative loss functional $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.
- A nonnegative penalty functional $P : \mathbb{R}^p \to \mathbb{R}$, with $p(0) = 0$.

Consider estimating a sparse vector of coefficients $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}_-)^T)^T$ with $\boldsymbol{\beta}_- = (\beta_1, \beta_2, \ldots, \beta_p)^\intercal$ based on training data, through penalized empirical loss minimization

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}_- \in \mathbb{R}^p}{\operatorname{argmin}} \left[L(\mathbf{y}, \mathcal{X}\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}) \right] \tag{1.1}$$

where $\lambda > 0$ is the regularization parameter; $\lambda = 0$ corresponds to no regularization and $\lim_{\lambda \to \infty} \widehat{\boldsymbol{\beta}}_-(\lambda) = 0$. Generally for a given vector $\mathbf{e}$, we use $\mathbf{e}_-$ to denote the subvector with the first entry of $\mathbf{e}$ omitted. In what follows, for any set $\mathcal{A} \subset \{1, 2, \ldots, p\}$ and vector $\mathbf{e} \in \mathbb{R}^p$, let $e_{\mathcal{A}}$ denote the p dimensional vector such that we only keep the coordinates of $\mathbf{e}$ when their indices are in $\mathcal{A}$ and replace others by 0. One often needs to compute the solution at a fine grid of λ's in order to pick a data-driven optimal λ for fitting a 'best' final model.

$P(\boldsymbol{\beta})$ is a sparsity-inducing penalty to produce a sparse estimator, which is especially preferred when $p \gg n$. Some widely used regularization methods include the LASSO, the elastic net and the grouped LASSO penalty.

The LASSO [23] is a very popular technique for high-dimensional modeling

$$\lambda P(\boldsymbol{\beta}) = \lambda ||\boldsymbol{\beta}_-||_1.$$

LASSO yields sparse estimates of $\boldsymbol{\beta}$ because it shrinks small lease squares estimates $\widehat{\beta}_j^{ols}$'s toward exact zero. [24] proposed the elastic net penalty as an improved variant of the LASSO for high-dimensional data when predictors are highly correlated. It connects the LASSO penalty and the ridge penalty

$$\lambda P(\boldsymbol{\beta}) = \lambda ||\boldsymbol{\beta}_-||_1 + \frac{1}{2}\lambda_2 ||\boldsymbol{\beta}_-||_2 \ (\lambda_2 > 0).$$

As mentioned above, two commonly used non-convex penalties are the SCAD penalty and the MCP. The SCAD penalty function [2] is defined by

$$\begin{aligned} p_\lambda(|\beta|) &= \lambda|\beta| I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } \ a > 2. \end{aligned}$$

The MCP [7] function has the form

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \ge a\lambda), \quad \text{for some } \ a > 1.$$

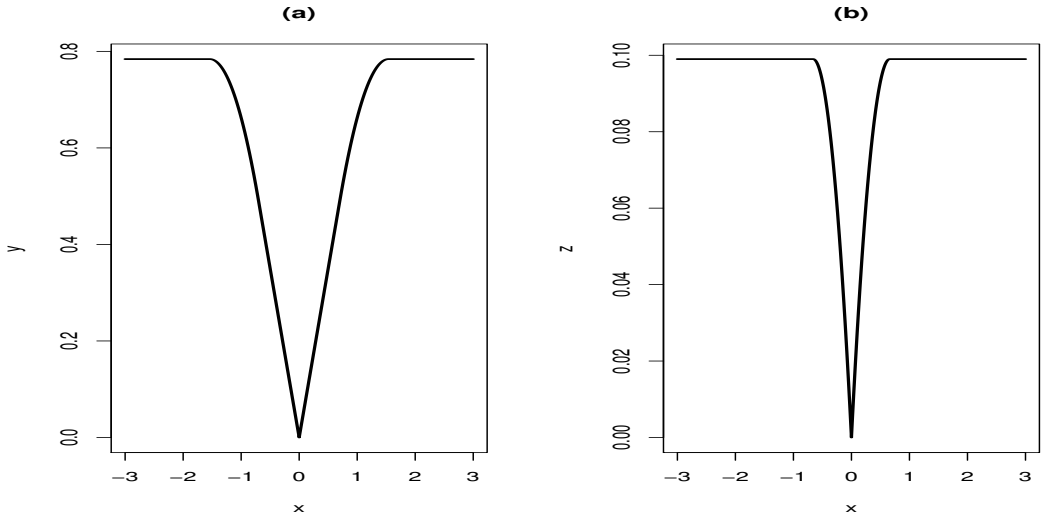

Figure 1.1: *(a) SCAD penalty function with* $\lambda = 0.7$,$a = 2.2$*; (b) MCP penalty function with* $\lambda = 0.3$*,* $a = 2.2$.

See Figure 1.2 for an illustration of the shape of the two penalty functions. In such a case, we have $\lambda P(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\lambda(|\beta_j|)$.

Many "modern" machine learning methods can be cast in the framework of penalized optimization [25]. In penalized regression problems, the loss function takes the form

$$L(\mathbf{y}, \mathcal{X}\boldsymbol{\beta}) = \sum_i l(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta})$$

where the residuals $Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta}$ quantifies the discrepancy between an observation Y_i and a linear predictor $\mathbf{X}_i^\intercal \boldsymbol{\beta}$. An example is the lasso penalized least squares:

$$\text{Least Squares :} \qquad \widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\frac{1}{2n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}_-||_1 \right]$$

In classification problems,

$$L(\mathbf{Y}, \mathcal{X}\boldsymbol{\beta}) = \sum_i l(Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})$$

where $(Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})$ are margins for classification. Examples are the lasso penalized logistic regression and support vector machine using hinge loss or squared hinge loss or

Huberized squared hinge loss with the lasso penalty.

$$\text{Logistic}: \quad \widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}}\right) + \lambda ||\boldsymbol{\beta}_-||_1$$

$$\text{l}_1 \text{ norm SVM}: \quad \widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+ + \lambda ||\boldsymbol{\beta}_-||_1$$

$$\text{Squared SVM}: \quad \widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+^2 + \lambda ||\boldsymbol{\beta}_-||_1$$

$$\text{Huberized SVM}: \quad \widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} H_c(Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}_-||_1$$

$$\text{where } H_c(t) = \begin{cases} 0 & t > 1 \\ (1-t)^2/2\delta & 1 - \delta < t \leq 1 \\ 1 - t - \delta/2 & t \leq 1 - \delta \end{cases}$$

1.3 Thesis outline

This thesis consists of preliminary research achievements in Chapter 2 and our two papers to be found in Chapters 3-4.

In Chapter 2, we consider the penalized quantile regression model under ultra-high dimensional settings. In this framework, we demonstrate the fundamental methodology and theory for non-convex penalty in [14, 26]. [27] provide an error bound estimation of l_1 penalized quantile regression estimator with given regular conditions . Similar estimation is also displayed in [28, 29]. Meanwhile, [14] prove that the oracle estimator is among the local minima of non-convex penalized quantile objective function even for ultra-high dimensional cases. Furthermore, [26] show that LLA algorithm initiated by l_1 penalized quantile regression estimator is able to obtain oracle estimator in two iterations with proper λ selected. In this section, a complete theoretical framework and computation methodology has been established for non-convex penalized quantile regression model.

In Chapter 3, we still work on the non-convex penalized quantile regression model but focus on its real application . To tackle the computational difficulty of LLA algorithm for high dimensional data, we introduce the QICD (iterative coordinate descent algorithm) as an alternative with high computational speed. QICD algorithm combines

the idea of MM (majorization and minimization) to transform the non-convex optimization problem to a series of convex ones and the potential of coordinate descent algorithm to increase the iteration speed. Under regular conditions, QICD are proved to converge to local minima for non-convex penalized quantile regression model. Simulation and real data examples are presented for comparison with LLA algorithm.

In Chapter 4, we extend the complete theory and methodology framework built before to non-convex penalized SVMs. Some shared properties of quantile regression and SVMs motivate us to explore the error bounds and oracle property under ultra-high dimensional settings. The effectiveness of LLA algorithm for non-convex penalized SVMs is studied to achieve oracle estimator in practice. Again, numerical experiments are presented in the end.

Finally, Chapter 5 discusses some potential future work in other different research fields.

Chapter 2

Non-convex Penalized Quantile Regression Model

2.1 Chapter Overview

In this chapter, we outline the theoretical framework of non-convex penalized quantile regression model in [26, 28]. Respectively, [28] provide a tight estimation of error bounds for l_1 penalized quantile regression estimator; and [26] show explicitly how the LLA algorithm can achieve the oracle estimator for non-convex penalized quantile regression model. Their efforts inspire us to explore the possibility of popularizing this framework to other non-convex penalized models and the development of potential on suitable algorithms in practice. Numerical results illustrate the advantage of non-convex penalty on both variable selection consistency and oracle estimator solution over LASSO.

2.2 Introduction

It is common to observe that real life ultra-high dimensional data displays heterogeneity due to either heteroscedastic variance or other forms of non-location-scale covariate effects. This type of heterogeneity is often of scientific importance but tends to be overlooked by exiting procedures which mostly focus on the mean of the conditional distribution. Furthermore, despite significant recent developments in ultra-high dimensional regularized regression, the statistical theory of the existing methods generally

requires conditions substantially stronger than those usually imposed in the classical $p < n$ framework. These conditions include homoscedastic random errors, Gaussian or near Gaussian distributions, and often hard-to-check conditions on the design matrix, among others. These two main concerns motivate us to study nonconvex penalized quantile regression in ultra-high dimension.

Quantile regression [30] has become a popular alternative to least squares regression for modeling heterogeneous data. [31, 32, 33] established nice asymptotic theory for high-dimensional M-regression with possibly nonsmooth objective functions. Their results apply to quantile regression (without the sparseness assumption) but require that $p = o(n)$.

To deal with the ultra high dimensionality, we regularize quantile regression with a non-convex penalty function, such as the SCAD penalty and the MCP. The choice of non-convex penalty is motivated by the well-known fact that directly applying the l_1 penalty tends to include inactive variables and to introduce bias. We advocate a more general interpretation of sparsity which assumes that only a small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may be different when we consider different segments of the conditional distribution. By considering different quantiles, this framework enables us to explore the entire conditional distribution of the response variable given the ultra-high dimensional covariates. In particular, it can provide a more realistic picture of the sparsity patterns, which may differ at different quantiles.

Regularized quantile regression with fixed p was recently studied by [34, 35, 36, 37]. Their asymptotic techniques, however, are difficult to extend to the ultra-high dimension. For high dimensional p, [27] recently derived a nice error bound for quantile regression with the l_1-penalty. They also showed that a post-l_1-quantile regression procedure can further reduce the bias. However, in general post-l_1-quantile regression does not possess the oracle property.

The main technical challenge of our work is to deal with both the non-smooth loss function and the non-convex penalty function in ultra-high dimension. This non-convex optimization problem usually contains multiple local minimizers. [17] has proposed the LLA algorithm to at least compute a local solution of the non-convex penalized problem. However, before declaring that the non-convex penalty is superior to LASSO,

we need to solve a missing puzzle in this picture. The oracle property of non-convex penalized quantile regression is established on a theoretical local solution. We need to prove that the employment of LLA algorithm is able to find such a local optimal solution possessing desired theoretical properties. Many have tried to address this issue [7, 6, 13]. The basic idea is to find conditions under which the non-convex penalized problem has a unique minimizer and hence eliminate the difficulty of multiple minimizers. Though this method is natural and intuitive, the imposed conditions for unique minimizer are always too stringent to be realistic.

In this chapter, we illustrate a direct approach in [26] to tackle the multiple local minimizers issue. We present a general procedure based on LLA algorithm for computing non-convex penalized quantile regression problem and derive a lower bound on the probability that this specific solution is equal to the oracle estimator. This probability lower bound equals to $1-\delta_0-\delta_1-\delta_2$ where δ_0 corresponds to the exception probability of the localizability of the underlying model, δ_1 and δ_2 represent the exception probability of the regularity of the oracle estimator and they are irrelevant to any actual estimation method. Explicit expressions of δ_0, δ_1 and δ_2 are demonstrated in Section 2.3. Under weak regular conditions, δ_1 and δ_2 are very small. In a sense, if δ_0 goes to zero then the LLA algorithm can find the oracle estimator with overwhelming probability. Hence, this theory suggests that as long as a reasonable initial estimator is given, the LLA algorithm is able to deliver an oracle estimator for non-convex penalized quantile regression model. In addition, once the oracle estimator is obtained, the LLA algorithm converges in the next iteration. Furthermore, we also display how to prove all exception probabilities δ_0, δ_1 and δ_2 go to zero at a fast rate under the ultra-high dimensional setting where $\log p = O(n^\eta)$ for some $\eta \in (0, 1)$.

Throughout this chapter the following notations are used. For $\mathbf{U} = (u_{ij})_{k\times l}$, denote $||\mathbf{U}||_{min} = \min_{(i,j)} |u_{ij}|$ as its minimal absolute value, and let $\lambda_{min}(\mathbf{U})$ and $\lambda_{max}(\mathbf{U})$ be its smallest and largest eigenvalues respectively. We also use some matrix norm: the l_1 norm $||\mathbf{U}||_{l_1} = \max_j \sum_i |u_{ij}|$, the l_2 norm $||\mathbf{U}||_{l_2} = \lambda^{1/2}_{max}(\mathbf{U}^\intercal\mathbf{U})$, the l_∞ norm $||\mathbf{U}||_{l_\infty} = \max_i \sum_j |u_{ij}|$, the entrywise l_1 norm $||\mathbf{U}||_1 = \sum_{(i,j)} |u_{ij}|$ and the entrywise l_∞ norm $||\mathbf{U}||_\infty = \max_{(i,j)} |u_{ij}|$.

2.3 Oracle Property of Non-convex Penalized Quantile Regression Estimator

2.3.1 The Methodology

Let us begin with the notation and statistical setup. Suppose that we have a random sample $\{Y_i, X_{i1}, \ldots, X_{ip}\}$, $i = 1, \ldots, n$, from the following model:

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \triangleq \mathbf{X}_i^\intercal \boldsymbol{\beta} + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1)$-dimensional vector of parameters, $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^\intercal$ with $X_{i0} = 1$, and the random errors ϵ_i satisfy $P(\epsilon_i \leq 0 | \mathbf{x}_i) = \tau$ for some specified $0 < \tau < 1$. Let $f_i(\cdot)$ be the density function and $F_i(\cdot)$ be the density function and distribution function of ϵ_i respectively. The case $\tau = 1/2$ corresponds to median regression. The number of covariates $p = p_n$ is allowed to increase with the sample size n. It is possible that p_n is much larger than n. Moreover, in this thesis we use the following notation for vector norms: for $\mathbf{x} \in \mathbb{R}^k$ and $q \geq 1$ is a real number, we define $||\mathbf{x}||_q = \left(\sum_{i=1}^k |x_i|^q\right)^{1/q}$, $||\mathbf{x}||_\infty = \max(|x_1|, \ldots, |x_k|)$ and $||\mathbf{x}||_0 = \sum_{i=1}^k I(x_i \neq 0)$.

The true parameter value $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^* \ldots, \beta_{p_n}^*)^\intercal$ is assumed to be sparse; that is, the majority of its components are exactly zero. Let $A = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients. Let $|A| = q_n = q$ be the cardinality of the set A, which is allowed to increase with n. The sparsity assumption means that $q \ll p$.

We consider the following penalized quantile regression model

$$Q(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(|\beta_j|), \tag{2.2}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function (or check loss function), and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. For convenience, we denote $l_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta})$ and $P_\lambda(|\boldsymbol{\beta}|) = \sum_{j=1}^p p_\lambda(|\beta_j|)$. The tuning parameter λ controls the model complexity and goes to zero at an appropriate rate. The penalty function $p_\lambda(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $p'_\lambda(t)$ on $(0, +\infty)$.

Assume that an oracle knows knows the true support set A of the underlying model,

and the oracle estimator is defined as

$$\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_A, \mathbf{0}) = \underset{\boldsymbol{\beta}:\boldsymbol{\beta}_{A^c}=\mathbf{0}}{\operatorname{argmin}}\, l_n(\boldsymbol{\beta}) \tag{2.3}$$

We assume throughout the thesis that the problem is regular such that the oracle solution is unique, satisfying

$$\nabla_j l_n(\tilde{\boldsymbol{\beta}}) = 0, \quad \forall j \in A \tag{2.4}$$

where ∇_j is the subgradient with respect to the $j-$th component of $\boldsymbol{\beta}$. The oracle estimator is not a feasible estimator as A is always unknown in practice but it can be used as a theoretical benchmark for other estimators to compare with. An estimator is said to have the oracle property if it has the same asymptotic properties as the oracle estimator [2, 10]. In addition, an estimator is said to have the strong oracle property if the estimator equals the oracle estimator with overwhelming probability [6].

It is well known that penalized regression with the convex l_1 penalty tends to over-penalize large coefficients and to include spurious variables in the selected model. This may not be of much concern for predicting future observations, but is nonetheless undesirable when the purpose of the data analysis is to gain insights into the relationship between the response variable and the set of covariates. Non-convex penalty has been put on board to overcome those issues of l_1 penalty. However, we still need impose some general conditions on our non-convex penalty function $P_\lambda(|t|)$.

(i) $P_\lambda(t)$ is increasing and concave in $t \in [0, +\infty)$ with $P_\lambda(0) = 0$;

(ii) $P_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $P'_\lambda(0) := P'_\lambda(0+) \geq a_1\lambda$;

(iii) $P'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;

(iv) $P'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with the pre-specified constant $a > a_2$.

where a_1 and a_2 are two fixed positive constants. The above definition contains the well-know SCAD and MCP penalties. The derivative of the SCAD penalty is

$$P'_\lambda(t) = \lambda I(t \leq \lambda) + \frac{(a\lambda - t)_+}{a-1} I(t > \lambda), \quad \text{for some } a > 2,$$

and the derivative of the MCP is

$$P'_\lambda(t) = (\lambda - \frac{t}{a})_+, \quad \text{for some } a > 1.$$

Algorithm 1 The local linear approximation (LLA) algorithm

1. Initialize $\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}^{initial}$ and compute the adaptive weight

$$\widehat{\mathbf{w}}^{(0)} = (\widehat{w}_1^{(0)}, \ldots, \widehat{w}_p^{(0)}) = (P'_\lambda(|\widehat{\beta}_1^{(0)}|), \ldots, P'_\lambda(|\widehat{\beta}_1^{(0)}|))'.$$

2. For $m = 1, 2, \ldots$, repeat the LLA iteration till convergence

   2.1. Obtain $\widehat{\boldsymbol{\beta}}^{(m)}$ by solving the following optimization problem

$$\widehat{\boldsymbol{\beta}}^{(m)} = \min_{\beta} l_n(\boldsymbol{\beta}) + \sum_j \widehat{w}_j^{m-1} \cdot \beta_j,$$

   2.2. Update the adaptive weight vector $\widehat{\mathbf{w}}^{(m)} = P'_\lambda(|\widehat{\beta}_j^{(m)}|)$.

It is easy to see that $a_1 = a_2 = 1$ for the SCAD, and $a_1 = 1 - a^{-1}$, $a_2 = 1$ for the MCP.

Till now, we confront the major problem: whether non-convex penalized estimator owns the oracle property, or even strong oracle property for quantile regression model? As a response, we claim in the following, although the estimator is defined via non-convex penalized problem, the computed estimator will possess oracle property even it is just a local solution. In a sense, it is absolutely fine that the computed local solution is not a global minimizer, which is not in most cases, as long as it has the optimal or desired statistical properties. We focus on one typical solution which is achieved by LLA algorithm [17]. Basically, the LLA algorithm takes advantage of the special non-convex structure of penalty functions and utilizes the majorization and minimization (MM) trick to transform non-convex optimization problem to a sequence of weighted l_1 penalized problems. In each iteration of LLA algorithm, the underlying local linear approximation is the best convex majorization of the non-convex penalty function (see Theorem 2 of [17]). Furthermore, the MM trick also provides theoretical guarantee to the convergence of the LLA algorithm to a stationary point of non-convex penalized problem (2.2).

Here, we display the details of the LLA algorithm as in Algorithm 1.

2.3.2 Asymptotic Properties

In the following section, we demonstrate the asymptotic analysis of the LLA algorithm for obtaining the oracle estimator in the non-convex penalized quantile regression problem if it is initiated by some appropriate initial estimator. Note that the check loss function $\rho_\tau(\cdot)$ is convex but non-differentiable. Thus we need to handle the subgradient $\nabla l_n(\boldsymbol{\beta}) = (\nabla_1 l_n(\boldsymbol{\beta}), \ldots, \nabla_p l_n(\boldsymbol{\beta}))$, where

$$\nabla_j l_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_i X_{ij} \cdot ((1-\tau) I(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta} > 0) - z_j I(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta} = 0) - \tau I(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta} < 0))$$

with $z_j \in [\tau - 1, \tau]$ is the subgradient of $\rho_\tau(u)$ when $u = 0$. Denote $M_A = \max_i \frac{1}{q}||\mathbf{X}_{iA}||_2$ and $m_{A^c} = \max_{(i,j):j\in A^c} |X_{ij}|$. Define $\delta_0 = \Pr(||\widehat{\boldsymbol{\beta}}^{initial} - \boldsymbol{\beta}^*||_2 > a_0\lambda)$, where $a_0 = \min(1, a_2)$. We have the following theorem in [26].

Theorem 2.3.1. *Suppose*

(1) there exist constant u_0 and $0 < f_{min} \le f_{max} < \infty$ such that for any u satisfying $|u| \le u_0$, $f_{min} \le \min_i f_i(u) \le \max_i f_i(u) \le f_{max}$.

If $\lambda = o(\frac{1}{n})$ such that $\log p = o(n\lambda^2)$, $(M_A q)^{1/2}(||\boldsymbol{\beta}_A^*||_{min} - a\lambda) \le u_0$, and $m_{A^c} M_A q = o(\frac{n^{1/2}\lambda}{\log^{1/2} n})$, the LLA algorithm initiated by $\widehat{\boldsymbol{\beta}}^{iniital}$ converges to $\tilde{\boldsymbol{\beta}}$ after two iterations with probability at least $1 - \delta_0 - \delta_1 - \delta_2$, where

$$\begin{aligned}
\delta_1 &= 4n^{-\frac{1}{2}} + C_1(p-q)\cdot\exp(-\frac{a_1 n\lambda}{104 m_{A^c}}) + 2(p-q)\cdot\exp(-\frac{a_1^2 n\lambda^2}{32 m_{A^c}^2}), \quad \textit{and} \\
\delta_2 &= 4\exp(-\frac{\lambda_{min}^2 f_{min}^2}{72 M_A}\cdot\frac{n}{q}\left(||\boldsymbol{\beta}_A^*||_{min} - a\lambda\right)^2)
\end{aligned}$$

with $\lambda_{min} = \lambda_{min}(\frac{1}{n}\mathcal{X}_A^\intercal \mathcal{X})$ and $C_1 > 0$ that does not depend on n, p or q.

Under fairly weak assumptions, both δ_1 and δ_2 go to zero very quickly. To bound δ_0, we need to search for an appropriate initial estimator in practice to activate the LLA algorithm even under ultra high dimensional settings. As we mentioned before, we recommend the l_1 penalized quantile regression estimator as the initial value, *i.e.*, a proper bound for δ_0. We will discuss this topic and summarize the main results in the following chapter.

2.4 l_1 Penalized Quantile Regression

In the above section, we have confirmed that the LLA algorithm successfully capture the oracle estimator in non-convex penalized quantile regression model if a proper bound for δ_0 is given. For simplification, we propose the following l_1 penalized quantile regression estimator when $\tau = 0.5$,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}||Y - \mathcal{X}\boldsymbol{\beta}||_1 + \lambda||\boldsymbol{\beta}_-||_1 \tag{2.5}$$

Since our theory can be easily extended to all other quantile, we will use $\tau = 0.5$ in this section. Our purpose is to explore the properties of the estimator $\widehat{\boldsymbol{\beta}}$ whether a good error bound can be estimated. [28] presents the analysis of the penalized median regression estimator and discuss the selection of the penalty level λ, which does not depend on any unknown parameters or the noise distribution. This feature inherits the similar property of λ in our oracle property discussion hence the l_1 penalized quantile regression estimator is able to seal the theoretical gap smoothly for ultra high dimensional data. Actually, we will show that the estimator $\widehat{\boldsymbol{\beta}}$ owns surprisingly good properties. The major results contains two parts:

1. We propose a penalty level, it is simply

   $$\lambda = c\sqrt{\frac{2A(\alpha)\log p}{n}}$$

   where $c > 1$ is a constant, α is a chosen small probability, and $A(\alpha)$ is a constant such that $2p^{-(A(\alpha)-1)} \leq \alpha$. In practice, we choose $\lambda = \sqrt{\frac{q\log p}{n}}$, which matches the one we used in the above analysis. This choice of penalty is universal and we only assume that the noises have median 0 and $P(\epsilon = 0) = 0$ for all i.

2. We show that with high probability, the estimator has the error bound with high probability

   $$||\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||_2 = O(\sqrt{\frac{q\log p}{n}})$$

It is important to notice that we do not have any assumptions on the moments of the noise, we only need a scale parameter to control the tail probability of the noise. Actually, even for Cauchy distributed noise, where the first moment does not exist, the results still hold.

2.4.1 Choice of Penalty

In this section, we discuss the penalty level for the l_1 penalized estimator and answer the motivation to choose this λ for this problem. Still, we use $l_n(\boldsymbol{\beta}) = \frac{1}{n}||Y - \mathcal{X}\boldsymbol{\beta}||_1$. An important quantity to determine the penalty level is the sub-differential of l_n evaluated at the point of true coefficient $\tilde{\boldsymbol{\beta}}$. Here we just assume the error ϵ_i satisfying $P(\epsilon = 0) = 0$ and the median of ϵ_i is 0 for $i = 1, 2, \ldots, n$. Assume that $\epsilon_i \neq 0$ for all i, then the sub-differential of $l_n(\boldsymbol{\beta})$ at point $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ can be written as

$$S = \frac{1}{n}\mathcal{X}^\intercal(sign(\epsilon_1), sign(\epsilon_2), \ldots, sign(\epsilon_n))^\intercal$$

where $sign(x)$ denote the sign of x, *i.e.* $sign(x) = 1$ if $x > 0$, $sign(x) = -1$ if $x < 0$ and $sign(0) = 0$. Let $I = sign(\boldsymbol{\epsilon})$, then $I = (I_1, I_2, \ldots, I_n)^\intercal$ where $I_i = sign(\epsilon_i)$. Since ϵs are independent and have median 0, we have that $P(I_i = 1) = P(I_i = -1) = 0.5$ and I_is are independent.

The sub-differential of $l_n(\boldsymbol{\beta})$ at the point of $\tilde{\boldsymbol{\beta}}$, $S = \mathcal{X}^\intercal I$, summarizes the estimation error in the setting of the linear regression model. We will choose a penalty λ that dominates the estimation error with large probability. The principle of selecting the penalty λ is motivated by [27, 29, 38]. The intuition of this choice is that when the true coeffients $\tilde{\boldsymbol{\beta}}$ is a vector of 0, then the estimator should also be 0 with a given high probability. This is a general principle of choosing the penalty and can be applied to many other problems. Specifically, we choose a penalty λ such that it is greater than the maximum absolute value of S with high probability, *i.e.* we need to find a penalty level λ satisfying

$$P(\lambda \geq c||S||_\infty) \geq 1 - \alpha \tag{2.6}$$

for a given constant $c > 1$ and a given small probability α. Since the distribution of I is known, the distribution of $||S||_\infty$ is known for any given $\mathcal{X}$ and does not depend on any unknown parameters.

Now for any random variable x let $q_\tau(x)$ denote the $1 - \tau$ quantile of x. Then if we choose $\lambda = cq_\tau(||S||_\infty)$, inequality (2.6) is satisfied. Note that this penalty is provided and discussed in [27]. To approximate this quantity, we propose the following choice of penalty

$$\lambda = c\sqrt{\frac{2A(\alpha)\log p}{n}} \tag{2.7}$$

where $A(\alpha) > 0$ is a constant such that $2p^{-(A(\alpha)-1)} \leq \alpha$.

To show that the above choice of penalty satisfies (2.6), we need to bound the tail probability of $\sum_i^n X_{ij} I_i$ for $j = 1, 2, \ldots, p$. This can be done by using the Hoeffding's inequality [39] and union bounds. We have the following lemma.

Lemma 2.4.1. *The choice of penalty* $\lambda = c\sqrt{\frac{2A(\alpha)\log p}{n}}$ *as in (2.7) satisfies*

$$P(\lambda \geq c||S||_\infty) \geq 1 - \alpha$$

From the proof of the previous lemma, we can see that if we use the following special choice of λ,

$$\lambda = c\sqrt{\frac{2\log p}{n}} \tag{2.8}$$

Then we have that

$$P(\lambda \geq c||S||_\infty) \geq 1 - \frac{2}{p}. \tag{2.9}$$

The above penalties are simple and have good theoretical properties. Moreover, they do not require any conditions on design matrix $\mathcal{X}$ or value of p and n. Note that these choices are based on union bound and concentration inequalities. Thus when the sample size n is relatively small, these inequalities are not very tight. Hence in practice, these penalty levels tend to be relatively large and can cause additional bias to the estimator. From practical point of view, we suggest to use a smaller penalty level when the sample size is not large.

To simplify our arguments, we will use (2.8) as our default choice of penalty in this chapter. It can be seen that the above choices of penalty levels do not depend on the distribution of random errors $\boldsymbol{\epsilon}$ or unknown coefficient $\tilde{\boldsymbol{\beta}}$. As long as ϵ_is are independent random variables with median 0 and $P(\epsilon_i = 0) = 0$, the choices satisfy our requirement. This is a big advantage over the traditional lasso method, which significantly relies on the Gaussian assumption and the variance of the errors.

2.4.2 Properties of the l_1 Penalized Estimator

In this section, we present the analysis of the error bound of l_1 penalized quantile regression estimator. We need to state the upper bound for the estimation error $\mathbf{h} = \widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ under l_2 norm $||\mathbf{h}||_2$. As the choice of penalty is described in the above section,

we assume throughout this chapter, the penalty λ satisfies $\lambda > c||S||_\infty$ for some fixed constant $c > 1$.

At first, we introduce some conditions on design matrix $\mathcal{X}$. Recall that we assume $\lambda > c||S||_\infty$, this implies the following event, namely $\mathbf{h} = \widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ belongs to $\Delta_{\bar{C}}$, where

$$\Delta_{\bar{C}} = \{\gamma \in \mathbf{R}^{p+1} : ||\gamma_{T_+}||_1 \geq \bar{C}||\gamma_{T_+^c}||_1, where\, T_+ = T \bigcup \{0\}\, and\, T \subset \{1, 2, \ldots, p\}\, and\, |T| \leq q\}$$

and $\bar{C} = (c-1)/(c+1)$. To prove this, recall that $\widehat{\boldsymbol{\beta}}$ minimize $\frac{1}{n}||Y - \mathcal{X}\boldsymbol{\beta}||_1 + \lambda||\boldsymbol{\beta}_-||_1$. Hence

$$\frac{1}{n}||\mathcal{X}\mathbf{h} + \boldsymbol{\epsilon}||_1 + \lambda||\widehat{\boldsymbol{\beta}}_-||_1 \leq \frac{1}{n}||\boldsymbol{\epsilon}||_1 + \lambda||\tilde{\boldsymbol{\beta}}||_1$$

Let T denote the set of significant coefficients. Then we have

$$\frac{1}{n}(||\mathcal{X}\mathbf{h} + \boldsymbol{\epsilon}||_1 - ||\boldsymbol{\epsilon}||_1) \leq \lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||).$$

Since the sub-differential of $l_n(\boldsymbol{\beta})$ at the point of $\tilde{\boldsymbol{\beta}}$ is $\mathcal{X}^\intercal I$,

$$\frac{1}{n}(||\mathcal{X}\mathbf{h}+\boldsymbol{\epsilon}||_1-||\boldsymbol{\epsilon}||_1) \geq \frac{1}{n}(\mathcal{X}\mathbf{h})^\intercal I \geq \frac{1}{n}\mathbf{h}^\intercal\mathcal{X}^\intercal I\frac{1}{n} \geq -||\mathbf{h}||_1||\mathcal{X}^\intercal I||_\infty \geq -\frac{\lambda}{c}(||\mathbf{h}_{T_+}||_1-||\mathbf{h}_{T_+^c}||_1)$$

Thus,

$$||\mathbf{h}_{T_+}||_1 \geq \bar{C}||\mathbf{h}_{T_+^c}||_1$$

where $\bar{C} = \frac{c-1}{c+1}$.

Now we define some quantities for the design matrix $\mathcal{X}$. Let λ_q^u be the smallest number such that for any $q+1$ sparse vector $\mathbf{d} \in \mathbb{R}^{p+1}$,

$$||\mathcal{X}\mathbf{d}||_2^2 \leq n\lambda_q^u||\mathbf{d}||_2^2.$$

Here the $q+1$ sparse vector $\mathbf{d}$ means that the vector $\mathbf{d}$ has at most $q+1$ nonzero coordinates, or $||\mathbf{d}||_0 \leq q+1$. Based on matrix theory, we know that $\lambda_q^u = \lambda_{max}(\frac{1}{n}\mathcal{X}_A^\intercal\mathcal{X}_A)$.

Similarly, we have λ_q^l to be the largest number such that for any $q+1$ sparse vector $\mathbf{d} \in \mathbb{R}^{p+1}$,

$$||\mathcal{X}\mathbf{d}||_2^2 \geq n\lambda_q^l||\mathbf{d}||_2^2.$$

and $\lambda_q^l = \lambda_{min}(\frac{1}{n}\mathcal{X}_A^\intercal\mathcal{X}_A)$.

Let θ_{q_1,q_2} be the smallest number such that for any q_1 and q_2 sparse vector $\mathbf{d}_1$ and $\mathbf{d}_2$ with disjoint support,

$$|\langle \mathcal{X}\mathbf{d}_1, \mathcal{X}\mathbf{d}_2\rangle| \leq n\theta_{q_1,q_2}||\mathbf{d}_1||_2||\mathbf{d}_2||_2.$$

The definition of the above constants is essentially the Restricted Isometry Constants [40, 41].

We also need to define the following restricted eigenvalue of design matrix $\mathcal{X}$. These conditions are motivated by [29]. Let

$$\kappa_q^l(\bar{C}) = \min_{\mathbf{h}\in\Delta_{\bar{C}}} \frac{||\mathcal{X}\mathbf{h}||_1}{n||\mathbf{h}_{T_+}||_2}$$

To show the error bound for our l_1 penalized estimator, we need $\kappa_q^l(\bar{C})$ to be bounded away from 0 or goes to 0 slow enough. To simplify the notations, we will write $\kappa_q^l(\bar{C})$ as κ_q^l.

Before presenting the main theorem in [28], we need to state the scale assumptions on ϵ_i. Suppose there exists a constant $a > 0$ such that

$$\begin{aligned} P(\epsilon_i \geq x) &\leq \frac{1}{2+ax} \quad \text{for all } x \geq 0 \\ P(\epsilon_i \leq x) &\leq \frac{1}{2+a|x|} \quad \text{for all } x \leq 0 \end{aligned} \tag{2.10}$$

Here a serves as a scale parameter of the distribution of ϵ_i. This is a very weak condition and even Cauchy distribution satisfies it. Furthermore, we require another two conditions on design matrix $\mathcal{X}$

$$\lambda_q^l > \theta_{q,q}(\frac{1}{\bar{C}} + \frac{1}{4}) \tag{2.11}$$

and

$$\frac{3\sqrt{n}}{16}\kappa_q^l > \lambda\sqrt{qn} + C_1\sqrt{2q\log p}\left(\frac{5}{4} + \frac{1}{\bar{C}}\right), \tag{2.12}$$

for some constant C_1 such that $C_1 > 1 + 2\sqrt{\lambda_q^u}$. We formulate the main theorem in [28] here

Theorem 2.4.2. *Consider the model (2.5), assume $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent and identically distributed random variables satisfying (2.10). Suppose (2.11) and (2.12) hold, the l_1 penalized quantile regression estimator $\widehat{\boldsymbol{\beta}}$ satisfies with probability at least $1 - 2p^{-4q(C_2^2-1)+1}$*

$$||\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||_2 \leq \sqrt{\frac{2q\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a(\lambda_q^l - \theta_{q,q}(\frac{1}{\bar{C}} + \frac{1}{4}))^2/\lambda_q^u}\sqrt{1 + \frac{1}{\bar{C}}}$$

where $C_1 = 1 + 2C_2\sqrt{\lambda_q^u}$ *and* $C_2 > 1$ *is a constant.*

From the theorem above, we can see that with high probability,

$$||\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}||_2 = O_p(\frac{2q \log p}{n}).$$

This means that asymptotically we have our error bound for l_1 penalized estimator to go to zero. Till now, we can combine the results with Theorem 2.3.1 and have the following corollary.

Corollary 2.4.3. *Under assumptions in Theorem 2.4.2 and* λ *also satisfies the conditions in Theorem 2.3.1, the LLA algorithm initiated by* $\widehat{\boldsymbol{\beta}}$ *converge to* $\widehat{\boldsymbol{\beta}}^{oracle}$ *after two iterations with probability at least* $1 - 2p^{-4q(C_2^2-1)+1} - \delta_1 - \delta_2$.

2.5 Numerical Results

In this chapter, we use simulation and real data in [14] to examine the finite sample properties of our non-convex penalized quantile regression model. We consider both SCAD and MCP in the study (denoted by Q-SCAD and Q-MCP, respectively). Generally, we fix $a = 3.7$ in the SCAD and $a = 2$ in the MCP as suggested in [2] and [7] respectively. We compare these two procedures with least-squares based high-dimensional procedures, including LASSO, adaptive LASSO, SCAD and MCP penalized least squares regression (denoted by LS-Lasso, LS-ALasso, LS-SCAD and LS-MCP, respectively). We also compare the proposed procedures with LASSO penalized and adaptive-LASSO penalized quantile regression (denoted by Q-Lasso and Q-ALasso, respectively). Our main interest is the performance of various procedures when $p > n$ and the ability of the nonconvex penalized quantile regression to identify signature variables that are overlooked by the least-squares based procedures.

2.5.1 Simulation Study

Predictors $X_1, X_2, \cdots, X_p$ are generated in two steps. We first generate $(\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)^T$ from the multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. The next step is to set $X_1 = \Phi(\tilde{X}_1)$ and $X_j = \tilde{X}_j$ for $j = 2, 3, \cdots, p$. The

scalar response is generated according to the heteroscedastic location-scale model

$$Y = X_6 + X_{12} + X_{15} + X_{20} + 0.7X_1\epsilon,$$

where $\epsilon \sim N(0,1)$ is independent of the covariates. In this simulation experiment, X_1 plays an essential role in the conditional distribution of Y given the covariates; but does not directly influence the center (mean or median) of the conditional distribution.

We consider sample size $n = 300$ and covariate dimension $p = 400$ and 600. For quantile regression, we consider three different quantiles $\tau = 0.3$, 0.5 and 0.7. We generate an independent tuning data set of size $10n$ to select the regularization parameter by minimizing the estimated prediction error (based on either the squared error loss or the check function loss, depending on which loss function is used for estimation) calculated over the tuning data set; similarly as in [21]. In the real data analysis in section 2.5.2, we use cross-validation for tuning parameter selection.

For a given method, we denote the resulted estimate by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$. Based on simulation of 100 repetitions, we compare the performance of the aforementioned different methods in terms of the following criteria.

Size: the average number of non-zero regression coefficients $\widehat{\beta}_j \neq 0$ for $j = 1, 2, \cdots, p$;

P1: the proportion of simulation runs including all true important predictors, namely $\widehat{\beta}_j \neq 0$ for any $j \geq 1$ satisfying $\beta_j \neq 0$. For the LS-based procedures and conditional median regression, this means the percentage of times including X_5, X_{12}, X_{15} and X_{20}; for conditional quantile regression at $\tau = 0.3$ and $\tau = 0.7$, X_1 should also be included.

P2: the proportion of simulation runs X_1 is selected.

AE: the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

Tables 2.1 and 2.2 depict the simulation results for $p = 400$ and $p = 600$, respectively. In these two tables, the numbers in the parentheses in the columns labeled 'Size' and 'AE' are the corresponding sample standard deviations based on the 100 simulations. The simulation results confirm satisfactory performance of the nonconvex penalized quantile regression when $p > n$ for selecting and estimating relevant covariates. In this example, the signature variable X_1 is often missed by least-squares based methods,

but has high probability of being included when several different quantiles are examined together. This demonstrates that by considering several different quantiles, it is likely to gain a more complete picture of the underlying structure of the conditional distribution. From Tables 2.1 and 2.2, it can be seen that the penalized quantile median regression improves the corresponding penalized least squares methods in terms of AE due to the heteoscedastic error. Furthermore, it is observed that LASSO-penalized quantile regression tends to select a much larger model; on the other hand, the adaptive-Lasso penalized quantile regression tends to select a sparser model but with substantially higher estimation error for $\tau = 0.3$ and 0.7.

Table 2.1: Simulation results for penalized quantile regression models ($p = 400$)

Method	Size	P1	P2	AE
LS-Lasso	25.08 (0.60)	100%	6%	1.37(0.03)
Q-Lasso ($\tau = 0.5$)	24.43 (0.97)	100%	6%	0.95 (0.03)
Q-Lasso ($\tau = 0.3$)	29.83 (0.97)	99%	99%	1.67 (0.05)
Q-Lasso ($\tau = 0.7$)	29.65 (0.90)	98%	98%	1.58 (0.05)
LS-ALASSO	5.02 (0.08)	100%	0%	0.38 (0.02)
Q-Alasso ($\tau = 0.5$)	4.66 (0.09)	100%	1%	0.18 (0.01)
Q-Alasso ($\tau = 0.3$)	6.98 (0.20)	100%	92%	0.63 (0.02)
Q-Alasso ($\tau = 0.7$)	6.43 (0.15)	100%	98%	0.61 (0.02)
LS-SCAD	5.83 (0.20)	100%	0%	0.37 (0.01)
Q-SCAD ($\tau = 0.5$)	5.86 (0.24)	100%	0%	0.19 (0.01)
Q-SCAD ($\tau = 0.3$)	8.29 (0.34)	99%	99%	0.32 (0.02)
Q-SCAD ($\tau = 0.7$)	7.96 (0.30)	97%	97%	0.30 (0.02)
LS-MCP	5.43 (0.17)	100%	0%	0.37 (0.01)
Q-MCP ($\tau = 0.5$)	5.33 (0.18)	100%	1%	0.19 (0.01)
Q-MCP ($\tau = 0.3$)	6.76 (0.25)	99%	99%	0.31 (0.02)
Q-MCP ($\tau = 0.7$)	6.66 (0.20)	97%	97%	0.29 (0.02)

Table 2.2: Simulation results for penalized quantile regression models ($p = 600$)

Method	Size	P1	P2	AE
LS-Lasso	24.30 (0.61)	100%	7%	1.40 (0.03)
Q-Lasso ($\tau = 0.5$)	25.76 (0.94)	100%	10%	1.05 (0.03)
Q-Lasso ($\tau = 0.3$)	34.02 (1.27)	93%	93%	1.82 (0.06)
Q-Lasso ($\tau = 0.7$)	32.74 (1.22)	90%	90%	1.78 (0.05)
LS-ALASSO	4.68 (0.08)	100%	0%	0.37(0.02)
Q-Alasso ($\tau = 0.5$)	4.53 (0.09)	100%	0%	0.18 (0.01)
Q-Alasso ($\tau = 0.3$)	6.58 (0.21)	100%	86%	0.67 (0.02)
Q-Alasso ($\tau = 0.7$)	6.19 (0.16)	100%	86%	0.62 (0.01)
LS-SCAD	6.04 (0.25)	100%	0%	0.38 (0.02)
Q-SCAD ($\tau = 0.5$)	6.14 (0.36)	100%	7%	0.19 (0.01)
Q-SCAD ($\tau = 0.3$)	9.02 (0.45)	94%	94%	0.40 (0.03)
Q-SCAD ($\tau = 0.7$)	9.97 (0.54)	100%	100%	0.38 (0.03)
LS-MCP	5.56 (0.19)	100%	0%	0.38 (0.02)
Q-MCP ($\tau = 0.5$)	5.33 (0.23)	100%	3%	0.18 (0.01)
Q-MCP ($\tau = 0.3$)	6.98 (0.28)	94%	94%	0.38 (0.03)
Q-MCP ($\tau = 0.7$)	7.56 (0.32)	98%	98%	0.37 (0.03)

2.5.2 Real Data Analysis

We now illustrate the proposed methods by an empirical analysis of a real data set. The data set came from a study that used expression quantitative trait locus (eQTL) mapping in laboratory rats to investigate gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease [42].

This microarray data set has expression values of 31042 probe sets on 120 twelve-week-old male offspring of rats. We carried out the following two preprocessing steps: remove each probe for which the maximum expression among the 120 rats is less than the 25th percentile of the entire expression values; and remove any probe for which the range of the expression among 120 rats is less than 2. After these two preprocessing steps, there are 18958 probes left. As in [5, 11], we study how expression of gene TRIM32 (a gene

identified to be associated with human hereditary diseases of the retina), corresponding to probe 1389163_at, depends on expressions at other probes. As pointed out in [42], "Any genetic element that can be shown to alter the expression of a specific gene or gene family known to be involved in a specific disease is itself an excellent candidate for involvement in the disease, either primarily or as a genetic modifier." We rank all other probes according to the absolute value of the correlation of their expression and the expression corresponding to 1389163_at and select the top 300 probes. Then we apply several methods on these 300 probes.

First, we analyze the complete data set of 120 rats. The penalized least squares procedures and the penalized quantile regression procedures studied in Section 3.1 were applied. We use five-fold cross validation to select the tuning parameter for each method. In the second column of Table 3, we report the number of nonzero coefficients (# nonzero) selected by each method.

There are two interesting findings. First, the sizes of the models selected by penalized least squares methods are different from that of models selected by penalized quantile regression. In particular, both LS-SCAD and LS-MCP, which focus on the mean of the conditional distribution, select sparser models compared to Q-SCAD and Q-MCP. A sensible interpretation is that a probe may display strong association with the target probe only at the upper tail or lower tail of the conditional distribution; it is also likely that a probe may display associations in opposite directions at the two tails. The least-squares based methods are likely to miss such heterogeneous signals. Second, a more detailed story is revealed when we compare the probes selected at different quantiles $\tau = 0.3, 0.5, 0.7$. The probes selected by Q-SCAD(0.3), Q-SCAD(0.5), and Q-SCAD(0.7) are reported in the first column of the left, center and right panels, respectively, of Table 2.4. Although Q-SCAD selects 23 probes at both $\tau = 0.5$ and $\tau = 0.3$, only 7 of the 23 overlap, and only 2 probes (1382835_at and 1393382_at) are selected at all three quantiles. We observe similar phenomenon with Q-MCP. This further demonstrates the heterogeneity in the data.

We then conduct 50 random partitions. For each partition, we randomly select 80 rats as the training data and the other 40 as the testing data. A five-fold cross-validation is applied to the training data to select the tuning parameters. We report the average number of nonzero regression coefficients (ave # nonzero), where numbers in

Table 2.3: Analysis of microarray data set

	all data	random partition	
Method	# nonzero	ave # nonzero	prediction error
LS-Lasso	24	21.66(1.67)	1.57(0.03)
Q-Lasso ($\tau = 0.5$)	23	18.36(0.83)	1.51(0.03)
Q-Lasso ($\tau = 0.3$)	23	19.34(1.69)	1.54(0.04)
Q-Lasso ($\tau = 0.7$)	17	15.54(0.71)	1.29(0.02)
LS-ALASSO	16	15.22(10.72)	1.65(0.27)
Q-ALasso ($\tau = 0.5$)	13	11.28(0.65)	1.53(0.03)
Q-ALasso ($\tau = 0.3$)	19	12.52(1.38)	1.57(0.03)
Q-ALasso ($\tau = 0.7$)	10	9.16(0.48)	1.32(0.03)
LS-SCAD	10	11.32(1.16)	1.72(0.04)
Q-SCAD ($\tau = 0.5$)	23	18.32(0.82)	1.51(0.03)
Q-SCAD ($\tau = 0.3$)	23	17.66(1.52)	1.56(0.04)
Q-SCAD ($\tau = 0.7$)	19	15.72(0.72)	1.30(0.03)
LS-MCP	5	9.08(1.68)	1.82(0.04)
Q-MCP ($\tau = 0.5$)	23	17.64(0.82)	1.52(0.03)
Q-MCP ($\tau = 0.3$)	15	16.36(1.53)	1.57(0.04)
Q-MCP ($\tau = 0.7$)	16	13.92(0.72)	1.31(0.03)

the parentheses are the corresponding standard errors across 50 partitions, in the third column of Table 3. We evaluate the performance over the test set for each partition. For Q-SCAD and Q-MCP, we evaluate the loss using the check function at the corresponding τ. As the squared loss is not directly comparable with the check loss function, we use the check loss with $\tau = 0.5$ (i.e. l_1 loss) for the LS-based methods. The results are reported in the last column of Table 3, where the prediction error is defined as $\sum_{i=1}^{40} \rho_\tau(Y_i - \widehat{Y}_i)$ and the numbers in the parentheses are the corresponding standard errors across 50 partitions. We observe similar patterns as when the methods are applied to the whole data set. Furthermore, the penalized quantile regression procedures improves the corresponding penalized least squares in terms of prediction error. The performance

of Q-Lasso, Q-ALasso, Q-SCAD and Q-MCP are similar in terms of prediction error, although the Q-Lasso tends to select less sparse models and the Q-ALasso tends to select sparser model, compared with Q-SCAD and Q-MCP.

As with every variable selection method, different repetitions may select different subsets of important predictors. In Table 2.4, we report in the left column the probes selected using the complete data set and in the right column the frequency these probes appear in the final model of these 50 random partitions for Q-SCAD(0.3), Q-SCAD(0.5), and Q-SCAD(0.7) in the left, middle and right panels, respectively. The probes are ordered such that the frequency is decreasing. From Table 2.4, we observe that some probes such as 1383996_at and 1382835_at have high frequencies across different τ's, while some other probes such as 1383901_at do not. This implies that some probes are important across all τ, while some probes might be important only for certain τ.

Wei and He (2006) proposed a simulation based graphical method to evaluate the overall lack-of-fit of the quantile regression process. We apply their graphical diagnosis method using the SCAD penalized quantile regression. More explicitly, we first generate a random $\tilde{\tau}$ from the uniform (0,1) distribution. We then fit the SCAD-penalized quantile regression model at the quantile $\tilde{\tau}$, where the regularization parameter is selected by five-fold cross-validation. Denote the penalized estimator by $\widehat{\boldsymbol{\beta}}(\tilde{\tau})$, and we generate a response $Y = \mathcal{X}^{\intercal}\widehat{\boldsymbol{\beta}}(\tilde{\tau})$, where $\mathcal{X}$ is randomly sampled from the set of observed vector of covariates. We repeat this process 200 times and produce a sample of 200 simulated responses from the postulated linear model. The QQ plot of the simulated sample vs the observed sample is given in Figure 2.1. The QQ plot is close to 45 degree line and thus indicates a reasonable fit.

Table 2.4: Frequency table for the real data

Q-SCAD(0.3)		Q-SCAD(0.5)		Q-SCAD(0.7)	
Probe	Frequency	Probe	Frequency	Probe	Frequency
1383996_at	31	1383996_at	43	1379597_at	38
1389584_at	26	1382835_at	40	1383901_at	34
1393382_at	24	1390401_at	27	1382835_at	34
1397865_at	24	1383673_at	24	1383996_at	34
1370429_at	23	1393382_at	24	1393543_at	30
1382835_at	23	1395342_at	23	1393684_at	27
1380033_at	22	1389584_at	21	1379971_at	23
1383749_at	20	1393543_at	20	1382263_at	22
1378935_at	18	1390569_at	20	1393033_at	19
1383604_at	15	1374106_at	18	1385043_at	18
1379920_at	13	1383901_at	18	1393382_at	17
1383673_at	12	1393684_at	16	1371194_at	16
1383522_at	11	1390788_a_at	16	1383110_at	12
1384466_at	10	1394399_at	14	1395415_at	6
1374126_at	10	1383749_at	14	1383502_at	6
1382585_at	10	1395415_at	13	1383254_at	5
1394596_at	10	1385043_at	12	1387713_a_at	5
1383849_at	10	1374131_at	10	1374953_at	3
1380884_at	7	1394596_at	10	1382517_at	1
1369353_at	5	1385944_at	9		
1377944_at	5	1378935_at	9		
1370655_a_at	4	1371242_at	8		
1379567_at	1	1379004_at	8		

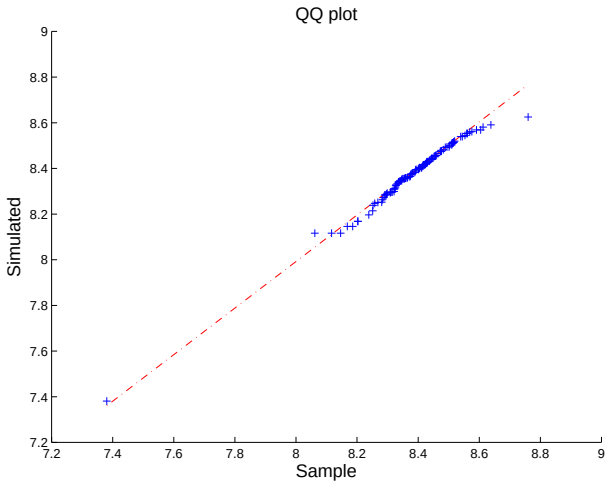

Figure 2.1: *Lack-of-fit diagnosis QQ plot for the real data.*

Chapter 3

A Iterative Coordinate Descent Algorithm for High-dimensional Non-convex Penalized Quantile Regression

3.1 Chapter Overview

In last chapter, we have solved the final puzzle on non-convex penalized quantile regression on searching for oracle estimator via the LLA algorithm. However, the computation for non-convex optimization is highly challenging when $p \gg n$. Specifically, the LLA algorithm can be exceedingly slow on solving non-convex penalized quantile regression model, which pluses a non-smooth loss function. In this chapter, we propose and study a new iterative coordinate descent algorithm (QICD) for solving this problem in ultra high dimension. Existing coordinate descent algorithms for least squares regression cannot be directly applied, hence we imbed the majorization-minimization (MM) idea in our method to tackle the non-convexity. We establish the convergence property of the proposed algorithm under some regular conditions for a general class of non-convex penalty functions including popular choices such as SCAD and MCP. Our simulation study confirms that QICD substantially improves the computational speed in the $p \gg n$

setting. We further illustrate the application by analyzing a microarray data set. More details can be found in [43].

3.2 Introduction

In the setting $p \gg n$, the theory for penalized quantile regression has been systematically studies in the last chapter, we switch to work on the breakthrough on increasing the computational speed for this problem. For unpenalized quantile regression, [44] proposed a useful interior point algorithm; and [45] developed an effective MM algorithm which majorizes the quantile loss function by a quadratic function. Several algorithms have also been developed for penalized quantile regression. For Lasso penalized quantile regression, [34] proposed an algorithm that computes the entire solution path; [46] includes a fast greedy coordinate descent algorithm for median regression. However, neither algorithm applies to non-convex penalties. A linear programming based modified LLA algorithm [17] was used in [14] for nonconvex penalized quantile regression, but its computation slows noticeably when p is large. Moreover, the aforementioned work have not studied the convergence theory of the proposed algorithm.

To tackle the computational challenges caused by the nonsmooth quantile loss function and the nonconvex penalty function, we propose a new iterative coordinate descent algorithm and study its convergence property. The new algorithm achieves fast computation by successively solving a sequence of univariate minimization subproblems. It combines the idea of the MM algorithm [18, 19, 47] with that of the coordinate descent algorithm. We refer to this new iterative coordinate descent algorithm as QICD, where Q stands for quantile. We consider a general class of nonconvex penalty functions and establish the convergence property of the QICD algorithm by extending [48]'s theory for the convergence of the coordinate descent algorithm. It is noteworthy that [48] requires a quasiconvexity condition, which is not met by nonconvex penalized quantile regression.

The coordinate descent algorithm was systematically investigated for convex problems, such as Lasso, in the independent work of [46, 49], the idea of which can be traced back to [50, 51]. For nonconvex penalized least squares regression, coordinate descent

algorithms and their convergence theory were recently investigated by [20, 21]. [22] proposed a new coordinate descent algorithm for non-convex penalized generalized linear models which enjoys the appealing property of avoiding the computation of a scaling factor in each update of the solutions. These algorithms are very effective in large-scale problems but do not apply to non-convex penalized quantile regression.

In this chapter, we describe the new QICD algorithm and establish the convergence property. Furthermore, we investigate the performance of the proposed algorithm through Monte Carlo studies and demonstrate its application using a real data example. The technical details are presented in the Appendix.

3.3 The QICD Algorithm

The QICD algorithm combines the idea of the MM algorithm with that of the coordinate descent algorithm. More specifically, we first replace the non-convex penalty function by its majorization function to create a surrogate objective function. Then we minimize the surrogate objective function with respect to a single parameter at each time and cycle through all parameters until convergence. For each univariate minimization problem, we only need to compute a one-dimensional weighted median, which ensures fast computation.

3.3.1 The Majorization Minimization step

We consider a *majorization function* $\phi_{\beta_0}(\beta)$, which majorizes $p_\lambda(|\beta|)$ at β_0 in the sense that

$$\phi_{\beta_0}(\beta) \quad \geq \quad p_\lambda(|\beta|) \quad \text{for all } \beta \text{ with equality when } \beta = \beta_0. \tag{3.1}$$

Let $\boldsymbol{\beta}^{(k)}$ denote the value of $\boldsymbol{\beta}$ after the kth iteration, $k = 1, 2, \ldots$. Let $p'_\lambda(|\beta|+)$ denotes the limit of $p'_\lambda(x)$ as $x \to |\beta|$ from the above. Furthermore, we assume that $p_\lambda(\cdot)$ is piecewise differentiable so that $p'_\lambda(|\beta|+)$ exists for all β. Then in the kth iteration,

$$\phi_{\beta_j^{(k-1)}}(|\beta_j|) = p'_\lambda(|\beta_j^{(k-1)}|+)|\beta_j| - p'_\lambda(|\beta_j^{(k-1)}|+)|\beta_j^{(k-1)}| + p_\lambda(|\beta_j^{(k-1)}|) \tag{3.2}$$

majorizes the penalty function $p_\lambda(|\beta_j|)$, $k = 1, 2, \ldots; j = 1, 2, \ldots, p$; that is,

$$\phi_{\beta_j^{(k-1)}}(|\beta_j|) \geq p_\lambda(|\beta_j|) \text{ for all } \beta_j \text{ with equality when } \beta_j = \beta_j^{(k-1)}, \tag{3.3}$$

see Proposition 3.4.1 below. Figure 2 illustrates the SCAD penalty function and its majorization function; the figure for the MCP penalty looks similar and is thus omitted.

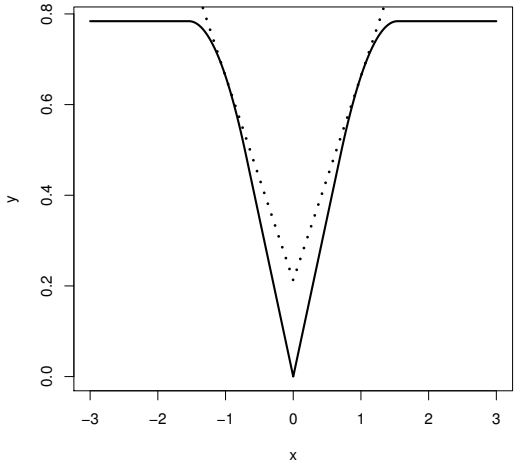

Figure 3.1: *SCAD penalty function (solid line) and its majorization function (dotted line)* $\phi_{\beta_0}(\beta)$ *with* $\lambda = 0.7$,$a = 2.2$.

Subsequently, the penalized objective function $Q(\boldsymbol{\beta})$ defined in (2.2) is majorized by

$$Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) \quad = \quad n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \sum_{j=1}^{p}\phi_{\beta_j^{(k-1)}}(|\beta_j|) \tag{3.4}$$

at the kth iteration. It can be shown that any decrease of the value of $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ results in a decrease of the value of $Q(\boldsymbol{\beta})$. Hence, we minimize the majorization function $Q_{\boldsymbol{\beta}^{k-1}}(\boldsymbol{\beta})$ at iteration k to update the value of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(k)} \quad = \quad \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) \tag{3.5}$$

The above iterative scheme decreases the value of $Q(\boldsymbol{\beta})$ monotonically in each iteration. This property is summarized in Proposition 3.4.1. We note that when considering nonconvex penalized generalized linear models, [22] applied a majorization step on the

loss function to avoid the computation of a scaling factor in each update of the solution. Different from their approach, our majorization step is applied on the nonconvex penalty function. The majorization step solves two problems at once. First, it transforms the problem of minimizing a nonconvex objective function to a sequence of convex minimization problems, for which the coordinate descent algorithm can be applied. Second, the majorized penalized quantile loss function is quasiconvex, which allows us to apply the results in [48] to further study the convergence property of the proposed algorithm.

3.3.2 The Coordinate Descent Step

To solve the minimization problem in (3.5), we employ the idea of the "one-at-a-time" coordinate descent algorithm; that is, to update the jth coordinate, we treat the other coordinates as fixed. We would incorporate the sub-iteration of the coordinate descent minimization within each iteration of the majorization minimization step.

Assume that at the beginning of the kth iteration, the value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(k-1)}$. To minimize $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$, we apply the coordinate descent algorithm, which at each sub-iteration cycles through all the $(p+1)$ covariates. Consider the rth sub-iteration of the kth iteration. Suppose we have finished updating the estimates of the coefficients for $x_{i0}, x_{i1}, \ldots, x_{i(j-1)}$ and obtain $\boldsymbol{\beta}_{j-1}^{(k)(r)} = (\beta_0^{(k)(r+1)}, \ldots, \beta_{j-1}^{(k)(r+1)}, \beta_j^{(k)(r)}, \ldots, \beta_p^{(k)(r)})^\intercal$. Next, we update the estimate for the coefficient of x_{ij} by

$$
\begin{aligned}
\beta_j^{(k)(r+1)} &= \underset{\beta_j}{\operatorname{argmin}}\, Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}_{j-1}^{(k)(r)}) \\
&= \underset{\beta_j}{\operatorname{argmin}} \Big\{ n^{-1} \Big[\sum_{i=1}^{n} \rho_\tau \big(Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - x_{ij}\beta_j \\
&\quad - \sum_{s>j} x_{is}\beta_s^{(k)(r)} \big) \Big] + \Big[\sum_{s<j} \phi_{\beta_s^{(k-1)}}(|\beta_s^{(k)(r+1)}|) + \phi_{\beta_j^{(k-1)}}(|\beta_j|) \\
&\quad + \sum_{s>j} \phi_{\beta_s^{(k-1)}}\big(|\beta_s^{(k)(r)}|\big) \Big] \Big\} \\
&= \underset{\beta_j}{\operatorname{argmin}} \Big\{ n^{-1} \Big[\sum_{i=1}^{n} \rho_\tau \big(Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - x_{ij}\beta_j \\
&\quad - \sum_{s>j} x_{is}\beta_s^{(k)(r)} \big) \Big] + p'_\lambda(|\beta_j^{(k-1)}|+)|\beta_j| \Big\}. \qquad (3.6)
\end{aligned}
$$

It is noteworthy that in the above minimization $\boldsymbol{\beta}^{(k-1)}$ and all the other coordinates

in $\boldsymbol{\beta}_{j-1}^{(k)(r)}$ are held fixed. An important observation is that (3.6) can be equivalently expressed as a minimization problem for weighted median regression. To see the connection, we rewrite (3.6) as

$$\min_{\beta_j} \Big\{ (n+1)^{-1} \sum_{i=1}^{n+1} \omega_{ij} |u_{ij}| \Big\}, \tag{3.7}$$

where

$$u_{ij} = \begin{cases} \frac{Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - \sum_{s>j} x_{is}\beta_s^{(k)(r)}}{x_{ij}} - \beta_j, & i = 1, 2, \cdots, n, \\ \beta_j, & i = n+1, \end{cases}$$

and

$$\omega_{ij} = \begin{cases} n^{-1}|x_{ij}(\tau - I(u_{ij}x_{ij} < 0))|, & i = 1, 2, \cdots, n, \\ p'_\lambda(|\beta_j^{(k-1)}|+), & i = n+1. \end{cases}$$

Therefore, $\beta_j^{(k)(r+1)}$ can be obtained by solving a single parameter quantile regression model using the above $n+1$ pseudo-observations, $j > 1$ with $\tau = 0.5$. In practice, after the rth sub-iteration of the kth iteration of this algorithm, we have the weights $\omega_{ij}^{(k)(r)}$

$$\omega_{ij}^{(k)(r)} = \begin{cases} n^{-1}|x_{ij}(\tau - I(u_{ij}^{(k)(r)}x_{ij} < 0))|, & i = 1, 2, \cdots, n, \\ p'_\lambda(|\beta_j^{(k-1)}|+), & i = n+1. \end{cases}$$

where

$$u_{ij}^{(k)(r)} = \begin{cases} \frac{Y_i - \sum_{s<j} x_{is}\beta_s^{(k)(r+1)} - \sum_{s>j} x_{is}\beta_s^{(k)(r)}}{x_{ij}} - \beta_j^{(k)(r)}, & i = 1, 2, \cdots, n, \\ \beta_j, & i = n+1, \end{cases}$$

Hence we can calculate $\beta_j^{(k)(r+1)}$ by using weighted median searching in (3.7).

When $j = 0$, $\beta_0^{(k)(r+1)}$ can be calculated by using only n pseudo-observations since no penalty is given to $\beta_0^{(k)(r+1)}$. A similar observation was made for the Lasso penalized median regression by [46]. The weighted median can be computed quickly by many statistical software packages such as the *quantreg* package in R. Actually, quicksort, also known as partition-exchange sort, is utilized in this algorithm to find the weighted median, ensuring the high speed in each update of $\beta_j^{(k)(r+1)}$.

The above computation yields

$$\boldsymbol{\beta}_j^{(k)(r)} = (\beta_0^{(k)(r+1)}, \ldots, \beta_j^{(k)(r+1)}, \beta_{j+1}^{(k)(r)}, \ldots, \beta_p^{(k)(r)})^T.$$

This process is repeated for $r = 1, 2, \ldots,$ until convergence. Then we update $\boldsymbol{\beta}^{(k-1)}$ to $\boldsymbol{\beta}^{(k)}$.

3.3.3 Choice of the Tuning Parameter

Algorithm 3.3.1 summarizes the details of the QICD algorithm. for a given tuning parameter λ. In real applications, the choice of λ is important. Cross-validation is popular but is observed to often result in overfitting [52]. Moreover, cross-validation is time-consuming when p is notably large.

High-dimensional BIC-type criterion for nonconvex penalized least-squares regression with diverging p has been recently investigated by [53, 54, 55, 56], among others. [57] recently proposed high-dimensional BIC for quantile regression when p is much larger than n. Motivated by their work, we consider the following high-dimensional BIC criterion. Let $\boldsymbol{\beta}_\lambda = (\beta_{\lambda,1}, ..., \beta_{\lambda,p})^T$ be the penalized estimator obtained with the tuning parameter λ, and let $\mathcal{S}_\lambda \equiv \{j : \beta_{\lambda,j} \neq 0, 1 \leq j \leq p\}$ be the index set of covariates with nonzero coefficients. Define

$$\text{HBIC}(\lambda) = \log\big(\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\lambda)\big) + |\mathcal{S}_\lambda| \frac{\log(\log n)}{n} C_n, \tag{3.8}$$

where $|\mathcal{S}_\lambda|$ is the cardinality of the set $\mathcal{S}_\lambda$, and C_n is a sequence of positive constants diverging to infinity as n increases. We select the value of λ that minimizes $\text{HBIC}(\lambda)$. In practice, we recommend to take $C_n = O(\log p)$, which we find to work well in a variety

of settings.

Algorithm 3.3.1: QICD ALGORITHM$(k, r, j, p, \boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})$

---

comment: Input an initial value $\boldsymbol{\beta}^{(0)}$

for $k \geq 0$

then {

repeat

for $r \geq 0$

then {

for $j \in \{0, 1, 2, \cdots, p\}$

repeat

{

comment: Calculate the weighted median in (3.7).

$\boldsymbol{\beta}_{j+1}^{(k)(r)} \leftarrow \boldsymbol{\beta}_{j}^{(k)(r)}$

for $j = p$

then $r \leftarrow r + 1$

$j \leftarrow j + 1 \pmod{p}$

}

until $\boldsymbol{\beta}_{j}^{(k)(r)}$ converge to $\boldsymbol{\beta}^{*}$

$\boldsymbol{\beta}^{(k)} \leftarrow \boldsymbol{\beta}^{*}$

}

until $\boldsymbol{\beta}^{(k)}$ converge to $\widehat{\boldsymbol{\beta}}$

}

then return $(\widehat{\boldsymbol{\beta}})$

3.4 The Convergence Theory

The main result in this section establishes that under some regularity conditions, the proposed QICD algorithm converges to a stationary point of the penalized objective function in (2.2).

Proposition 3.4.1 below summarizes the properties of the majorization minimization step.

Proposition 3.4.1. *Assume that in the penalized quantile loss function $Q(\boldsymbol{\beta})$ defined in (2.2), $p_\lambda(\cdot)$ is piecewise differentiable, nondecreasing and concave on $(0, \infty)$, and $p_\lambda(\cdot)$ is continuous at 0 with $p'_\lambda(0+) < \infty$. Then*
(1) the function $\phi_{\beta_j^{(k-1)}}(\beta)$ defined in (3.2) majorizes $p_\lambda(|\beta|)$ at the points $\pm|\beta_j^{(k-1)}|$;

(2) the function $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ *defined in (3.4) majorizes* $Q(\boldsymbol{\beta})$ *at the points* $\pm|\boldsymbol{\beta}^{(k-1)}|$*; (3) the majorization minimization step has the descent property, that is, for all* $k = 1, 2, , \ldots$

$$Q(\boldsymbol{\beta}^{(k)}) \leq Q(\boldsymbol{\beta}^{(k-1)}). \tag{3.9}$$

The general theory of [48] on the coordinate descent algorithm does not apply to the penalized quantile objective function in (2.2). This is because non-convex penalized quantile regression does not meet the *quasiconvex* condition. Lemma 3.4.2 below suggests that if we consider the majorized loss function in $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$, then the coordinate descent step yields a coordinate-wise minimum (see Appendix A for the definition), by applying Tseng's theory.

Lemma 3.4.2. *If* $p_\lambda(\cdot)$*satisfies the conditions in Proposition 3.4.1, then the* $\boldsymbol{\beta}^{(k)}$ *defined in Section 3.3.2 is a coordinate-wise minimum point of* $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$*,* $k = 1, 2, \ldots$.

Lemma 3.4.3 below describes the convergence behavior of $Q(\boldsymbol{\beta}^{(k)})$. It indicates that $Q(\boldsymbol{\beta}^{(k)})$ follows similar convergence behavior as its majorization function $Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})$ under some weak conditions.

Lemma 3.4.3. *If* $Q(\boldsymbol{\beta}^{(0)}) < +\infty$*, then* $\{Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})\}$ *is a bounded and decreasing sequence with respect to* k*. If we denote* $\lim_{k\to\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)})$ *by* A*, then* $Q(\boldsymbol{\beta}^*) = A$*, where* $\boldsymbol{\beta}^*$ *be an arbitrary cluster point of* $\{\boldsymbol{\beta}^{(k)}\}$.

The convergence property of the QICD algorithm is established by combining the results of the two preceding lemmas and utilizing a result in [58] (see Lemma B.0.1 in Appendix B). Theorem 3.4.4 below states that every cluster point of the QICD algorithm is a stationary point of the penalized quantile loss function $Q(\boldsymbol{\beta})$.

Theorem 3.4.4. *(Convergence property of QICD) Consider the penalized quantile loss function* $Q(\boldsymbol{\beta})$ *in (2.2), where the given data* $(Y, \mathbf{X})$ *lie on a compact set and* $Q(\boldsymbol{\beta}^{(0)}) < +\infty$ *for an initial value* $\boldsymbol{\beta}^{(0)}$*. Suppose that the penalty function* $p_\lambda(\cdot)$ *satisfies the conditions in Proposition 3.4.1 and* $p'_\lambda(|\theta|+) = p'_\lambda(|\theta|-)$ *on* $(0, \infty)$*. Consider an arbitrary cluster point* $\boldsymbol{\beta}^{**}$ *of* $\{\boldsymbol{\beta}^{(k-1)}\}$*, that is, there exists a sequence* $\{k_m\}$ *such that* $\lim_{m\to\infty} \boldsymbol{\beta}^{(k_m-1)} = \boldsymbol{\beta}^{**}$*. Let* $\boldsymbol{\beta}^*$ *be an arbitrary cluster point of* $\{\boldsymbol{\beta}^{(k_m)}\}$*. Assume* $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$

is regular at $\boldsymbol{\beta}^*$ *and* $\boldsymbol{\beta}^{**}$*. Then* $\boldsymbol{\beta}^{**}$ *is a stationary point of* $Q(\boldsymbol{\beta})$*. In particular, every cluster point of the sequence generated by the QICD algorithm* $\{\boldsymbol{\beta}^{(k)}\}$ *is a stationary point of* $Q(\boldsymbol{\beta})$*.*

Note that the condition on $(Y, \mathbf{X})$ is a mild assumption. And the conditions on the penalty functions are satisfied by the popular nonconvex SCAD and MCP penalties. The proof of Theorem 3.4.4 is provided in Appendix B.

3.5 Numerical Examples

3.5.1 Monte Carlo Simulations

To compare the QICD algorithm with existing methods, we use the same simulation data settings as in Section 2.5.1 but increase the dimension p to even higher level. In this study, we consider sample size $n = 300$, covariates dimension $p = 1000$ and 2000, and three different quantiles $\tau = 0.3, 0.5$ and 0.7. For each simulation scenario, we have 100 simulation runs. Specifically, as the comparison of Lasso penalized quantile regression and non-convex penalized quantile regression has been performed in [14], we focus on comparing the performance of the QICD with that of LLA, with SCAD and MCP penalty functions. For both procedures, the high-dimensional BIC defined in Section 3.3.3 is applied to choose the tuning parameter.

The convergence criteria used in implementing the QICD algorithm are as follows: (i) the coordinate descent minimization step in each iteration stops if the absolute difference of the successive sub-iterations is less than 10^{-6} (convergence of coefficients in sub-iteration) and the number of sub-iterations exceeds p; (ii) the majoriaztion minimization step stops if the absolute difference of the successive iterations is less than 10^{-6} (convergence of coefficients in iteration)and the number of iterations exceeds 100.

We evaluate the two algorithms by the estimation error and the model selection ability of the resulted estimators, and their respective computational speed. For a given estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^\intercal$, we consider the following five similar criteria as in 2.5.1:

Size: the average number of non-zero regression coefficients $\widehat{\beta}_j \neq 0$ for $j = 1, 2, \cdots, p$;

P1: the proportion of simulation runs including all true important predictors, namely

$\widehat{\beta}_j \neq 0$ satisfying $\beta_j \neq 0$, $j \geq 1$. Note that for $\tau = 0.5$ the true model includes X_6, X_{12}, X_{15} and X_{20}; for $\tau = 0.3$ and $\tau = 0.7$, X_1 should also be included.

P2: the proportion of simulation runs X_1 is selected.

AE: the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

Time: the running times (CPU seconds) for each method in each repetition (the process of calculating the estimate $\hat{\boldsymbol{\beta}}$).

Table 3.1: QICD Simulation results ($p = 1000$)

Method	Size	P1	P2	AE	Time
QICD-SCAD ($\tau = 0.5$)	5.15 (0.51)	100%	0%	0.04 (0.01)	1.53 (0.40)
QICD-SCAD ($\tau = 0.3$)	7.53 (2.11)	100%	94%	0.11 (0.02)	1.57 (0.39)
QICD-SCAD ($\tau = 0.7$)	8.02 (2.42)	100%	93%	0.11 (0.03)	1.56 (0.33)
LLA-SCAD ($\tau = 0.5$)	5.00 (0.00)	100%	0%	0.04 (0.01)	24.78 (1.54)
LLA-SCAD ($\tau = 0.3$)	7.68 (1.44)	100%	91%	0.11 (0.03)	39.43 (3.27)
LLA-SCAD ($\tau = 0.7$)	10.86 (2.06)	100%	94%	0.13 (0.02)	28.59 (1.94)
QICD-MCP ($\tau = 0.5$)	5.25 (0.72)	100%	0%	0.04 (0.01)	1.52 (0.41)
QICD-MCP ($\tau = 0.3$)	7.57 (1.95)	100%	96%	0.12 (0.03)	1.94 (0.60)
QICD-MCP ($\tau = 0.7$)	8.40 (2.78)	100%	96%	0.12 (0.03)	1.77 (0.36)
LLA-MCP ($\tau = 0.5$)	5.00 (0.00)	100%	0%	0.04 (0.01)	29.88 (6.92)
LLA-MCP ($\tau = 0.3$)	8.58 (1.68)	100%	93%	0.12 (0.03)	38.54 (8.50)
LLA-MCP ($\tau = 0.7$)	9.89 (1.81)	100%	95 %	0.12 (0.02)	69.69 (19.70)

Tables 3.1 and 3.2 summarize the simulation results for p=1000 and 2000, respectively. We observe that both the QICD algorithm and the LLA algorithm have satisfactory performance in terms of estimation error and model selection accuracy. The QICD algorithm is remarkably faster than the LLA algorithm. For $p = 1000$, the QICD algorithm takes less than 2 seconds to finish a repetition, which is $\frac{1}{15}$ of the runtime of the LLA algorithm; for $p = 2000$ case, the LLA always needs 200 seconds or more to finish one repetition, which is much longer than the runtime of the QICD algorithm

Table 3.2: QICD Simulation results ($p = 2000$)

Method	Size	P1	P2	AE	Time
QICD-SCAD ($\tau = 0.5$)	5.23 (0.88)	100%	0%	0.04 (0.01)	3.37 (1.08)
QICD-SCAD ($\tau = 0.3$)	8.00 (2.49)	100%	93%	0.11 (0.03)	2.96 (0.76)
QICD-SCAD ($\tau = 0.7$)	8.52 (2.16)	100%	93%	0.12 (0.03)	3.19 (0.72)
LLA-SCAD ($\tau = 0.5$)	13.48 (2.93)	100%	0%	0.07 (0.02)	235.17 (17.57)
LLA-SCAD ($\tau = 0.3$)	8.41 (1.68)	100%	87%	0.11 (0.03)	148.94 (8.00)
LLA-SCAD ($\tau = 0.7$)	9.67 (2.21)	100%	92%	0.12 (0.03)	214.23 (20.21)
QICD-SCAD ($\tau = 0.5$)	5.33 (1.18)	100%	0%	0.04 (0.02)	3.05 (0.80)
QICD-SCAD ($\tau = 0.3$)	8.21 (2.72)	100%	92%	0.12 (0.03)	3.20 (0.76)
QICD-SCAD ($\tau = 0.7$)	8.48 (2.17)	100%	93%	0.12 (0.03)	3.72 (1.05)
LLA-MCP ($\tau = 0.5$)	13.67 (2.97)	100%	0%	0.07 (0.02)	166.12 (38.98)
LLA-MCP ($\tau = 0.3$)	8.45 (2.03)	100%	88%	0.12 (0.03)	219.03 (46.37)
LLA-MCP ($\tau = 0.7$)	8.65 (1.77)	100%	92 %	0.12 (0.02)	354.01 (83.76)

(under 4 seconds). Furthermore, the LLA algorithm owns vast variation on running time; especially in case $p = 2000$, the standard deviation of then running time when $\tau = 0.7$ could be as large as 80. However, the running time of QICD algorithm is much stabler with standard deviation less than 1. In other words, the QICD algorithm can keep high level running speed consistently in various situations.

The QICD algorithm tends to select a sparser model but with comparable estimation error comparing with the model selected by the LLA algorithm. In particular, for $p = 2000$, the estimation error associated the QICD algorithm is slightly smaller than that of the LLA algorithm; meanwhile, the size of none zero coefficients for the QICD algorithm when $\tau = 0.5$, around 5, is also moderately smaller than that of the LLA algorithm, around 13.

Furthermore, the QICD algorithm has a smaller error of selecting X_1 at the median comparing with the LLA algorithm. In summary, the fast computation of the QICD algorithm does not come at the cost of sacrificing its performance.

3.5.2 An Application

We next analyze the same microarray dataset of [42] in Section 2.5.2 for studying expression quantitative trait locus (eQTL) mapping in the laboratory rats. After the same preprocessing, 18,958 probes remained. We are interested in how the expression of gene TRIM32 (a gene identified to be associated with human hereditary diseases of the retina), corresponding to probe 1389163_at, depends on expressions at other probes. As pointed out by [42], "Any genetic element that can be shown to alter the expression of a specific gene or gene family know to be involved in a specific disease is itself an excellent candidate for involvement in the disease, either primarily or as a genetic modifier."

We rank all remaining 18,958 probes according to the absolute value of the correlation of their expression and the expression corresponding to 1389163_at and select the top 3000 probes. On this subset ($n = 120$, $p = 3000$), we applied the QICD algorithm to study the relationships between the expression of TRIM32 and expression of the 3000 genes. First, we analyze the data on all 120 rats using SCAD or MCP penalized quantile regression and consider three quantiles $\tau = 0.3$, 0.5 and 0.7. Still, we use the HBIC to select the tuning parameter λ for each case. In the second column of Table 3.3, we report the number of nonzero coefficients (# nonzero) selected in each case. An interesting finding is that different sets of probes are selected at different quantiles. Specifically, in the SCAD cases, though 18 and 21 probes have been selected at $\tau = 0.3$ and $\tau = 0.7$ respectively, only 6 of them overlap ("1368887_at" "1382291_at" "1390048_at" "1380371_at" "1395973_at" "1374786_at"),but none of them is selected at $\tau = 0.5$. We could find the similar phenomenon on MCP case. This reveals the heterogeneity of this dataset.

Then we randomly partition the 120 rats 50 times. In each partition, we randomly select 80 rats for the training set and have the rest 40 as a test set. We fit penalized quantile regression model and compute the tuning parameter on the training set. The we the prediction error of the selected model using the test set. In the third column of Table 3.3, we report the average number of nonzero regression coefficients of the selected model and their associated robust standard deviations over the 50 repetitions. In the last column of Table 3.3, we report the prediction error (and its standard deviation) on the test data. the predication error is computed using the check function at the

Table 3.3: Analysis of microarray data set

	all data	random partition	
Method	# nonzero	ave # nonzero	prediction error
QICD-SCAD ($\tau = 0.5$)	18	16.53 (6.59)	1.72 (0.22)
QICD-SCAD ($\tau = 0.3$)	21	18.08 (7.09)	1.57 (0.28)
QICD-SCAD ($\tau = 0.7$)	13	12.06 (6.63)	1.53 (0.19)
QICD-MCP ($\tau = 0.5$)	19	15.61 (6.32)	1.73 (0.22)
QICD-MCP ($\tau = 0.3$)	23	16.86 (6.60)	1.56 (0.26)
QICD-MCP ($\tau = 0.7$)	12	11.12 (4.87)	1.52 (0.18)

corresponding τ, that is, $\sum_{i=1}^{40} \rho_\tau(y_i - \hat{y}_i)$. We observe that the performance of SCAD penalty and MCP penalty is similar. At quantiles $\tau = 0.5$ and 0.7, we tend to select fewer probes than at $\tau = 0.3$; however, at $\tau = 0.3$ and 0.7, we have a smaller prediction error than at $\tau = 0.5$.

3.6 Discussion

The chapter describes two timely contributions to nonconvex penalized quantile regression analysis of high-dimensional data. It proposes a fast iterative coordinate descent algorithm which is shown empirically to significantly improve the computational speed. Furthermore, it extends [48]'s theory to establish the convergence properties.

We emphasize here the focus of this paper on extending the coordinate descent algorithm for fast computation with high-dimensional data. Although the majorization step is adopted, this step alone does not lead to the desirable computational efficiency gain for nonconvex penalized quantile regression. We note that the LLA algorithm[17] for penalized least squares regression also falls into the framework of MM algorithm. A stable and versatile class of MM algorithms applicable to a wide variety of penalization problems with non-convex penalty was given in [59]. They established a local convergence theory but required a strict convexity condition, thus excluded the more interesting $p > n$ case which is under study in the current paper.

Chapter 4

Non-convex Penalized Support Vector Machines

4.1 Chapter Overview

The support vector machine is a powerful binary classification tool with high accuracy and great flexibility. It has achieved great success, but its performance can be seriously impaired if many redundant covariates are included. Some efforts have been devoted to studying variable selection for SVMs, but asymptotic properties, such as variable selection consistency, are largely unknown under high dimensional settings. Fortunately, the hinge loss function in SVM owns quite similar properties as quantile regression model, which stimulate us to try to extend our theory in Chapter 2 for non-convex penalized SVM. In this chapter, we first demonstrate the results of [60], in ultrahigh dimensions, there is one local minimizer to the objective function of non-convex penalized SVMs having the desired oracle property. We further address the problem of non-unique local minimizers by showing that the local linear approximation algorithm is guaranteed to converge to the oracle estimator even in the ultrahigh dimensional setting if an appropriate initial estimator is available. Then we illustrate that a l_1 penalized SVM provides a proper initial estimator even for ultra high dimensional data. Numerical examples provide supportive evidence.

4.2 Introduction

Support vector machines (SVMs), originally introduced by [61, 62] and subsequently investigated by many others, are a popular and highly powerful technique for classification and have a solid mathematical foundation in statistical learning. In modern applications, we often face the problem of classification at the presence of a very large number of redundant features. For example, in genomics it is of fundamental importance to build a classifier using a small number of genes from thousands of candidate genes for the purpose of disease diagnosis and drug discovery; in spam email classification, it is desirable to build an accurate classier using a relatively small number of words from a dictionary that contains a huge number of different words. For such applications, although the standard l_2-norm SVM avoids overfitting to some extent, it does not automatically have dimension reduction of feature space built in. Hence, it usually does not yield an interpretable sparse decision rule. Furthermore, numerical evidence in the literature (e.g.,[63]) suggests that including many redundant features may seriously impair the generalization performance of l_2-norm SVMs.

The standard l_2-norm SVM has the well known *hinge loss*+l_2 *norm penalty* formulation. One effective way to preform simultaneous variable selection and classification using SVM is to replace the l_2 norm penalty with the l_1 norm penalty, which results in the l_1-norm SVM. See the earlier work of [64, 65]. Important advancement on the methodology and theory of l_1-norm SVMs has been obtained in recent years. Interested readers may consult [63] who proposed a path-following algorithm and effectively demonstrated the advantages of l_1-norm SVMs in high-dimensional sparse scenario, [66] who investigated the adaptivity of SVMs with l_1 penalty and derived its adaptive rates, [67] who obtained an oracle inequality involving both model complexity and margin for l_1-norm SVMs, [68] who extended the l_1-norm SVM to multi-class classification problems, [69] who proposed to use adaptive l_1 penalty with the SVM, and [70] who considered l_1-norm SVMs with a built-in reject option, among others.

Though the convex l_1 -penalty can also induce sparsity, it is well known that its variable selection consistency in linear regression relies on the stringent irrepresentability condition on the design matrix. This condition, however, can easily be violated in practice; see the examples in [8, 71]. Moreover, the regularization parameter for model

selection consistency in this case is not optimal for prediction accuracy [9, 72]. Hence we still turn to consider the penalized SVM with a general class of non-convex penalties, such as the SCAD penalty [2] or the minimax concave penalty (MCP) [7]. For the non-convex penalty, [11] investigated the oracle property of SCAD-penalized least squares regression in the high dimensions. However, a different set of proving techniques is needed for the non-convex penalized SVMs because the hinge loss in the SVM is not a smooth function. The Karush-Kuhn-Tucker local optimality condition is generally not sufficient for the set-up of a non-smooth loss plus a non-convex penalty. A new sufficient optimality condition based on subgradient calculation is used in the technical proof in this paper. We prove that under some general conditions, with probability tending to 1, the oracle estimator is a local minimizer of the non-convex penalized SVM objective function where the number of variables may grow exponentially with the sample size. By oracle estimator, we mean an estimator obtained by minimizing the empirical hinge loss with only relevant covariates. As one referee pointed out, with a finite sample, the empirical hinge loss may have multiple minimizers because the objective function is piecewise linear. This issue will vanish asymptotically because we assume that the population hinge loss has a unique minimizer. Such an assumption on the population hinge loss has been made in the existing literature [73].

Even though non-convex penalized SVMs are shown to enjoy the aforementioned local oracle property, it is largely unknown whether numerical algorithms can identify this local minimizer since the objective function involved is non-convex and typically multiple local minimizers exist. Existing methods rely heavily on conditions that guarantee that the local minimizer is unique. In general, when the convexity of the hinge loss function dominates the concavity of the penalty, the non-convex penalized SVM actually has a unique minimizer due to global convexity. Recently [12] gave sufficient conditions for a unique minimizer of the non-convex penalized least square regression when global convexity is not satisfied. However, for ultrahigh dimensional cases, it would be unrealistic to assume the existence of a unique local minimizer. See [56] for relevant discussion and a possible solution to non-convex penalized regression.

In this chapter, we further address the non-uniqueness issue of local minimizers by verifying that, with probability tending to 1, the LLA algorithm is guaranteed to yield

an estimator with the desired oracle property in merely two iterations under the localizability condition [26]. This convergence result extends the work of [26] by relaxing the differentiability assumption of the loss function and holds in the ultra high dimensional setting with $p = o(\exp(n\delta))$ for some positive constant δ. We further show that the localizability condition is automatically valid if an appropriate initial estimator is chosen. Finally we prove that the l_1 penalized SVM estimator is a proper initial estimator to enable the LLA algorithm to find the oracle estimator under ultra high dimensional settings.

4.3 Oracle Property of Non-convex Penalized Support Vector Machine

We begin with the basic set-up and notation. In binary classification, we are typically given a random sample $\{(Y_i, \mathbf{X}_i)\}_n^{i=1}$ from an unknown population distribution $P(\mathbf{X}, Y)$. Here $Y_i \in \{1, -1\}$ denotes the categorical label and $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^\intercal = (X_{i0}, (\mathbf{X}_{i-})^\intercal)^\intercal$ denotes the input covariates with $X_{i0} = 1$ corresponding to the intercept term. The goal is to estimate a classification rule that can be used to predict output labels for future observations with input covariates only. With potentially varying misclassification cost specified by weight $W_i = w$ if $Y_i = 1$ and $W_i = 1 - w$ if $Y_i = -1$ for some $0 < w < 1$, the linear weighted SVM [74] estimates the classification boundary by solving

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+ + \lambda \boldsymbol{\beta}_-^\intercal \boldsymbol{\beta}_-$$

where $(1 - u)_+ = \max\{1 - u, 0\}$ denotes the hinge loss, $\lambda > 0$ is a regularization parameter and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}_-)^\intercal)^\intercal$ with $\boldsymbol{\beta}_- = (\beta_1, \beta_2, \ldots, \beta_p)$. The standard SVM is a special case of the weighted SVM with weight parameter $w = 0.5$. In this chapter, we consider the standard SVM without losing generaity. The problem becomes

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+ + \lambda \boldsymbol{\beta}_-^\intercal \boldsymbol{\beta}_-$$

In general, the corresponding decision rule, $sign(\mathbf{X}^\intercal \boldsymbol{\beta})$, uses all covariates and is not capable of selecting relevant covariates.

Towards variable selection for the linear weighted SVM, we consider the population linear hinge loss $E(1 - Y\mathbf{X}^\intercal\boldsymbol{\beta})_+$. Let $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^\intercal = (\beta_1^*, (\boldsymbol{\beta}_-^*)^\intercal)^\intercal$ denote the true parameter, which is defined as the minimizer of the population weighted hinge loss, namely

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E(1 - Y\mathbf{X}^\intercal\boldsymbol{\beta})_+ \tag{4.1}$$

Still, the number of covariates $p = p_n$ is allowed to increase with the sample size n. It is even possible that p_n is much larger than n. Again, we assume the true parameter $\boldsymbol{\beta}^*$ is sparse. Let $A = \{1 \leq j \leq p_n; \beta_j^* \neq 0\}$ be the index set of the nonzero coefficients. Let $q = q_n = |A|$ be the cardinality of set A, which is also allowed to increase with n. For convenience, we assume that $p_n - q_n$ components of $\boldsymbol{\beta}^*$ are 0, *i.e.* $\boldsymbol{\beta}^{*\intercal} = (\boldsymbol{\beta}_1^{*\intercal}, \mathbf{0}^\intercal)$. Correspondingly, we write $\mathbf{X}_i^\intercal = (\mathbf{Z}_i^\intercal, \mathbf{R}_i^\intercal)$, where $\mathbf{Z}_i = (X_{i0}, X_{i1}, \ldots, X_{iq})^\intercal = (1, (\mathbf{Z}_{i-})^\intercal)^\intercal$ and $\mathbf{R}_i = (X_{i[q+1]}, \ldots, X_{ip})^\intercal$. Further we denote π_+ and π_- respectively to be the marginal probability of the label $Y = 1$ and $Y = -1$.

To facilitate our theoretical analysis, we introduce the gradient vector and Hessian matrix of the population linear weighted hinge loss. Let $L(\boldsymbol{\beta}_1) = E(1 - Y\mathbf{Z}^\intercal\boldsymbol{\beta}_1)_+$ be the population linear hinge loss by only relevant covariates. Define $S(\boldsymbol{\beta}_1) = (S(\boldsymbol{\beta}_1)_j)$ to be the $q_n + 1$-dimensional vector given by

$$S(\boldsymbol{\beta}_1) = -E(I(1 - Y\mathbf{Z}^\intercal\boldsymbol{\beta}_1 \geq 0)Y\mathbf{Z})$$

where $I(\cdot)$ is the indicator function. Also define $H(\boldsymbol{\beta}_1) = (H(\boldsymbol{\beta}_1)_{jk})$ to be the $(q_n + 1) \times (q_n + 1)$ matrix given by

$$H(\boldsymbol{\beta}_1) = E(\delta(1 - Y\mathbf{Z}^\intercal\boldsymbol{\beta}_1)\mathbf{Z}\mathbf{Z}^\intercal),$$

where $\delta(\cdot)$ is the Dirac delta function. It can be shown, $S(\boldsymbol{\beta}_1)$ and $H(\boldsymbol{\beta}_1)$ can be considered to be the gradient vector and Hessian matrix of $L(\boldsymbol{\beta}_1)$ respectively. See Lemma 2 of [73] for details.

4.3.1 Non-convex Penalized Support Vector Machine

By acting as if the true sparsity structure is known in advance, the oracle estimator is defined as $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\intercal, \mathbf{0}^\intercal)^\intercal$, where

$$\tilde{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}_1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\mathbf{Z}_i^\intercal\boldsymbol{\beta}_1)_+. \tag{4.2}$$

Here the objective function is piecewise linear. With a finite sample, it may have multiple minimizers. In that case, $\tilde{\boldsymbol{\beta}}_1$ can be chosen to be any minimizer. Our forthcoming theoretical results still hold. In the limit as $n \to \infty$, $\tilde{\boldsymbol{\beta}}_1$ minimizes the population version of the objective function $E(1 - Y\mathbf{Z}^\intercal\boldsymbol{\beta}_1)_+$. [75] shows that, when the misclassification cost are equal, the minimizer of $E(1 - Yf(\mathbf{Z}))_+$ over measurable $f(\mathbf{Z})$ is the Bayes rule $sign(p(\mathbf{z}) - \frac{1}{2})$, where $p(\mathbf{z}) = P(Y = 1|\mathbf{Z} = \mathbf{z})$. This suggests that the oracle estimator is aiming at approximating the Bayes rule. In practice, achieving an estimator with the desired oracle property is very challenging, because the sparsity structure of the true parameter $\boldsymbol{\beta}^*$ is largely unknown. Later we shall show that, under some regularity conditions, our proposed algorithm can find an estimator with oracle property and we claim convergence with high probability. Indeed, the numerical examples in Section 4.5 demonstrate that the estimator selected by our proposed algorithm has performance that is close to that of the Bayes rule. Note that the Bayes rule is unattainable here because we assume no knowledge on the high dimensional conditional density $P(\mathbf{X}|Y)$.

In this chapter, we consider the non-convex penalized hinge loss objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta})_+ + \sum_{j=1}^{p_n} p_\lambda(|\beta_j|), \tag{4.3}$$

where $p_n(\cdot)$ is a symmetric penalty function with tuning parameter λ. Let $p'_\lambda(t)$ be the derivative of $p_\lambda(t)$ with respect to t. We consider a general class of non-convex penalties that satisfy the following conditions.

(A1) The symmetric penalty $p_\lambda(t)$ is assumed to be non-decreasing and concave for $t \in [0, \infty)$, with a continuous derivative $p'_\lambda(t)$ on $(0, \infty)$ and $p_\lambda(0) = 0$.

(A2) There exists $a > 1$ such that $\lim_{t\to 0+} p'_\lambda(t) = \lambda$, $p'_\lambda(t) \geq \lambda - t/a$ for $0 < t < a\lambda$ and $p'_\lambda(t) = 0$ for $t \geq a\lambda$

The motivation for such a non-convex penalty is that the convex l_1 penalty lacks the oracle property owing to the overpenalization of large coefficients in the model selected. Consequently it is undesirable to use the l_1 penalty when the purpose of the data analysis is to select the relevant covariates among potentially high dimensional candidates in classification. Note that p, q, λ and other related quantities are allowed to depend on n, and we suppress the subscript n whenever there is no confusion. As expected, we

will use the two common non-convex penalties which satisfy the assumptions (A1) and (A2): SCAD and MCP.

4.3.2 Oracle Property

To facilitate our technical proofs, we impose the following regularity conditions.

(C1) The density of $\mathbf{Z}^*$ given $Y = 1$ and $Y = -1$ are continuous and have common support in $\mathbb{R}^q$;

(C2) $E(X_j^2) < \infty$ for $1 \leq j \leq q$;

(C3) The true parameter $\boldsymbol{\beta}^*$ is unique and a nonzero vector;

(C4) $q_n = O(n^{c_1})$, namely $\lim_{n\to\infty} q_n/n^{c_1} < \infty$, for some $0 \leq c_1 < \frac{1}{3}$;

(C5) There is a constant $M_1 > 0$ such that $\lambda_{max}(\frac{1}{n}\mathcal{X}_A^\intercal \mathcal{X}_A) \leq M_1$, where $\mathcal{X}_A$ is the first $q_n + 1$ columns of the design matrix and λ_{max} denotes the largest eigenvalue. It is further assumed that $\max_{1\leq i\leq n} ||\mathbf{Z}_i|| = O_p(\sqrt{q_n \log n})$, $(\mathbf{Z}_i, Y_i)$ are in general position [76, Section 2.2] and X_{ij} are sub-Gaussian random variables for $1 \leq i \leq n, q_n + 1 \leq p_n$;

(C6) $\lambda_{min}(H(\boldsymbol{\beta}_1^*)) \geq M_2$ for some constant $M_2 > 0$, where λ_{min} denotes the smallest egienvalue;

(C7) $n^{(1-c_2)/2} \min_{1\leq j\leq q_n} |\beta_j^*| \leq M_3$ for some constant $M_3 > 0$ and $2c_1 < c_2 \leq 1$;

(C8) Denote the conditional density of $\mathbf{Z}^\intercal \boldsymbol{\beta}_1^*$ given $Y = 1$ and $Y = -1$ as f and g respectively. It is assumed that f is uniformly bounded away from 0 and ∞ in a neighborhood of 1 and g is uniformly bounded away from 0 and ∞ in a neighborhood of -1.

Remark 1. Conditions (C1)-(C3) and (C6) were also assumed for fixed p in [73]. We need these assumptions to ensure that the oracle estimator is consistent in the scenario of diverging p. Condition (C3) states that the optimal classification decision function is not constant, which is required to ensure that $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ are a well-defined gradient vector and Hessian matrix of the hinge loss; see Lemma 2 and Lemma 3 of [73].

Conditions (C4) and (C7) are common in the literature of high dimensional inference[11]. More specifically, condition (C4) states that the divergence rate of the number of nozero coefficients cannot be faster than $\sqrt{n}$ and condition (C7) simply states that the signals cannot decay too quickly. The condition on the largest eigenvalues of the design matrix in condition (C5) is similar to the sparse Riesz condition and was also assumed in [77], [78] and [7]. Note that the bound on the smallest eigenvalue is not specified. The condition on the maximum norm in condition (C5) holds when $\mathbf{Z}_-$ given Y follows a multivariate normal distribution. $(\mathbf{Z}_i, Y_i)$ are in general position if with probability 1 there are exactly $q_n + 1$ elements in $D = \{i : 1 - Y_i \mathbf{Z}_i^\intercal \tilde{\boldsymbol{\beta}}_1 = 0\}$ [76, Section 2.2]. The condition for general position is true with probability 1 with respect to Lebesgue measure. Condition (C8) requires that there is enough information around the non-differentiable point of the hinge loss, similarly to condition (C3) in [14] for quantile regression.

For illustrative examples that satisfy all the above conditions, assume that $0 < \pi_+ = \pi_- < 1$ and let the number of signals be fixed. The first example is that the conditional distributions of $\mathbf{X}_-$ given Y have unbounded support $\mathbb{R}^p$ with sub-Gaussian tails. It can be easily seen that the Fisher discriminant analysis is one special case when $\mathbf{X}_-$ given Y are Gaussian. C1-C4 and C7 are trivial. C5 holds by the properties of sub-Gaussian random variables. [73] showed that C6 holds if the supports of the conditional densities of $\mathbf{Z}_-$ given Y are convex, which are naturally satisfied in $\mathcal{R}^q$. C8 is trivially satisfied by the unbounded support of the conditional distribution of $\mathbf{Z}_-$ given Y. Another example is the probit model that $\mathbf{X}_-$ has unbounded support $\mathbb{R}^p$ with sub-Gaussian tails and $Pr(Y = 1|\mathbf{X}_-) = \Phi(\mathbf{X}^\intercal \boldsymbol{\beta})$ for some $\boldsymbol{\beta} \neq \mathbf{0}$. It can be easily checked that the conditional distributions of $\mathbf{X}_-$ given Y also have unbounded supports $\mathbb{R}^p$ and hence all the conditions are satisfied.

Now In this subsection, we establish the theory of the local oracle property for the non-convex penalized SVMs, namely the oracle estimator is one of the local minimizers of the objective function $Q(\boldsymbol{\beta})$ defined in equation (4.3).

Theorem 4.3.1. *Assume that conditions (C1)-(C8) hold. Let* $\mathbf{B}_n(\lambda)$ *be the set of local minimizers of the objective function* $Q_(\boldsymbol{\beta})$ *with regularization parameter* λ*. The oracle*

estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\intercal, \mathbf{0}^\intercal)^\intercal$ *satisfies*

$$Pr(\tilde{\boldsymbol{\beta}} \in \mathbf{B}_n(\lambda)) \to 1$$

as $n \to 1$, *if* $\lambda = o(n^{-(1-c_1)/2})$, *and* $\log(p)q\log(n)n^{-1/2} = o(\lambda)$.

It can be shown that, if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < c_2/2$, then the oracle property holds even for $p = o(\exp(n^{(\delta-c_1)/2}))$. Therefore, the local oracle property holds for the non- convex penalized SVM even when the number of covariates grows exponentially with the sample size.

4.3.3 Implementation and Tuning

To solve the non-convex penalized SVMs, we use the LLA algorithm as discussed before. In details, we start with an initial value $\widehat{\boldsymbol{\beta}}^{(0)}$. At each step $t \geq 1$, we update by solving

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta})_+ + \sum_{j=1}^{p} p'_\lambda(|\widehat{\beta}_j^{(t-1)}|)|\beta_j| \right), \tag{4.4}$$

where $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$. Following the literature, when $\widehat{\beta}_j^{(t-1)} = 0$, we take $p'_\lambda(0)$ to be $p'_\lambda(0+) = \lambda$. The LLA algorithm is an instance of the majorization-minimization algorithm and converges to a local minimizer of the non-convex objective function.

With slack variables, the convex optimization problem (4.4) can be easily recast as a linear programming problem

$$\min_{\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{\beta}} \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{p} p'_\lambda(|\widehat{\beta}_j^{(t-1)}|)\eta_j \right)$$

subject to

$$\begin{aligned}
\xi_i &\geq 0, \quad i = 1, 2, \ldots, n, \\
\xi_i &\geq 1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \\
\eta_j &\geq \beta_j,\ \eta_j \geq -\beta_j, \quad i = 1, 2, \ldots, p.
\end{aligned}$$

We stop the algorithm when $\sum_{j=1}^{p}(p'_\lambda(|\widehat{\beta}_j^{(t-1)}|) - p'_\lambda(|\widehat{\beta}_j^{(t)}|))^2$ is sufficiently small.

For the choice of tuning parameter λ, [79] suggested that the SVM information criterion SVMIC which, for a subset S of $\{1, 2, \ldots, p\}$, is defined as

$$\text{SVMIC}(S) = \sum_{i=1}^{n} \xi_i + \log(n)|S|,$$

where $|S|$ is the cardinality of S and ξ_i, $i = 1, 2, \ldots, n$, denote the corresponding slack variables. This criterion directly follows the spirit of the Bayesian information criterion BIC by [80]. [54] showed that BIC can be too liberal when the model space is large and proposed the extended BIC

$$\text{EBIC}_\gamma(S) = -2\text{logLikelihood} + \log(n)|S| + 2\gamma \binom{p}{|S|}, \quad 0 \leq \gamma \leq 1.$$

By combining these ideas, [60] suggest the SVM-extended BIC

$$\text{SVMIC}_\gamma = \sum_{i=1}^{n} 2\xi_i + \log(n)|S| + 2\gamma \binom{p}{|S|}, \quad 0 \leq \gamma \leq 1.$$

We use $\gamma = 0.5$ as suggested by [54] and choose the λ that minimize SVMIC_γ.

4.3.4 The LLA Algorithm with Convergence to Oracle Estimator

In this subsection we need to verify the validity of the LLA algorithm in non-convex penalized SVMs. Theorem 4.3.1 indicates that one of the local minimizers has the oracle property. However, there can potentially be multiple local minimizers and it remains challenging to identify the oracle estimator. In the high dimensional setting, assuming that the local minimizer is unique would not be realistic.

In this subsection, instead of assuming the uniqueness of solutions, we work directly on the conditions under which the oracle estimator can be identified by some numerical algorithms that solve the non-convex penalized SVM objective function. As we suggested in last section, the LLA algorithm remains to be a handy tool to solve our problem. Recently the LLA has been shown to be capable of identifying the oracle estimator in the set-up of folded concave penalized estimation with a differentiable loss function [26, 56]. We generalize their results to non-differentiable loss functions, so that they can fit in the framework of the non-convex penalized SVMs. Similarly to their work, the main condition required is the existence of an appropriate initial estimator

inputted in the iterations of the LLA algorithm. Denote the initial estimator as $\widehat{\boldsymbol{\beta}}^{(0)}$. Intuitively, if the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ lies in a small neighborhood of the true value $\boldsymbol{\beta}^*$, the algorithm should converge to the good local minimizer around $\boldsymbol{\beta}^*$. This localizability will be formalized in terms of l_∞ distance later. With such an appropriate initial estimator, under the aforementioned regularity conditions, one can prove that the LLA algorithm converges to the oracle estimator with probability tending to 1 even in ultrahigh dimensions.

Let $\widehat{\boldsymbol{\beta}}^{(0)} = (\widehat{\beta}_0^{(0)}, \ldots, \widehat{\beta}_p^{(0)})^\intercal$. Consider the following events:

(i) $F_{n1} = \{|\widehat{\beta}_j^{(0)} - \beta_j^*| > \lambda, \text{for some } 1 \leq j \leq p\}$;

(ii) $F_{n2} = \{|\beta_j^* < (a+1)\lambda|\}, \text{for some } 1 \leq j \leq q\}$;

(iii) $F_{n3} = \{\text{for all subgradients } s(\tilde{\boldsymbol{\beta}}), |s_j(\tilde{\boldsymbol{\beta}})| > (1 - 1/a)\lambda \text{ for some } q+1 \leq j \leq p \text{ or } |s_j(\tilde{\boldsymbol{\beta}})| \neq 0 \text{ for some } 0 \leq j \leq q\}$, where $s(\tilde{\boldsymbol{\beta}}) = (s_0(\tilde{\boldsymbol{\beta}}), \ldots, s_p(\tilde{\boldsymbol{\beta}}))$ with

$$s_j(\tilde{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^{n} Y_i X_{ij} I(1 - Y_i \mathbf{X}_i^\intercal \tilde{\boldsymbol{\beta}} > 0) - \frac{1}{n}\sum_{i=1}^{n} Y_i X_{ij} v_j,$$

where $-1 \leq v_i \leq 0$ if $1 - Y_i \mathbf{X}_i^\intercal \tilde{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise, $j = 0, \ldots, p$;

(iv) $F_{n4} = \{|\tilde{\beta}_j| < a\lambda, \text{ for some } 1 \leq j \leq q\}$.

Denote the corresponding probability as $P_{ni} = Pr(F_{ni}), i = 1, 2, 3, 4$. P_{n1} represents the localizability of the problem. When we have an appropriate initial estimator, we expect P_{n1} to converge to 0 as $n \to +\infty$. P_{n2} is the probability that the true signal is too small to be detected by any method. P_{n3} describes the behaviour of the subgradients at the oracle estimator. P_{n4} is concerned with the magnitude of the oracle estimator on relevant variables. Under regularity conditions, the oracle estimator will detect the true signals and hence P_{n4} will be very small.

Now we can have the following theorem for the LLA algorithm to identify the oracle estimator $\tilde{\boldsymbol{\beta}}$ in the non-convex penalized SVMs based on P_{n1}, P_{n2}, P_{n3} and P_{n4}.

Theorem 4.3.2. *With probability at least* $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$, *the LLA algorithm initiated by* $\widehat{\boldsymbol{\beta}}^{(0)}$ *finds the oracle estimator* $\tilde{\boldsymbol{\beta}}$ *after two iterations. Furthermore, if conditions (C1)-(C8) hold,* $\lambda = o(n^{-(1-c_2)/2})$ *and* $\log(p) q \log(n) n^{-1/2} = o(\lambda)$, *then* $P_{n2} \to 0$, $P_{n3} \to 0$ *and* $P_{n4} \to 0$ *as* $n \to \infty$.

The first part of Theorem 4.3.2 provides a non-asymptotic lower bound on the probability that the LLA algorithm converges to the oracle estimator. As we shall show in Appendix B, if none of the events F_{ni} happen, the LLA algorithm initiated with $\widehat{\boldsymbol{\beta}}^{(0)}$ will find the oracle estimator in the first iteration, and in the second iteration it will find the oracle estimator again and thus claim convergence. Only a single correction is required in the first iteration and the second iteration is needed to stop the algorithm. Therefore, the LLA algorithm can identify the oracle estimator after two iterations and this result holds generally without conditions (C1)-(C8).

The second part of Theorem 4.3.2 indicates that, under conditions (C1)-(C8), the lower bound is determined only by the limiting behavior of the initial estimator. As long as an appropriate initial estimator is available, the problem of selecting the oracle estimator from potential multiple local minimizers is addressed.

4.4 Error Bound for l_1 Penalized Support Vector Machine

In this section, we study the asymptotic behavior of the estimated l_1-norm SVM coefficients in the ultra-high dimension and derive that the error bound is of near-oracle rate $O(\sqrt{q \log p/n})$, with q being the number of features with nonzero coefficients and n is the sample size. As an important application, we show that this result helps greatly extend the applicability of the recent algorithm and theory of high-dimensional nonconvex-penalized SVM [60] by providing a statistically valid and computationally convenient initial value. This will solve our final puzzle and validate the application of the LLA algorithm on non-convex penalized SVMs.

Explicitly, the use of nonconvex penalty function aims to further reduce the bias associated with the l_1 penalty and accurately identify the set of relevant features for classification. However, the presence of nonconvex penalty results in computational complexity. [60] proposed an algorithm and showed that given an appropriate initial value, in two iterative steps the algorithm is guaranteed to produce an estimator that possesses the oracle property in the ultra-high dimension and consequently with probability approaching one the zero coefficients are estimated as exactly zero. However, the availability of a qualified initial estimator is itself a challenging issue in high dimension. [60] provided an initial estimator that would satisfy the requirement when $p = o(\sqrt{n})$.

We shows that the l_1-SVM can be a valid initial estimator under general conditions when p grows at an exponential rate of n, which completes the algorithm and theory of [60] in last section.

4.4.1 l_1-norm support vector machine

We inherit the notations and definitions above. The standard linear SVM can be expressed as the following regularization problem

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+ + \lambda ||\boldsymbol{\beta}_-||_2^2, \tag{4.5}$$

where $(1-u)_+ = \max\{1-u, 0\}$ is often called the hinge loss function, λ is a tuning parameter and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}_-)^\intercal)^\intercal$ with $\boldsymbol{\beta}_- = (\beta_1, \beta_2, \ldots, \beta_p)^\intercal$. Generally for a given vector $\mathbf{e}$, we use $\mathbf{e}_-$ to denote the subvector with the first entry of $\mathbf{e}$ omitted. Actually, optimization problem in (4.5) is known as the primal problem of the SVM, which can be converted into an equivalent dual problem for further solution.

The l_1-norm SVM replaces the l_2 penalty in (4.5) by the l_1 penalty. That is, we consider the objective function

$$l_n(\boldsymbol{\beta}, \lambda) = n^{-1} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta})_+ + \lambda ||\beta_-||_1, \tag{4.6}$$

and define

$$\widehat{\boldsymbol{\beta}}(\lambda) = \operatorname*{argmin}_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta}, \lambda). \tag{4.7}$$

For a given data point X_i, it is classified into class + (corresponding to $\hat{Y}_i = 1$) if $\mathbf{X}_i^\intercal \widehat{\boldsymbol{\beta}}(\lambda) > 0$ and into class − (corresponding to $\hat{Y}_i = -1$) if $\mathbf{X}_i^\intercal \widehat{\boldsymbol{\beta}}(\lambda) < 0$.

By introducing the slack variables, we can transform our optimization problem (4.7) as a linear programming problem

$$\begin{aligned} \min_{\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\beta}} \quad & \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} \zeta_j \right) \\ \text{subject to} \quad & \xi_i \geq 0, \qquad i = 1, 2, \ldots, n, \\ & \xi_i \geq 1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}, \qquad i = 1, 2, \ldots, n, \\ & \zeta_j \geq \beta_j, \zeta_j \geq -\beta_j, \qquad j = 1, 2, \ldots, p. \end{aligned} \tag{4.8}$$

Same case as Section 4.3.3, we can utilize the LLA algorithm to solve this problem. Several R packages have been designed to solve such a standard problem, such as `lpSolve` in the core and `linprog`.

4.4.2 The Choice of the Tuning Parameter λ and a Fact About $\widehat{\boldsymbol{\beta}}$

The key result in this section is an error bound of $||\widehat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^*||$, where $\boldsymbol{\beta}^*$ is the minimizer of the population version of the hinge loss function, that is,

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\beta}), \tag{4.9}$$

where $L(\boldsymbol{\beta}) = E(1 - Y\mathbf{X}^\intercal\boldsymbol{\beta})_+$. [75] suggested that there is a close connection between the minimizer of the population hinge loss function and the Bayes rule. The definition of $\boldsymbol{\beta}^*$ above is also used in [16, 73], both of which only considered the fixed p case. We are interested in the error bound of $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||$ when $p \gg n$. In the ultra-high dimensional settings, it is often reasonable to assume that $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\intercal$ is sparse in the sense that most of its components are exactly zero. We define the index set of active features as $T = \{1 \leq j \leq p : \beta_j^* \neq 0\}$. We denote the cardinality of T by $|T| = q$. To incorporate the intercept term, we also define $T_+ = T \bigcup \{0\}$.

Next, we introduce the gradient vector and Hessian matrix of the population hinge loss function $L(\boldsymbol{\beta})$. We define

$$S(\boldsymbol{\beta}) = -E(I(1 - Y\mathbf{X}^\intercal\boldsymbol{\beta} \geq 0)Y\mathbf{X})$$

as the $(p+1)$-dimensional gradient vector and

$$H(\boldsymbol{\beta}) = E(\delta(1 - Y\mathbf{X}^\intercal\boldsymbol{\beta})\mathbf{X}\mathbf{X}^\intercal)$$

as the $(p+1) \times (p+1)$-dimensional Hessian matrix where $I(\cdot)$ is the indicator function and $\delta(\cdot)$ is the Dirac delta function. Section 6.1 of [73] has explained more details of the explicit forms of $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ under certain conditions.

Throughout this section, we assume the following regularity condition.

(B1) The densities of $\mathbf{X}_-$ given $Y = \pm 1$ are continuous and have common support in $\mathbb{R}^{p+1}$, and there exists a constant $M > 0$ such that $|X_j| \leq M$, $j \in \{1, \ldots, p\}$.

Remark 2. Condition (B1) ensures that $H(\boldsymbol{\beta})$ is well defined and continuous in $\boldsymbol{\beta}$. The bound of $\mathbf{X}_-$ can be relaxed with further technical complexity. More details can be found in [16, 73].

The estimated l_1-norm SVM parameter $\widehat{\boldsymbol{\beta}}(\lambda)$ defined in (4.7) depends on the tuning parameter λ. We will first show that a universal choice

$$\lambda = c\sqrt{2A(\alpha)\log p/n}, \tag{4.10}$$

where c is some given constant, α is a small probability and $A(\alpha) > 0$ is a constant such that $4p^{-\frac{A(\alpha)}{M^2}+1} \leq \alpha$, can provide theoretical guarantee on the good performance of $\widehat{\boldsymbol{\beta}}(\lambda)$.

The above choice of λ is motivated by a principle in the setting of penalized least squares regression [29], which advocates to choose the penalty level λ to dominate the subgradient of the loss function evaluated at the true value. Intuitively, the subgradient evaluated at $\boldsymbol{\beta}^*$ summaries the estimation noise. See also the application of the same principle to choose the penalty level for quantile regression [28, 38]. Another more technical motivation of this principle comes from the KKT condition in convex optimization theory. Let $\tilde{\boldsymbol{\beta}}$ be the oracle estimator (formally defined the following section) that minimizes the sample hinge loss function when the index set T is known in advance. Define the subgradient function

$$\widehat{S}(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n} I(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta} \geq 0)Y_i\mathbf{X}_i.$$

Then it follows from the argument as in Theorem 3.1 of [60] that under some weak regularity conditions $||\widehat{S}(\tilde{\boldsymbol{\beta}})||_\infty \leq \lambda$ with probability approaching one. It follows from [73] that the oracle estimator $\tilde{\boldsymbol{\beta}}$ provides a consistent and asymptotic normal estimate of $\boldsymbol{\beta}^*$.

Hence, in the ideal case where the population parameter $\boldsymbol{\beta}^*$ is known, an intuitive choice of λ is to set its value to be larger than the supremum norm of $\widehat{S}(\boldsymbol{\beta}^*)$ with large probability, that is

$$P(\lambda \geq c||\widehat{S}(\beta^*)||_\infty) \geq 1 - \alpha, \tag{4.11}$$

where $c > 1$ is some given constant and α is a small probability. Lemma 4.4.1 below shows that the choice of λ given in (4.10) satisfies this requirement.

Lemma 4.4.1. *Assume that condition (B1) is satisfied. Suppose* $\lambda = c\sqrt{2A(\alpha)\log p/n}$, *we have*

$$P(\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty) \geq 1 - \alpha$$

with α *being a given small probability defined earlier in this section.*

The proof of Lemma 4.4.1 is given in the Appendix B. The crux of the proof is to bound the tail probability of $\sum_{i-1}^{n} I(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^* \geq 0)Y_i\mathbf{X}_i$ by applying Hoeffding's inequality and the union bound. Later in this section, we will show that this choice of λ warrants near-oracle rate performance of $\widehat{\boldsymbol{\beta}}(\lambda)$.

Let $\mathbf{h} = \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}(\lambda)$. We state below an interesting fact on $\mathbf{h}$.

Lemma 4.4.2. *For* $\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty$ *and* $\bar{C} = \frac{c-1}{c+1}$, *we have*

$$\mathbf{h} \in \Delta_{\bar{C}},$$

where

$$\Delta_{\bar{C}} = \big\{\boldsymbol{\gamma} \in \mathbf{R}^{p+1} : ||\boldsymbol{\gamma}_{T_+}||_1 \geq \bar{C}||\boldsymbol{\gamma}_{T_+^c}||_1, \textit{where } T_+ = T \cup \{0\},\ T \subset \{1, 2, \ldots, p\} \textit{ and } |T| \leq q\big\},$$

with T_+^c *denoting the complement of* T_+, *and* $\boldsymbol{\gamma}_{T_+}$ *denoting the* $(p+1)$*-dimensional vector that has the same coordinates as* $\boldsymbol{\gamma}$ *on* T_+ *and zero coordinates on* T_+^c.

We call $\Delta_{\bar{C}}$ the *restricted set*. The proof of (4.4.2) is given in the Appendix B.

4.4.3 Regularity conditions

Let $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ denote the feature design matrix. We define *restricted eigenvalues* as follows

$$\lambda_{max} = \max_{\mathbf{d} \in \mathbb{R}^{p+1} : ||\mathbf{d}||_0 \leq q+1} \frac{\mathbf{d}^T\mathcal{X}^T\mathcal{X}\mathbf{d}}{n||\mathbf{d}||_2^2} \tag{4.12}$$

and

$$\lambda_{min}(H(\boldsymbol{\beta}^*); q) = \min_{\mathbf{d} \in \Delta_{\bar{C}}} \frac{\mathbf{d}^T H(\boldsymbol{\beta}^*)\mathbf{d}}{||\mathbf{d}||_2^2}. \tag{4.13}$$

These can be considered as extension of the sparse eigenvalue notion in [29] for analyzing l_1 penalized least squares regression and the Dantzig selector [3] and the restricted isometry constants in [81].

In addition to condition (B1) introduced in Section 4.4.2, we require the following regularity conditions for the main theory of this paper.

(B2) $q = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.

(B3) There exists a constant M_1 such that $\lambda_{max} \leq M_1$ almost surely.

(B4) $\lambda_{min}(H(\boldsymbol{\beta})^*; q) \geq M_2$, for some constant $M_2 > 0$.

(B5) $n^{(1-c_2)/2} \min_{j \in T} |\beta_j^*| \geq M_3$ for some constants $M_3 > 0$ and $2c_1 < c_2 \leq 1$.

(B6) Denote the conditional density of $\mathbf{X}^\intercal \boldsymbol{\beta}^*$ given $Y = +1$ and $Y = -1$ as f^* and g^*, respectively. It is assumed that f^* is uniformly bounded away from 0 and ∞ in a neighborhood of 1 and g^* is uniformly bounded away from 0 and ∞ in a neighborhood of -1.

Remark 2. Conditions (B2) and (B5) are very common in high dimensional literatures. Basically, condition (B2) states that the number of non-zero variables cannot diverge at a rate larger than $\sqrt{n}$. Condition (B5) controls the decay rate of true parameter $\boldsymbol{\beta}^*$. Condition (B3) is one of the *restricted eigenvalue* (RE) assumptions in [28, 29]. In our case, we only need an upper bound for this restricted eigenvalue. Condition (B4) requires the positive-definiteness of $H(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$. We provide a thorough discussion of this condition in Appendix C, including an example that demonstrates the validity of this condition. Condition (B6) warrants that there is sufficient information around the non-differentiable point of the hinge loss, similarly to condition (C3) in [14] for quantile regression.

4.4.4 An error bound of $\widehat{\boldsymbol{\beta}}(\lambda)$ in ultra-high dimension

Before stating the main theorem, we first present an important lemma, which has to do with the empirical process behavior of the hinge loss function.

Lemma 4.4.3. *Assume that conditions (B1)-(B2) and (B3) are satisfied. Let*

$$\begin{aligned} B(\mathbf{h}) \;=\; & \frac{1}{n}\Big| \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^* + Y_i \mathbf{X}_i^\intercal \mathbf{h})_+ - \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^*)_+ \\ & - E\left(\sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^* + Y_i \mathbf{X}_i^\intercal \mathbf{h})_+ - \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^\intercal \boldsymbol{\beta}^*)_+ \right) \Big|. \end{aligned}$$

Assume $p > n$ *and* $p > C_1\sqrt{q}$*, then*

$$P\left(\sup_{||\mathbf{h}||_0=q+1,||\mathbf{h}||_2=1} B(\mathbf{h}) \geq (1+2C_2\sqrt{2M_1})\sqrt{\frac{2q\log p}{n}}\right) \leq 2p^{-4q(C_2^2-1)},$$

where $C_2 > 1$ *and* C_1 *are constants.*

Lemma 4.4.3 guarantees that $n^{-1}\left(\sum_{i=1}^n(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+-\sum_{i=1}^n(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+\right)$ is close to its expected value with high probability. This provides an important tool to handle the nonsmoothness of the hinge loss function in proving the main theory, which is stated below.

Theorem 4.4.4. *Suppose that conditions (B1)-(B6) hold, then the estimated* l_1*-norm SVM coefficients vector* $\widehat{\boldsymbol{\beta}}(\lambda)$ *satisfies*

$$||\widehat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^*||_2 \leq \sqrt{1+\frac{1}{\bar{C}}}\Big(\frac{2\lambda\sqrt{q}}{M_2}+\frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\bar{C}})\Big)$$

with probability at least $1-2p^{-4q(C_2^2-1)+1}$*, where* C *is a constant.*

From this theorem, we can easily capture the nearly oracle property for l_1 penalized SVM estimator, such that with high probability,

$$||\widehat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^*||_2 = O_p\left(\sqrt{\frac{q\log p}{n}}\right)$$

when $\lambda = c\sqrt{2A(\alpha)\log p/n}$. Actually, in the inequality of Theorem 4.4.4, the first term satisfies $\frac{\lambda\sqrt{q}}{M_2} = \frac{2}{M_2}\sqrt{\frac{2A(\alpha)q\log p}{n}} = O\left(\sqrt{\frac{q\log p}{n}}\right)$ and it is also trivial to have the second term of the same order. Hence the nearly oracle property of $\widehat{\beta}(\lambda)$ will hold given λ above.

To numerically evaluate the above error bound of the L_1-norm SVM, we consider the simulation setting in Model 4 of Section 4.5.1. We choose $p = 0.1 * n^2$, $q = \lfloor n^{1/3} \rfloor$ and $\boldsymbol{\beta}^*_- = ((1.1,\ldots,1.1)_q, 0,\ldots,0)^T$, which allows p and q to vary with sample size n. Figure 4.1 depicts the average of $||\widehat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^*||_2$ across 200 simulation runs for different values of n for L_1-norm SVM and compare the curve with the theoretical error bound $(\sqrt{\frac{q\log p}{n}})$. We observe that these two curves display similar decreasing pattern and approach each other as n gets larger.

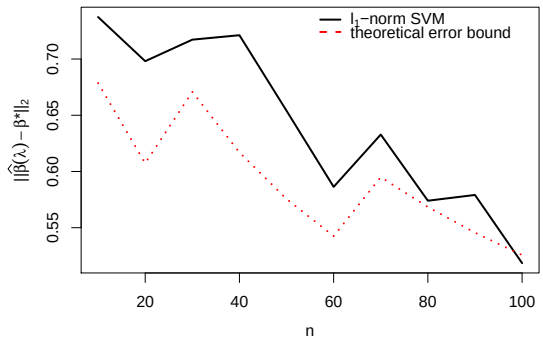

Figure 4.1: *L_2-norm estimation error comparison*

As we discussed before, Theorem 4.3.2 shows that if an appropriate initial estimator exists, *i.e.* $P_{n1} \to 0$, then under general regularity conditions, the LLA algorithm can identify the oracle estimator with probability approaching one in just two iterative steps. Now we have a complete systematic framework for non-convex penalized SVM in ultra high dimension. As the error bound that we derived on l_1 norm SVM ensures that the $\widehat{\boldsymbol{\beta}}$ is a qualified initial value, we states the following theorem for ultra high dimensional cases.

Theorem 4.4.5. *Assume $\widehat{\beta}(\lambda)$ is the solution to the l_1 penalized SVM with tuning parameter $\lambda = c\sqrt{2A(\alpha)\log p/n}$ defined above. Suppose that conditions (B1)-(B6) hold, then we have $P(|\widehat{\beta}_j(\lambda) - \beta_j^*| > \lambda,\ \text{for some } 1 \le j \le p) \to 0$ as $n \to \infty$. Furthermore, under the regular conditions stated in Theorem 4.3.2, the LLA algorithm initiated by $\widehat{\beta}(\lambda)$ finds the oracle estimator in two iterations with probability tending to 1, i.e., $P(\widehat{\beta}^{nc}(\lambda) = \tilde{\beta})$, where $\widehat{\beta}^{nc}(\lambda)$ is the solution for non-convex penalized SVM with given λ.*

4.5 Numerical Results

In this section, we will investigate the finite sample performance of the l_1-norm SVM. We will also study its application to non-convex penalized SVM in high dimension.

4.5.1 Monte Carlo results for l_1-norm SVM

We generate random data from each of the following four models.

- Model 1: $Pr(Y=1)=Pr(Y=-1)=0.5$, $\mathbf{X}_-|(Y=1)\sim MN(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{X}_-|(Y=-1)\sim MN(-\boldsymbol{\mu},\boldsymbol{\Sigma})$, $q=5$, $\boldsymbol{\mu}=(0.1,0.2,0.3,0.4,0.5,0,\ldots,0)^T\in\mathbb{R}^p$, $\boldsymbol{\Sigma}=(\sigma_{ij})$ with diagonal entries equal to 1, nonzero entries $\sigma_{ij}=-0.2$ for $1\le i\neq j\le q$ and other entries equal to 0. The Bayes rule is $\text{sign}(1.39X_1+1.47X_2+1.56X_3+1.65X_4+1.74X_5)$ with Bayes error 6.3%.

- Model 2: $Pr(Y=1)=Pr(Y=-1)=0.5$, $\mathbf{X}_-|(Y=1)\sim MN(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{X}_-|(Y=-1)\sim MN(-\boldsymbol{\mu},\boldsymbol{\Sigma})$, $q=5$, $\boldsymbol{\mu}=(0.1,0.2,0.3,0.4,0.5,0,\ldots,0)^T\in\mathbb{R}^p$, $\boldsymbol{\Sigma}=(\sigma_{ij})$ with $\sigma_{ij}=-0.4^{|i-j|}$ for $1\le i,j\le q$ and other entries equal to 0. The Bayes rule is $\text{sign}(3.09X_1+4.45X_2+5.06X_3+4.77X_4+3.58X_5)$ with Bayes error 0.6%.

- Model 3: model stays the same as Model 2, but $\boldsymbol{\Sigma}=(\sigma_{ij})$ with nonzero elements $\sigma_{ij}=-0.4^{|i-j|}$ for $1\le i,j\le q$ and $\sigma_{ij}=0.4^{|i-j|}$ for $q<i,j\le p$. The Bayes rule is still $sign(3.09X_1+4.45X_2+5.06X_3+4.77X_4+3.58X_5)$ with Bayes error 0.6%.

- Model 4: $\mathbf{X}_-\sim MN(\mathbf{0}_p,\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}=(\sigma_{ij})$ with nonzero elements $\sigma_{ij}=0.4^{|i-j|}$ for $1\le i,j\le p$, $Pr(Y=1|\mathbf{X}_-)=\Phi(\mathbf{X}_-^T\boldsymbol{\beta}_-^*)$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $\boldsymbol{\beta}_-^*=(1.1,1.1,1.1,1.1,0,\ldots,0)^T$ and $q=4$. The Bayes rule is $\text{sign}(1.1X_1+1.1X_2+1.1X_3+1.1X_4)$ with Bayes error 10.4%.

Model 1 and Model 4 are identical to the ones in [60]. In particular, Model 1 focuses on a standard linear discriminate analysis setting. On the other hand, Model 4 is a typical probit regression case. Models 2 and 3 are designed with autoregressive covariance as correlation decaying off-diagonal-wise. We consider sample size $n=100$ with $p=1000$ and 1500, and $n=200$ with $p=1500$ and 2000. Similarly as in [82], we use an independent tuning data set of size $2n$ to tune our λ by minimizing the prediction error using five-fold cross validation. The tuning range spans from 2^{-6} to 2 as equally-spaced sequence with 100 elements. For each simulation scenario, we conduct 200 runs. Then we generate an independent test data set of size n to report the estimated test error.

We evaluate the performance of l_1-norm SVM by its testing misclassification error rate, estimator error and variable selection ability. In particular, we measure the estimation accuracy by two criteria: the L_2 estimation error $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$ where Appendix C provides details on the calculation of $\boldsymbol{\beta}^*$ and the absolute value of the sample correlation $c(\cdot, \cdot)$ between $\mathbf{X}^T\widehat{\boldsymbol{\beta}}(\lambda)$ and $\mathbf{X}^T\boldsymbol{\beta}^*$. The absolute value of the sample correlation (AAC) is also used as accuracy measure in [83]. To summarize, we will report

- **Test error**: The misclassification error rate.
- L_2 **error**: $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$.
- **AAC**: Absolute absolute correlation $corr(\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{X}^\intercal\boldsymbol{\beta}^*)$.
- **Signal**: the average of number of non-zero regression coefficients $\widehat{\boldsymbol{\beta}}_i \neq 0$ with $i = 1, 2, 3, 4, 5$ for Model 1-3 and with $i = 1, 2, 3, 4$ for Model 4. This measures the ability of l_1-norm SVM selecting relevant features.
- **Noise**: the average of number of non-zero regression coefficients $\widehat{\boldsymbol{\beta}}_i(\lambda) \neq 0$ with $i \notin \{1, 2, 3, 4, 5\}$ for Model 1-3 and with $i \notin \{1, 2, 3, 4\}$ for Model 4. This measures the ability of l_1-norm SVM not selecting noise features.

Table 1 summarizes the simulation results for all four models. The numbers in the parentheses are the corresponding standard errors based on 200 replications. Overall, the l_1-norm SVM performs satisfactorily for classification with relatively low error rates in all the models. It is also successful in eliminating the majority of the irrelevant features. The performance increases with increased sample size. In terms of estimation accuracy, the L_2 error decreases as p decreases and n increases, which echoes the result in the Theorem 4.4.4. The measurement AAC is greater than 0.9 in most cases, implying the direction of $\widehat{\boldsymbol{\beta}}(\lambda)$ matching that of bayes rule.

It is worth noting that the earlier literature have already performed thorough numerical analysis to compare the performance of the L_1-norm SVM with L_2-norm SVM and logistic regression.For example, in a numerical experiment [63] observes that the performance of L_1-norm SVM and the L_2-norm SVM when there is no redundant features; however, the performance of L_2-norm SVM can be adversely affected by the presence of redundant features. From similar simulation study, [84] demonstrated that

Table 4.1: Simulation results for L_1-norm SVMs

Model	n	p	*Test error*	L_2 error	*AAC*	*Signal*	*Noise*
Model 1	100	1000	0.17(0.06)	0.53(0.14)	0.89(0.03)	4.84(0.41)	38.20(5.50)
	100	1500	0.19(0.05)	0.59(0.14)	0.89(0.03)	4.75(0.47)	40.27(5.41)
	200	1500	0.10(0.03)	0.27(0.07)	0.96(0.02)	5.00(0.07)	19.80(4.12)
	200	2000	0.10(0.02)	0.27(0.06)	0.96(0.02)	5.00(0.00)	23.61(4.80)
Model 2	100	1000	0.06(0.04)	0.34(0.12)	0.95(0.02)	4.88(0.35)	21.25(4.22)
	100	1500	0.07(0.04)	0.39(0.12)	0.95(0.02)	4.79(0.41)	28.80(4.61)
	200	1500	0.02(0.01)	0.21(0.07)	0.97(0.01)	4.99(0.10)	5.41(2.25)
	200	2000	0.02(0.02)	0.22(0.07)	0.97(0.01)	4.99(0.10)	6.88(2.50)
Model 3	100	1000	0.06(0.05)	0.36(0.14)	0.95(0.02)	4.8.(0.40)	19.93(3.87)
	100	1500	0.06(0.04)	0.37(0.13)	0.95(0.02)	4.83(0.40)	27.55(4.85)
	200	1500	0.02(0.02)	0.22(0.07)	0.97(0.02)	5.00(0.07)	5.18(2.19)
	200	2000	0.02(0.02)	0.20(0.08)	0.97(0.02)	5.00(0.07)	6.72(2.67)
Model 4	100	1000	0.16(0.04)	0.52(0.13)	0.94(0.03)	3.88(0.33)	12.87(3.65)
	100	1500	0.17(0.05)	0.55(0.14)	0.93(0.03)	3.81(0.42)	12.09(3.56)
	200	1500	0.13(0.03)	0.33(0.09)	0.97(0.01)	4.00(0.00)	11.12(3.53)
	200	2000	0.15(0.03)	0.43(0.07)	0.94(0.02)	4.00(0.00)	48.34(7.71)

L_1-norm SVM did perform remarkably better than logistic regression for finite sample sizes. While in most cases, the two methods are comparable on variable selection consistency in large sample cases. See similar observation in [69], [**?**], among others. Although L_1-norm SVM can outperform regular L_2-norm SVM when there are many redundant features, it shares the drawback of L_1 penalized least squares regression that it overpenalizes large coefficients and tends to have larger false positives (including more noise features) comparing with the non-convex penalized SVM, which will be investigated in Section 4.5.2.

4.5.2 Monte Carlo results for nonconvex penalized SVM

In this subsection, we consider the same four models as in Section 5.1. Instead of the l_1-norm SVM, we use it as the initial value for the nonconvex penalized SVM algorithm proposed in [60]. We consider two popular choices of nonconvex penalty functions: SCAD penalty (with $a = 3.7$) and MCP penalty (with $a = 3$). As suggested in [60], we used the recently developed high-dimensional BIC criterion to choose the tuning parameter for non-convex penalized SVMs. More specifically, the SVM-extended BIC is defined as

$$SVMIC_\gamma(T) = \sum_{i=1}^{n} 2\xi_i + \log(n)|T| + 2\gamma \binom{p}{|T|}, \qquad 0 \le \gamma \le 1,$$

where in practice we can set $\gamma = 0.5$ as suggested by [54] and choose the λ that minimizes the above $SVMIC_\gamma$ for non-convex penalized SVMs.

Tables 4.2 and 4.3 summarize the simulation results for SCAD and MCP penalty functions, respectively. We observe that the SCAD-penalized SVM and MCP-penalized MCP have similar performance, both demonstrating a clear advantage of selecting the relevant features and excluding irrelevant ones over l_1-norm SVM. The Noise size decreases dramatically to less than 3 as the sample size gets larger. The Signal size is almost 5 when $n = 200$ for Model 1-3 and 4 for Model 4, implying the success of selecting the exact true model. We also observe that non-convex penalized SVM has uniformly smaller L_2 error and larger AAC than L_1-norm SVM. This resonates with the observation in the literature that eliminating irrelevant features enhances classification performance. The Monte Carlo study confirms the effectiveness of the algorithm of [60] for feature selection for SVM in high dimension when using l_1-norm SVM as an initial value.

4.6 Conclusions

We investigate the statistical properties of l_1-norm SVM coefficients in ultra-high dimension. We proved that l_1-norm SVM coefficients achieve a near-oracle rate of estimation error. To deal with the nonsmoothness of the hinge loss function, we employ empirical processes techniques to derive the theory. Furthermore, we showed that under some

Table 4.2: Simulation results for SCAD penalized SVM

Model	n	p	Test error	L_2 error	AAC	Signal	Noise
Model 1	100	1000	0.10(0.05)	0.25(0.17)	0.95(0.04)	4.88(0.38)	4.92(5.82)
	100	1500	0.12(0.06)	0.35(0.20)	0.93(0.05)	4.84(0.53)	9.31(8.89)
	200	1500	0.08(0.03)	0.15(0.10)	0.98(0.03)	4.99(0.12)	0.48(0.51)
	200	2000	0.07(0.02)	0.10(0.05)	0.99(0.01)	5.00(0.00)	0.66(0.80)
Model 2	100	1000	0.04(0.05)	0.25(0.17)	0.95(0.05)	4.73(0.51)	1.47(1.38)
	100	1500	0.05(0.05)	0.28(0.18)	0.94(0.05)	4.64(0.55)	1.42(1.38)
	200	1500	0.03(0.03)	0.19(0.10)	0.96(0.03)	4.91(0.29)	2.77(3.53)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.07)	1.40(1.81)
Model 3	100	1000	0.05(0.04)	0.30(0.16)	0.94(0.04)	4.53(0.58)	0.58(0.84)
	100	1500	0.04(0.04)	0.24(0.15)	0.95(0.04)	4.75(0.46)	1.08(1.15)
	200	1500	0.02(0.01)	0.14(0.06)	0.98(0.01)	4.99(0.10)	1.30(1.53)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.00)	1.32(1.83)
Model 4	100	1000	0.15(0.05)	0.51(0.20)	0.94(0.04)	3.50(0.59)	7.54(5.20)
	100	1500	0.17(0.05)	0.61(0.18)	0.93(0.04)	3.57(0.71)	8.86(6.37)
	200	1500	0.12(0.03)	0.19(0.10)	0.99(0.01)	3.98(0.14)	3.19(2.45)
	200	2000	0.14(0.03)	0.39(0.19)	0.97(0.03)	3.69(0.51)	0.95(1.07)

general regularity conditions, the l_1-norm SVM provides an appropriate initial value for the recent algorithm developed by [60] for noncovex penalized SVM in high dimension. Combined with the theory in Section 4.3, we extended the applicability and validity of their result to the ultra-high dimension.

Table 4.3: Simulation results for MCP penalized SVM

Model	n	p	Test error	L_2 error	AAC	Signal	Noise
Model 1	100	1000	0.11(0.05)	0.28(0.17)	0.95(0.04)	4.87(0.42)	5.46(5.45)
	100	1500	0.13(0.07)	0.36(0.20)	0.93(0.05)	4.84(0.47)	9.00(8.49)
	200	1500	0.07(0.02)	0.11(0.07)	0.99(0.02)	4.99(0.10)	0.48(0.51)
	200	2000	0.07(0.02)	0.10(0.04)	0.99(0.01)	5.00(0.00)	0.83(0.83)
Model 2	100	1000	0.03(0.03)	0.20(0.12)	0.96(0.03)	4.84(0.38)	0.88(0.97)
	100	1500	0.11(0.10)	0.47(0.27)	0.89(0.08)	4.08(0.85)	3.56(2.65)
	200	1500	0.02(0.01)	0.14(0.05)	0.98(0.01)	5.00(0.00)	1.50(2.22)
	200	2000	0.02(0.01)	0.14(0.06)	0.98(0.02)	5.00(0.07)	1.38(1.80)
Model 3	100	1000	0.04(0.04)	0.26(0.15)	0.95(0.04)	4.67(0.54)	0.60(0.82)
	100	1500	0.04(0.04)	0.24(0.15)	0.95(0.04)	4.75(0.46)	1.01(1.07)
	200	1500	0.02(0.01)	0.14(0.06)	0.98(0.01)	5.00(0.07)	1.27(1.72)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.00)	1.47(2.04)
Model 4	100	1000	0.15(0.05)	0.50(0.20)	0.94(0.04)	3.66(0.52)	7.20(4.49)
	100	1500	0.17(0.05)	0.62(0.16)	0.92(0.04)	3.35(0.68)	4.96(3.58)
	200	1500	0.12(0.03)	0.20(0.12)	0.99(0.01)	3.98(0.12)	1.99(1.72)
	200	2000	0.13(0.03)	0.34(0.17)	0.97(0.02)	3.83(0.43)	0.86(0.80)

Chapter 5

Conclusion

5.1 Discussion

In this thesis we have built up a complete theoretical framework for non-convex penalties on a general class of models.

To choose a proper model automatically, researchers tend to utilized penalized regression model to analyze ultra-high dimensional data. Though LASSO penalized regression is computationally efficient and owns attractive theoretical properties, it requires very stringent conditions on design matrix to ensure variable selection consistency. Hence, we select non-convex penalty as a promising alternative to identify sparsity pattern for popular models. We investigate non-convex penalized quantile regression model and support vector machines to solve heterogeneity issue and fulfill classification tasks respectively. We establish the theory for proposed models under very relaxed conditions. In particularly, the theory claims that non-convex penalized quantile regression model and SVMs have the oracle property even under ultra high dimensional settings.

This framework consists of two indispensable components. First, the LASSO penalized model is able to create an estimator to sufficiently approach to the true parameter when an appropriate tuning parameter λ is chosen. This gives us a significant upper bound for the estimation error of LASSO penalized model. Then, we propose the LLA algorithm to solve the non-convex penalized model by transforming this non-convex problem to a series of convex optimization sub problems. Under rather weak conditions, the LLA algorithm is capable to find the oracle estimator in two iterations if

using the LASSO estimator as initial value. This theory not only points out the oracle estimator is a local minimum to the non-convex penalized model, but also validate the LLA algorithm to identify it among all those potential minima.

However, the LLA algorithm faces inevitable computational difficulties when it comes to ultra-high dimensions. To tackle this computing speed issue, we also propose a brand new QICD algorithm to solve non-convex penalized quantile regression model. By combining the MM technique and coordinate descent procedures, we demonstrate the considerable increase on computational speed and a local convergence theory for our QICD algorithm. Basically, the QICD algorithm owns fast iterative speed over the LLA algorithm at little cost of sacrificing its performance. The extension of this algorithm to other general models is quite promising and valuable to explore.

5.2 Future Works

This thesis provides an inspiring start for non-convex penalty research from theoretical and computational perspective simultaneously. We can extend our framework to censored quantile regression model, which is becoming rather attractive on biological area. As medical and gene sets data are commonly censored, censored quantile regression model becomes more and more popular on consistent variable selection and accurate prediction. Generally, censored case may require more stringent regular conditions for penalized model to achieve oracle property, but still maintain similar properties as usual quantile regression models. It is potentially to build up a modified theoretical framework based on our results.

Another vision is to create a unified algorithm for the application of non-convex penalty on a general class of models. QICD algorithm provides a novel mixture of coordinate descent algorithm and majorization-minimization idea to guarantee the fast speed and proper convergence at the same time. To apply this algorithm on other models, such as SVMs and grouping penalized models, we need to search for an efficient method to solve each sub problem in each iteration. Fast weight sorting for median in QICD offers a good insight into this exploration and versatile iterative algorithms are needed for each specific model.

References

[1] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

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

[3] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[4] EJ Candès, MB Wakin, and SP Boyd. Enhancing sparsity by reweighting l_1. Technical report, Tech. Rep., California Institute of Technol., 2007.

[5] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618, 2008.

[6] Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484, 2011.

[7] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[8] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[9] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[10] Jianqing Fan, Heng Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[11] Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

[12] Yongdai Kim and Sunghoon Kwon. Global optimality of nonconvex penalized estimators. *Biometrika*, 99:315–325, 2012.

[13] Cun-Hui Zhang, Tong Zhang, et al. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

[14] Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

[15] Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for cox?s proportional hazards model with np-dimensionality. *The Annals of Statistics*, 39(6):3092–3120, 2011.

[16] Changyi Park, Kwang-Rae Kim, Rangmi Myung, and Ja-Yong Koo. Oracle properties of scad-penalized support vector machine. *Journal of Statistical Planning and Inference*, 142(8):2257–2270, 2012.

[17] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1566, 2008.

[18] D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[19] David R Hunter and Runze Li. Variable selection using mm algorithms. *The Annals of statistics*, 33(4):1617, 2005.

[20] Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011.

[21] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

[22] Dingfeng Jiang and Jian Huang. Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and Computing*, 24(5):871–883, 2014.

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

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[25] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

[26] Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014.

[27] Alexandre Belloni, Victor Chernozhukov, et al. l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

[28] Lie Wang. The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

[29] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[30] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[31] AH Welsh. On m-processes and m-estimation. *The Annals of Statistics*, pages 337–361, 1989.

[32] ZD Bai and Y Wu. Limiting behavior of m-estimators of regression coefficients in high dimensional linear models i. scale dependent case. *Journal of Multivariate Analysis*, 51(2):211–239, 1994.

[33] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

[34] Youjuan Li and Ji Zhu. L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[35] Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, pages 1108–1126, 2008.

[36] Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817, 2009.

[37] Bo Kai, Runze Li, and Hui Zou. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, 39(1):305–332, 2011.

[38] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[39] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[40] T Tony Cai, Lie Wang, and Guangwu Xu. Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308, 2010.

[41] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[42] Todd E Scheetz, Kwang-Youn A Kim, Ruth E Swiderski, Alisdair R Philp, Terry A Braun, Kevin L Knudtson, Anne M Dorrance, Gerald F DiBona, Jian Huang, Thomas L Casavant, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

[43] Bo Peng and Lan Wang. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24(3):676–694, 2014.

[44] Roger Koenker and Beum J Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283, 1996.

[45] David R Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

[46] T.T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[47] Kenneth Lange. Elementary optimization. In *Optimization*, pages 1–17. Springer, 2004.

[48] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[49] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[50] W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[51] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics,*, 57:1413–1457, 2004.

[52] Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

[53] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

[54] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[55] Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, 13(1):1037–1057, 2012.

[56] Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41(5):2505, 2013.

[57] Eun Ryung Lee, Hohsuk Noh, and Byeong U Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

[58] MS Bazaraa, HD Sherali, and CM Shetty. *Convex Optimization*. Wiley-Interscience, 2006.

[59] Elizabeth D Schifano, Robert L Strawderman, Martin T Wells, et al. Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics*, 4:1258–1299, 2010.

[60] Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):53–76, 2014.

[61] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[62] Vladimir N Vapnik. The nature of statistical learning theory. 1995.

[63] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems*, 16(1):49–56, 2004.

[64] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[65] Minghu Song, Curt M Breneman, Jinbo Bi, Nagamani Sukumar, Kristin P Bennett, Steven Cramer, and Nihal Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, 42(6):1347–1357, 2002.

[66] B Tarigan and SA van de Geer. Adaptivity of support vector machines with l1. Technical report, Penalty, Tech. Rep. of Math. Inst., Univ. of Leiden, Leiden, 2004, no. MI 2004-14., 2004.

[67] Bernadetta Tarigan, Sara A Van De Geer, et al. Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

[68] Lifeng Wang and Xiaotong Shen. On l_1-norm multiclass support vector machines. *Journal of the American Statistical Association*, 102(478):583–594, 2007.

[69] Hui Zou. An improved 1-norm svm for simultaneous classification and variable selection. In *AISTATS*, volume 2, pages 675–681. Citeseer, 2007.

[70] Marten Wegkamp, Ming Yuan, et al. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.

[71] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[72] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[73] Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *The Journal of Machine Learning Research*, 9:1343–1368, 2008.

[74] Yi Lin, Yoonkyung Lee, and Grace Wahba. Support vector machines for classification in nonstandard situations. *Machine learning*, 46(1-3):191–202, 2002.

[75] Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

[76] Roger Koenker. *Quantile regression*. Number 38. Cambridge University Press, 2005.

[77] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36:1567–1594, 2008.

[78] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

[79] Gerda Claeskens, Christophe Croux, and Johan Van Kerckhoven. An information criterion for variable selection in support vector machines. *The Journal of Machine Learning Research*, 9:541–558, 2008.

[80] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[81] T Tony Cai, Lie Wang, and Guangwu Xu. New bounds for restricted isometry constants. *Information Theory, IEEE Transactions on*, 56(9):4388–4394, 2010.

[82] Tony Cai, Weidong Liu, and Xi Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[83] R Dennis Cook, Bing Li, and Francesca Chiaromonte. Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584, 2007.

[84] Guilherme V Rocha, Xing Wang, and Bin Yu. Asymptotic distribution and sparsistency for l1-penalized parametric m-estimators with applications to linear svm and logistic regression. *arXiv preprint arXiv:0908.1940*, 2009.

[85] CA Rogers. Covering a sphere with spheres. *Mathematika*, 10(02):157–164, 1963.

[86] Jean Bourgain and Vitaly D Milman. New volume ratio properties for convex symmetric bodies in $\mathbb{R}^n$. *Inventiones Mathematicae*, 88(2):319–340, 1987.

Appendix A

Useful Definitions and Notations

Let $\mathbb{R}^m$ denote the m-dimensional real space. For any $h : \mathbb{R}^m \mapsto \mathbb{R} \bigcup \infty$, we denote by dom h the effective domain of h, i.e.,

$$dom\ h \quad = \quad \{\mathbf{x} \in \mathbb{R}^m | h(\mathbf{x}) < \infty\}$$

For any $\mathbf{x} \in dom\ h$ and any $d \in \mathbb{R}^m$, we denote the *(lower) directional derivative* of h at $\mathbf{x}$ in the direction d by

$$h'(\mathbf{x}; d) \quad = \quad \liminf_{\lambda \downarrow 0} [h(\mathbf{x} + \lambda d) - h(\mathbf{x})]/\lambda.$$

We say that h is *quasiconvex* if

$$h(\mathbf{x} + \lambda d) \quad \leq \quad \max\{h(\mathbf{x}), h(\mathbf{x} + d)\},$$

and that h is *hemivariate* if h is not constant on any line segment belonging to dom h. A function f is said to be *lower semicontinuous (lsc)* on dom f if

$$\liminf_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) \quad \geq \quad f(\mathbf{x}_0), \quad \text{for each } \mathbf{x}_0 \in \text{dom } f.$$

We define $\mathbf{z}$ as a *stationary point* of f if $\mathbf{z} \in$ dom f and

$$f'(\mathbf{z}; d) \quad \geq \quad 0, \quad \forall\ d.$$

We say that $\mathbf{z}$ is a *coordinate-wise minimum point* of f if $\mathbf{z} \in$ dom f and

$$f(\mathbf{z} + (0, \ldots, d_k, \ldots, 0)) \quad \geq \quad f(\mathbf{z}), \quad \forall\ d_k \in \mathbb{R}.$$

for all $k = 1, \ldots, N$. Here, we denote by $(0, \ldots, d_k, \ldots, 0)$ the vector in $\mathbb{R}^N$ whose kth coordinate is d_k and whose other coordinates are zero. We say that f is *regular* at $\mathbf{z} \in \text{dom } f$ if

$$\begin{aligned} f'(\mathbf{z}; d) \geq 0, \quad \forall d = (d_1, \ldots, d_N), \\ if\ f'(\mathbf{z}; (0, \ldots, d_k, \ldots, 0)) \geq 0, k = 1, \ldots, N. \end{aligned} \tag{A.1}$$

We consider a generalized penalized loss function of the form

$$f(x_1, \ldots, x_N) \quad = \quad f_0(x_1, \ldots, x_N) + \sum_{k=1}^{N} f_k(x_k) \tag{A.2}$$

where $f_0 : \mathbb{R}^N \mapsto \mathbb{R} \bigcup \infty$ and $f_k : \mathbb{R} \mapsto \mathbb{R} \bigcup \infty$, $k = 1, 2, \ldots, N$, with We assume that f is proper, i.e., $f \not\equiv \infty$. We adopt the following assumptions on $f, f_0, f_1, \ldots, f_N$.

(D1) f_0 is continuous on dom f_0.

(D2) For each $k \in \{1, \ldots, N\}$ and $(x_j)_{j \neq k}$, the function $x_k \mapsto f(x_1, \ldots, x_N)$ is *quasiconvex* and *hemivariate*.

(D3) $f_0, f_1, \ldots, f_N$ is *lower semicontinuous*.

(D4) dom f_0 is open and f_0 tends to ∞ at every boundary point of dom f_0.

(D5) dom $f_0 = Y_1 \times \cdots \times Y_N$, for some $Y_k \subseteq \mathbb{R}$, $k = 1, \ldots, N$.

In the following, we state a useful result of [48].

Proposition A.0.1. *[48] Consider an objective function of the form (A.2). Assume that $f, f_0, f_1, \ldots, f_N$ satisfy conditions (D1)-(D3) and that f_0 satisfies either condition (D4) or (D5). Let $\mathbf{x}^r = (x_1^r, \ldots, x_N^r)_{r=0,1,\ldots}$ be a sequence generated by the coordinate descent algorithm for minimizing (A.2) using the cyclic rule such as the one in (3.6). Then, either $\{f(\mathbf{x}^r)\} \downarrow -\infty$, or every cluster point $\mathbf{z}$ of $\{\mathbf{x}^r\}$ is a coordinatewise minimum point of f.*

Appendix B

Technical Proofs

Proof of Theorem 2.3.1. It follows by the result of Theorem 7 in [26]. □

Proof of Lemma 2.4.1. It follows by the result of Lemma 1 in [28]. □

Proof of Theorem 2.4.2. It follows by the result of Theorem 1 in [28]. □

Proof of Corollary 2.4.3. It follows by combining the result of Theorem 1 in [28] and Theorem 7 in [26] . □

Proof of Proposition 3.4.1. (1) It is easy to show that $\phi_{\beta_j^{(k-1)}}(\beta_j^{(k-1)}) = p_\lambda(|\beta_j^{(k-1)}|)$. If $|\beta| > |\beta_j^{(k-1)}|$, by the mean value theorem, we have

$$p_\lambda(|\beta|) - p_\lambda(|\beta_j^{(k-1)}|) \;=\; p'_\lambda(\xi+)(|\beta| - |\beta_j^{(k-1)}|)$$

for some $\xi \in [|\beta_j^{(k-1)}|, |\beta|]$. Since $p_\lambda(\cdot)$ is concave, we have $p_\lambda{}'(|\beta_j^{(k-1)}|+) \geq p_\lambda{}'(\xi+)$. Hence,

$$\begin{aligned} p_\lambda(|\beta|) &= p_\lambda(|\beta_j^{(k-1)}|) + p'_\lambda(\xi+)(|\beta| - |\beta_j^{(k-1)}|) \\ &\leq p_\lambda(|\beta_j^{(k-1)}|) + p'_\lambda(|\beta_j^{(k-1)}|+)(|\beta| - |\beta_j^{(k-1)}|) \\ &= \phi_{\beta_j^{(k-1)}}(\beta). \end{aligned}$$

Similarly, we can show that if $|\beta| < |\beta_j^{(k-1)}|$, then $p_\lambda(|\beta|) \leq \phi_{\beta_j^{(k-1)}}(\beta)$. Therefore, $\phi_{\beta_j^{(k-1)}}(\beta)$ *majorizes* $p_\lambda(\beta)$ at the points $\pm|\beta_j^{(k-1)}|$.

(2) It follows from (1) that

$$\begin{aligned} Q(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i-\mathbf{x}_i^\intercal\boldsymbol{\beta})+\sum_{j=1}^{p}p_\lambda(|\beta_j|) \\ &\leq n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i-\mathbf{x}_i^\intercal\boldsymbol{\beta})+\sum_{j=1}^{p}\phi_{\beta_j^{(k-1)}}(|\beta_j|) \\ &= Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}) \end{aligned}$$

Hence, $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ majorizes $Q(\boldsymbol{\beta})$ at the points $\pm|\boldsymbol{\beta}^{(k-1)}|$.
(3) We have

$$Q(\boldsymbol{\beta}^{(k)}) \leq Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)}) \leq Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k-1)}) = Q(\boldsymbol{\beta}^{(k-1)}),$$

where the first inequality and the last equality follow from the property of the majorization function in (3.1), while the second inequality follows from (3.5). This proves the descent property. □

Proof of Lemma 3.4.2. The result follows directly from Proposition A.0.1. It is easy to check that $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$ has the form (A.2) with components $f_0 = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})$ and $f_j = \phi_{\beta_j^{(k-1)}}(|\beta_j|)$ for $i \geq 1$ and $j \geq 1$, which satisfy conditions (D1)-(D5). Our algorithm implies that $\boldsymbol{\beta}^{(k)}$ is a cluster point of $\boldsymbol{\beta}_j^{(k)(r)}$. In addition, $Q(\boldsymbol{\beta}^{(k)}) \leq Q(\boldsymbol{\beta}^{(0)}) < +\infty$, and $Q_{\boldsymbol{\beta}^{(k-1)}}(\beta) \geq 0$, $\{Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}_j^{(k)(r)})\} \not\to -\infty$ as $r \to \infty$. Hence, by Proposition A.0.1, $\boldsymbol{\beta}^{(k)}$ is a *coordinatewise minimum point* of $Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta})$. □

Proof of Lemma 3.4.3. We have

$$\begin{aligned} \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)}) &\geq \lim_{k\to+\infty} Q(\boldsymbol{\beta}^{(k)}) \qquad \text{(due to } \textit{majorization}\text{)}. \\ &= \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k)}) \\ &\geq \lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k)}}(\boldsymbol{\beta}^{(k+1)}), \end{aligned}$$

where the last inequality follows because $\boldsymbol{\beta}^{(k)}$ is a coordinate minimum point. Hence $Q(\boldsymbol{\beta}^*) = A$. □

To prove Theorem 3.4.4, we state below a useful result from Bazaraa, Sherali and Shetty (2006, Theorem 3.3.10).

Lemma B.0.1. *(Bazaraa, Sherali and Shetty, 2006) Given a function f:* $\mathbb{R}^n \mapsto \mathbb{R}$*, let* $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda) = f(\bar{\mathbf{x}} + \lambda\mathbf{d})$*, where* $\bar{\mathbf{x}}$ *is some point in* $\mathbb{R}^n$ *and* $\mathbf{d} \in \mathbb{R}^n$ *is a nonezero direction. Then f is (strictly) convex if and only if* $F_{(\bar{\mathbf{x}};\mathbf{d})}(\cdot)$ *is a (strictly) convex function of* λ *for all* $\bar{\mathbf{x}}$ *and* $\mathbf{d} \neq \mathbf{0}$ *in* $\mathbb{R}^n$*.*

Proof. We include the proof here for completeness, Given any $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^n$, we write $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$ as $F(\lambda)$ for notational simplicity. If f is convex, then for any λ_1 and λ_2 in $\mathbb{R}$ and for any $0 \leq \alpha \leq 1$, we have

$$\begin{aligned} F(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= f(\alpha[\bar{\mathbf{x}} + \lambda_1\mathbf{d}] + (1-\alpha)[\bar{\mathbf{x}} + \lambda_2\mathbf{d}]) \\ &\leq \alpha f(\bar{\mathbf{x}} + \lambda_1\mathbf{d}) + (1-\alpha)f(\bar{\mathbf{x}} + \lambda_2\mathbf{d}) \\ &= \alpha F(\lambda_1) + (1-\alpha)F(\lambda_2) \end{aligned}$$

Hence, F is convex. Conversely, suppose that $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$, $\lambda \in \mathbb{R}$, convex for all $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^n$. Then, for any $\mathbf{x}_1$ and $\mathbf{x}_2$ and $0 \leq \lambda \leq 1$, we have

$$\begin{aligned} \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) &= \lambda f[\mathbf{x}_1 + 0(\mathbf{x}_2 - \mathbf{x}_1)] + (1-\lambda)f[\mathbf{x}_1 + 1(\mathbf{x}_2 - \mathbf{x}_1)] \\ &= \lambda F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(0) + (1-\lambda)F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1) \\ &\geq F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1-\lambda) \\ &= f[\mathbf{x}_1 + (1-\lambda)(\mathbf{x}_2 - \mathbf{x}_1)] \\ &= f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]. \end{aligned}$$

So f is convex. The argument for the strictly convex case is similar. □

Proof of Theorem 3.4.4. Let $\{r_m\}$ be a subsequence of $\{k_m\}$ such that $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m)} = \boldsymbol{\beta}^*$ and $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m-1)} = \boldsymbol{\beta}^{**}$. Denote $\lim_{k\to+\infty} Q_{\boldsymbol{\beta}^{(k-1)}}(\boldsymbol{\beta}^{(k)})$ by A. By Lemma 3.4.3, we have

$$\begin{aligned} Q(\boldsymbol{\beta}^*) = Q(\boldsymbol{\beta}^{**}) &= \lim_{m\to+\infty} Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta}^{(r_m)}) \\ &= Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*) = A. \end{aligned}$$

Note that $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$. By Lemma B.0.1, $R(\lambda) = Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta} + \lambda\mathbf{d})$ is convex in λ for all $\boldsymbol{\beta}$ and $\mathbf{d} \neq \mathbf{0}$. Moreover, by Lemma 3.4.2, $\boldsymbol{\beta}^{(r_m)}$ is the coordinate-wise minimum point of $Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta})$ for $m = 1, 2, \ldots$. Since $\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m)} = \boldsymbol{\beta}^*$ and

$\lim_{m\to+\infty} \boldsymbol{\beta}^{(r_m-1)} = \boldsymbol{\beta}^{**}$, $\boldsymbol{\beta}^*$ is a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ by the continuity of $Q_{\boldsymbol{\beta}^{(r_m-1)}}(\boldsymbol{\beta}^{(r_m)})$ and $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Since $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$ is *regular* at $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}^*$ is a stationary point as well by (A.1). Hence, we have

$$Q'_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*; \mathbf{d}) \quad \geq \quad 0, \quad \forall\ \mathbf{d},$$

which is equivalent to

$$R'(\lambda; \boldsymbol{\beta}^*, \mathbf{d})|_{\lambda=0} \quad \geq \quad 0, \quad \forall\ \mathbf{d}. \tag{B.1}$$

If $\boldsymbol{\beta}^{**}$ is not a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$, then $\forall\ a > 0$, $\exists\ \mathbf{d}^{(1)} = (0, \ldots, d_i, \ldots, 0)$, with $|d_i| < a$, such that

$$Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)}) < Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**}) = Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*) \tag{B.2}$$

Let $\mathbf{d}^{(2)} = \boldsymbol{\beta}^{**} + \mathbf{d}^{(1)} - \boldsymbol{\beta}^*$. Then we have, $\forall\ \lambda \in (0, 1)$,

$$\begin{aligned} Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^* + \lambda\mathbf{d}^{(2)}) &= Q_{\boldsymbol{\beta}^{**}}((1-\lambda)\boldsymbol{\beta}^* + \lambda(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)})) \\ &\leq (1-\lambda)Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*) + \lambda Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**} + \mathbf{d}^{(1)}) \\ &< Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*), \end{aligned}$$

where the last inequality follows from (B.2). Note that although $R(\lambda; \boldsymbol{\beta}^*)$ is not differentiable everywhere, it is non-differentiable at only a finite number of points; hence, there exits a constant λ, $R(\lambda; \boldsymbol{\beta}^*)$ is differentiable in $(0, \lambda)$. Then, by the mean value theorem there exits $\lambda_1 \in (0, \lambda)$ such that

$$\begin{aligned} R'(\lambda_1; \boldsymbol{\beta}^*, \mathbf{d}^{(2)}) &= \frac{R(\lambda; \boldsymbol{\beta}^*, \mathbf{d}^{(2)}) - R(0; \boldsymbol{\beta}^*, \mathbf{d}^{(2)})}{\lambda} \\ &= \frac{Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^* + \lambda\mathbf{d}^{(2)}) - Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^*)}{\lambda} \\ &< 0. \end{aligned}$$

However, $R(\cdot; \boldsymbol{\beta}^*, \mathbf{d}^{(2)})$ is convex. Hence, $R'(0; \boldsymbol{\beta}^*, \mathbf{d}^{(2)}) \leq R'(\lambda_1; \boldsymbol{\beta}^*, \mathbf{d}^{(2)}) < 0$. This contradicts (B.1). Therefore, $\boldsymbol{\beta}^{**}$ is a coordinatewise minimum point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Similarly, $\boldsymbol{\beta}^{**}$ is a stationary point of $Q_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta})$. Furthermore, since $p'_\lambda(|\theta|+) = p'_\lambda(|\theta|-)$ on $(0, \infty)$, we have

$$Q'(\boldsymbol{\beta}^{**}; \mathbf{d}) = Q'_{\boldsymbol{\beta}^{**}}(\boldsymbol{\beta}^{**}; \mathbf{d}) \geq 0, \qquad \forall\ \mathbf{d}.$$

Hence, $\boldsymbol{\beta}^{**}$ is a *stationary point* of $Q(\boldsymbol{\beta})$. Since $\boldsymbol{\beta}^{**}$ is an arbitrary cluster point of $\{\boldsymbol{\beta}^{(k-1)}\}$, we conclude that every cluster point of the sequence generated by the QICD algorithm is a stationary point of $Q(\boldsymbol{\beta})$. □

Proof of Theorem 4.3.1. It follows by the result of Theorem 2 in [60]. □

Proof of Theorem 4.3.2. It follows by the result of Theorem 3 in [60]. □

Proof of Lemma 4.4.1. By the union bound, we have

$$P\big(c\sqrt{2A(\alpha)\log p/n} \le c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty\big) \le \sum_{j=0}^{p} P\Big(\sqrt{2A(\alpha)\log p/n} \le \frac{1}{n}\big|\sum_{i=1}^{n} I(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^* \ge 0)Y_iX_{ij}\big|\Big)$$

Notice that we have $S(\boldsymbol{\beta}^*) = 0$ because of minimizer $\boldsymbol{\beta}^*$ and the definition of gradient vector. Then, for each i and j, $E(Y_iX_{ij}I(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^* \ge 0)) = 0$, by Hoeffding's inequality,

$$\begin{aligned} & P\Big(\sqrt{2A(\alpha)\log p/n} \le n^{-1}\big|\sum_{i=1}^{n} I(1 - Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^* \ge 0)Y_iX_{ij}\big|\Big) \\ \le \;& 2\exp(-\frac{4A(\alpha)n\log p}{4nM^2}) = 2p^{-\frac{A(\alpha)}{M^2}}. \end{aligned}$$

Thus,

$$P\big(c\sqrt{2A(\alpha)\log p/n} \le c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty\big) \le (p+1)\cdot 2p^{-\frac{A(\alpha)}{M^2}} \le \alpha.$$

□

Lemma 4.4.2 *For $\lambda \ge c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty$ and $\bar{C} = \frac{c-1}{c+1}$, we have*

$$\mathbf{h} \in \Delta_{\bar{C}},$$

where

$$\Delta_{\bar{C}} = \big\{\boldsymbol{\gamma} \in \mathbf{R}^{p+1} : ||\boldsymbol{\gamma}_{T_+}||_1 \ge \bar{C}||\boldsymbol{\gamma}_{T_+^c}||_1, \text{where } T_+ = T\cup\{0\},\ T \subset \{1,2,\ldots,p\} \text{ and } |T| \le q\big\},$$

with T_+^c denoting the complement of T_+, and $\boldsymbol{\gamma}_{T_+}$ denoting the $(p+1)$-dimensional vector that has the same coordinates as $\boldsymbol{\gamma}$ on T_+ and zero coordinates on T_+^c.

Proof of Lemma 4.4.2. Since $\widehat{\beta}$ minimizes $l_n(\boldsymbol{\beta})$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\widehat{\boldsymbol{\beta}})_+ + \lambda||\widehat{\boldsymbol{\beta}}_-||_1 \leq \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ + \lambda||\boldsymbol{\beta}^*_-||_1,$$
$$\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ \leq \lambda||\boldsymbol{\beta}^*_-||_1 - \lambda||\widehat{\boldsymbol{\beta}}_-||_1.$$

Let T denote the set of significant coefficients, *i.e.*, non-zero coefficients, we have

$$\begin{aligned} ||\boldsymbol{\beta}^*_-||_1 - ||\widehat{\boldsymbol{\beta}}_-||_1 &\leq ||\boldsymbol{\beta}^*_{T_+}||_1 - |\widehat{\boldsymbol{\beta}}_-||_1 \\ &\leq ||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1. \end{aligned}$$

This implies

$$\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ \leq \lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1).$$

Since the sub-differential of $l_n(\boldsymbol{\beta})$ at the point of $\boldsymbol{\beta}^*$ is $\widehat{S}(\boldsymbol{\beta}^*)$ and recall the assumption $\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty$, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ \\ \geq\ & \widehat{S}^\intercal(\boldsymbol{\beta}^*)\mathbf{h} \\ \geq\ & -||\mathbf{h}||_1 \cdot ||\widehat{S}(\boldsymbol{\beta}^*)||_\infty \\ \geq\ & -\frac{\lambda}{c}(||\mathbf{h}_{T_+}||_1 + ||\mathbf{h}_{T_+^c}||_1). \end{aligned}$$

Hence, we have

$$\begin{aligned} \lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1) &\geq -\frac{\lambda}{c}(||\mathbf{h}_{T_+}||_1 + ||\mathbf{h}_{T_+^c}||_1), \\ ||\mathbf{h}_{T_+}||_1 &\geq \bar{C}||\mathbf{h}_{T_+^c}||_1, \end{aligned}$$

where $\bar{C} = \frac{c-1}{c+1}$. We have thus proved that $\mathbf{h} \in \Delta_{\bar{C}}$. □

Proof of Lemma 4.4.3. We have

$$\left|(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+ - (1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+\right| \leq |\mathbf{X}_i^\intercal\mathbf{h}|.$$

By Hoeffding's inequality, we have

$$P\left(B(\mathbf{h}) \geq \frac{t}{\sqrt{n}}|\mathcal{X}\right) \leq 2\exp\Big(-\frac{2nt^2}{4||\mathcal{X}\mathbf{h}||_2^2}\Big).$$

By the definition of λ_{max},

$$P\left(B(\mathbf{h}) \geq \frac{t}{\sqrt{n}}|\mathcal{X}\right) \leq 2\exp\Big(-\frac{t^2}{2\lambda_{max}||\mathbf{h}||_2^2}\Big).$$

Hence,

$$P\left(B(\mathbf{h}) \geq \frac{t}{\sqrt{n}}\right) \leq 2\exp\Big(-\frac{t^2}{2M_1||\mathbf{h}||_2^2}\Big).$$

Let $t = C\sqrt{2q\log p}||\mathbf{h}||_2$, then

$$\begin{aligned} & P\big(B(\mathbf{h}) \geq C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}||_2\big) \\ \leq\ & 2\exp\big(-\frac{C^2 2q\log p||\mathbf{h}||_2^2}{2M_1||\mathbf{h}||_2^2}\big) \\ \leq\ & 2p^{-qC^2/M_1} \\ \leq\ & 2p^{-(q+1)C^2/2M_1} \end{aligned}$$

for all $C > 0$. Next we will derive an upper bound for $\sup_{||\mathbf{h}||_0=q+1,||\mathbf{h}||_2=1} B(\mathbf{h})$. Consider the ϵ-Net of the set $\{\mathbf{h} \in \mathbb{R}^{p+1}, ||\mathbf{h}||_0 = q+1, ||\mathbf{h}||_2 = 1\}$. We know that the covering number of $\{\mathbf{h} \in \mathbb{R}^{q+1}, ||\mathbf{h}||_2 = 1\}$ by balls of radius ϵ is at most $(C_1/\epsilon)^{q+1}$ for $\epsilon < 1$, see for example [85] and [86]. Therefore, the covering number of $\{\mathbf{h} \in \mathbb{R}^{p+1}, ||\mathbf{h}||_0 = q+1, ||\mathbf{h}||_2 = 1\}$ by ϵ balls is at most $(C_1 p/\epsilon)^{q+1}$ for $\epsilon < 1$. Suppose N is such a ϵ-Net of $\{\mathbf{h} \in \mathbb{R}^{p+1}, ||\mathbf{h}||_0 = q+1, ||\mathbf{h}||_2 = 1\}$. By union bound,

$$P\left(\sup_{\mathbf{h}\in N} B(\mathbf{h}) \geq C\sqrt{\frac{2q\log p}{n}}\right) \leq 2\big(\frac{C_1}{\epsilon}\big)^{q+1} p^{q+1} p^{-qC^2/M_1}.$$

And we have

$$\begin{aligned} & \sup_{\mathbf{h}_1,\mathbf{h}_2\in\mathbb{R}^{p+1}, ||\mathbf{h}_1-\mathbf{h}_2||_0\leq 2q+2, ||\mathbf{h}_1-\mathbf{h}_2||_2\leq\epsilon} |B(\mathbf{h}_1) - B(\mathbf{h}_2)| \\ \leq\ & \frac{2}{n}||\mathcal{X}(\mathbf{h}_1 - \mathbf{h}_2)||_1 \\ \leq\ & \frac{2}{n}\max_{j=0,1,2,\ldots,p}(||\mathcal{X}_{\cdot j}||_1)||\mathbf{h}_1 - \mathbf{h}_2||_1 \\ \leq\ & \frac{2}{n}nM\sqrt{2q+2}\epsilon \\ =\ & 2\sqrt{2q+2}M\epsilon, \end{aligned}$$

where $\mathcal{X}_{\cdot j}$ denotes the jth column vector of $\mathcal{X}$. Therefore,

$$\sup_{||\mathbf{h}||_0=q+1,||\mathbf{h}||_2=1} B(\mathbf{h}) \leq \sup_{\mathbf{h}\in N} B(\mathbf{h}) + 2\sqrt{2q+2}M\epsilon$$

Let $\epsilon = \sqrt{2q\log p}\frac{1}{2\sqrt{(2q+2)n}M} = \frac{\sqrt{\log p}}{2M\sqrt{n}}\sqrt{\frac{q}{q+1}}$, we have

$$\begin{aligned}
& P\left(\sup_{||\mathbf{h}||_0=q+1,||\mathbf{h}||_2=1} B(\mathbf{h}) \geq C\sqrt{\frac{2q\log p}{n}}\right) \\
\leq\ & P\left(\sup_{\mathbf{h}\in N} B(\mathbf{h}) \geq (C-1)\sqrt{\frac{2q\log p}{n}}\right) \\
\leq\ & 2\big(\frac{C_1}{\epsilon}\big)^{q+1} p^{q+1} p^{-(q+1)(C-1)^2/2M_1} \\
\leq\ & 2\big(\frac{C_1 p\sqrt{nq}}{p^{(C-1)^2/2M_1}}\big)^{q+1}
\end{aligned}$$

Since $p > n$ and $p > C_1\sqrt{q}$,

$$P\Big(\sup_{||\mathbf{h}||_0=q+1,||\mathbf{h}||_2=1} B(\mathbf{h}) \geq (1+2C_2\sqrt{2M_1})\sqrt{\frac{2q\log p}{n}}\Big) \leq 2p^{-4q(C_2^2-1)}.$$

□

Lemma B.0.2. *For any $x \in \mathbb{R}^n$,*

$$||x||_2 - \frac{||x||_1}{\sqrt{n}} \leq \frac{\sqrt{n}}{4}\big(\max_{1\leq i\leq n}|x_i| - \min_{1\leq i\leq n}|x_i|\big).$$

Proof of Lemma B.0.2. This proof is given in [81]. We include it here for completeness and easy reference. It is obvious that the result holds when $|x_1| = |x_2| = \ldots = |x_n|$. Without loss of generality, we now assume that $x_1 \geq x_2 \geq \ldots \geq x_n \geq 0$ and not all x_i are equal. Let

$$f(x) = ||x||_2 - \frac{||x||_1}{\sqrt{n}}.$$

Note that for any $i \in \{2, 3, \ldots, n-1\}$

$$\frac{\partial f}{\partial x_i} = \frac{x_i}{||x||_2} - \frac{1}{\sqrt{n}}.$$

This implies that when $x_i \leq \frac{||x||_2}{\sqrt{n}}$, $f(x)$ is decreasing w.r.t x_i; otherwise $f(x)$ is increasing w.r.t x_i. Hence, if we fix x_1 and x_n, when $f(x)$ achieves its maximum, x must be of the

form that $x_1 = x_2 = \ldots = x_k$ and $x_{k+1} = \ldots = x_n$ for some $1 \leq k \leq n$. Now,

$$f(x) = \sqrt{k(x_1^2 - x_n^2) + nx_n^2} - \frac{k}{\sqrt{n}}(x_1 - x_n) - \sqrt{n}x_n.$$

Treat this as a function of k for $k \in (0, n)$.

$$g(x) = \sqrt{k(x_1^2 - x_n^2) + nx_n^2} - \frac{k}{\sqrt{n}}(x_1 - x_n) - \sqrt{n}x_n.$$

By taking the derivatives, it is easy to see that

$$\begin{aligned} g(k) &\leq g\left(n\frac{(\frac{x_1+x_n}{2})^2 - x_n^2}{x_1^2 - x_n^2}\right) \\ &= \sqrt{n}(x_1 - x_n)\left(\frac{1}{2} - \frac{x_1 + 3x_n}{4(x_1 + x_n)}\right). \end{aligned}$$

Since $\frac{1}{2} - \frac{x_1+3x_n}{4(x_1+x_n)} \geq \frac{1}{4}$, we have

$$||x||_2 \leq \frac{||x||_1}{\sqrt{n}} + \frac{\sqrt{n}}{4}(x_1 - x_n).$$

We can also see that the above inequality becomes an equality if and only if $x_{k+1} = \ldots = x_n = 0$ and $k = \frac{n}{4}$. □

Proof of Theorem 4.4.4. Suppose $\mathbf{h} \in \Delta_{\bar{C}}$, then assume $|h_0| \geq |h_1| \geq \ldots \geq |h_p|$. Create a trivial partition of $\{0, 1, 2, \ldots, p\}$ as

$$S_0 = \{0, 1, 2, \ldots, q\}, S_1 = \{q + 1, q + 2 \ldots, 2q\}, \ldots,$$

where each set has the cardinality to be q+1. Easy to prove that $||\mathbf{h}_{S_0}||_1 \geq ||\mathbf{h}_{T_+}||_1 \geq \bar{C}||\mathbf{h}_{T_+^C}||_1 \geq \bar{C}||\mathbf{h}_{S_0^C}||_1$. According to the Lemma B.0.2, we have

$$\begin{aligned} \sum_{i\geq 1} ||\mathbf{h}_{S_i}||_2 &\leq \sum_{i\geq 1} \frac{||\mathbf{h}_{S_i}||_1}{\sqrt{q+1}} + \frac{\sqrt{q+1}}{4}|h_q| \\ &\leq \frac{||\mathbf{h}_{S_0^C}||_1}{\sqrt{q+1}} + \frac{||\mathbf{h}_{S_0}||_1}{4\sqrt{q+1}} \\ &\leq (\frac{1}{\sqrt{q+1}\bar{C}} + \frac{1}{4\sqrt{q+1}})||\mathbf{h}_{S_0}||_1 \\ &\leq (\frac{1}{4} + \frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2, \end{aligned} \tag{B.3}$$

and

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+-\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ \\
= \; & \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h}_{S_0})_+-\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+ \\
& +\sum_{j\geq 1}\frac{1}{n}(\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j}\mathbf{h}_{S_k})_+ \\
& -\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j-1}\mathbf{h}_{S_k})_+).
\end{aligned}$$

By the Lemma 4.4.3, with probability at least $1-2p^{-4q(C_2^2-1)}$,

$$\begin{aligned}
& \frac{1}{n}(\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j}\mathbf{h}_{S_k})_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j-1}\mathbf{h}_{S_k})_+) \\
\geq \; & \frac{1}{n}E(\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j}\mathbf{h}_{S_k})_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\sum_{k=1}^{j-1}\mathbf{h}_{S_k})_+) \\
& -C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_j}||_2.
\end{aligned}$$

The above inequality holds for each j. Hence with probability at least $1-2p^{-4q(C_2^2-1)+1}$,

$$\begin{aligned}
& \frac{1}{n}\big(\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+\big) \\
\geq \; & M(\mathbf{h})-C\sqrt{\frac{2q\log p}{n}}\sum_{j\geq 0}||\mathbf{h}_{S_j}||_2
\end{aligned}$$

where $M(\mathbf{h})=\frac{1}{n}E(\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*+Y_i\mathbf{X}_i^\intercal\mathbf{h})_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^\intercal\boldsymbol{\beta}^*)_+)$.

By the definition of $\mathbf{h}$ and (B.3), we have

$$\begin{aligned}
M(\mathbf{h}) & \leq \lambda(||\mathbf{h}_{T_+}||_1-||\mathbf{h}_{T_+^C}||_1)+(\frac{1}{4}+\frac{1}{\overline{C}})C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_0}||_2+C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_0}||_2 \\
& \leq \lambda\sqrt{q}||\mathbf{h}_{S_0}||_2+C\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\overline{C}})||\mathbf{h}_{S_0}||_2. \qquad \text{(B.4)}
\end{aligned}$$

By Taylor series expansion of $L(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$, we have

$$M(\mathbf{h})=\frac{1}{2}\mathbf{h}^\intercal H(\boldsymbol{\beta}^*)\mathbf{h}+o_p(||\mathbf{h}||_2^2).$$

By the condition (A5), we have

$$\mathbf{h}^{\intercal} H(\boldsymbol{\beta}^*)\mathbf{h} \geq M_2||\mathbf{h}||_2^2.$$

Substitute it in (B.4), we have

$$\frac{1}{2}M_2||\mathbf{h}||_2^2 + o_p(||\mathbf{h}||_2^2) \quad \leq \quad \lambda\sqrt{q}||\mathbf{h}_{S_0}||_2 + C\sqrt{\frac{2q\log p}{n}}(\frac{5}{4} + \frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2$$

Next we establish an inequality between $||\mathbf{h}||_2^2$ and $||\mathbf{h}_{S_0}||_2^2$. Actually, we have $||\mathbf{h}||_2^2 = ||\mathbf{h}_{S_0}||_2^2 + \sum_{j\geq 1}||\mathbf{h}_{S_j}||_2^2 \geq ||\mathbf{h}_{S_0}||_2^2$, and

$$\begin{aligned}\sum_{j\geq 1}||\mathbf{h}_{S_j}||_2^2 &\leq |\mathbf{h}_q|\sum_{j\geq 1}||\mathbf{h}_{S_j}||_1 \\ &\leq \frac{1}{\bar{C}}|\mathbf{h}_q|||\mathbf{h}_{S_0}||_1 \\ &\leq \frac{1}{\bar{C}}||\mathbf{h}_{S_0}||_2^2.\end{aligned}$$

So $||\mathbf{h}||_2^2 \leq (1+\frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2^2$, and then $o_p(||\mathbf{h}||_2^2) = o_p(||\mathbf{h}_{S_0}||_2^2)$.

To wrap up, we have

$$||\mathbf{h}_{S_0}||_2 + o_p(||\mathbf{h}||_2) \leq \frac{2\lambda\sqrt{q}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4} + \frac{1}{\bar{C}}).$$

Hence,

$$||\mathbf{h}||_2 + o_p(||\mathbf{h}||_2) \leq \sqrt{1+\frac{1}{\bar{C}}}\left(\frac{2\lambda\sqrt{q}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4} + \frac{1}{\bar{C}})\right). \tag{B.5}$$

Let $\mathbf{h} = \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \in \Delta_{\bar{C}}$, we have

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leq \sqrt{1+\frac{1}{\bar{C}}}\Big(\frac{2\lambda\sqrt{q}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4} + \frac{1}{\bar{C}})\Big)$$

with probability at least $1 - 2p^{-4q(C_2^2-1)+1}$. □

Proof of Theorem 4.4.5. It follows by combining the result of Theorem 4.4.4 with that of Theorem 4 in of [60]. □

Appendix C

Discussions of Condition (B4)

Condition (B4) requires the lower bound of the eigenvalues of $H(\boldsymbol{\beta})$ to be positive around $\boldsymbol{\beta}^*$. In the following, we provide a set of sufficient conditions for (B4).

(B1*) For some $1 \leq k \leq p$,

$$\int_{\mathcal{S}} I(X_k \geq V_k^-) X_i g(\mathbf{X}) d\mathbf{X} < \int_{\mathcal{S}} I(X_k \leq U_k^+) X_i f(\mathbf{X}) d\mathbf{X}$$

or

$$\int_{\mathcal{S}} I(X_k \leq V_k^+) X_i g(\mathbf{X}) d\mathbf{X} > \int_{\mathcal{S}} I(X_k \geq U_k^-) X_i f(\mathbf{X}) d\mathbf{X}$$

Here U_k^+, $V_k^+ \in [-\infty, +\infty]$ are upper bounds such that $\int_{\mathcal{S}} I(X_k \leq U_k^+) f(\mathbf{X}) d\mathbf{X} = \min(1, \frac{\pi_-}{\pi_+})$ and $\int_{\mathcal{S}} I(X_k \leq V_k^+) f(\mathbf{X}) d\mathbf{X} = \min(1, \frac{\pi_+}{\pi_-})$. Similarly, lower bounds U_k^-, $V_k^- \in [-\infty, +\infty]$ and are defined as $\int_{\mathcal{S}} I(X_k \geq U_k^-) f(\mathbf{X}) d\mathbf{X} = \min(1, \frac{\pi_-}{\pi_+})$ and $\int_{\mathcal{S}} I(X_k \geq V_k^-) g(\mathbf{X}) d\mathbf{X} = \min(1, \frac{\pi_+}{\pi_-})$.

(B2*) For an orthogonal transformation A_j that maps $\frac{\boldsymbol{\beta}_-^*}{||\boldsymbol{\beta}_-^*||_2}$ to the j-th unit vector $\mathbf{e}_j$ for some $j \in \{1, 2, 3, \ldots, p\}$, there exists rectangles

$$D^+ = \{x \in M^+ : l_i \leq (A_j x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq j\}$$

and

$$D^- = \{x \in M^- : l_i \leq (A_j x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq j\}$$

such that $f(x) \geq B_1 > 0$ on D^+, and $g(x) \geq B_2 > 0$ on D^-, where $M^+ = \{x \in \mathbf{R}^p | x^T \boldsymbol{\beta}_-^* + \boldsymbol{\beta}_0^* = 1\}$ and $M^- = \{x \in \mathbf{R}^p | x^T \boldsymbol{\beta}_-^* + \boldsymbol{\beta}_0^* = -1\}$.

Moreover, Condition (B1) can be further relaxed and no bound restriction is needed on X_j. We refer the modified Condition (B1) as

(B3*) The densities f and g are continuous with common support $\mathcal{S} \subset \mathbb{R}^p$ and have finite second moments.

As a side result, Lemma 5 in [73] showed that Condition (B4) holds under (B1*)-(B3*). Although their paper's results on the Bahadur representation of L_1-norm SVM coefficients are restricted to the classical fixed p case, a careful examination of the derivation showed that this particular lemma holds irrespective of the dimension of p. We refer to [73] for more discussions on the implications for these two conditions.

In the following, we demonstrate that Conditions (B1*)-(B3*) hold in a nontrivial example where we have two multivariate normal distributions in $\mathcal{R}^p$. The marginal distribution of Y is given by $\pi_+ = \pi_- = 1/2$. Let f and g be the density functions of $\mathbf{X}_-$ given $Y = 1$ and -1, respectively. Here, we assume f and g are multivariate normal densities with different mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and a common covariance matrix $\boldsymbol{\Sigma}$. This setup was also considered in [73] but we will provide more details to show condition (B4) is satisfied in our high-dimensional setting. In particular, we will provide some details for deriving the analytic forms of $\boldsymbol{\beta}^*$ and $H(\boldsymbol{\beta}^*)$, which complements the results in [73].

For normal density functions f and g, it is straightforward to check Condition (B3*) is satisfied. While $U_k^+ = V_k^+ = +\infty$ and $U_k^- = V_k^- = -\infty$, Condition (B1*) also holds. Since D^+ and D^- are bounded rectangles in $\mathbb{R}^p$, the normal densities f and g are always bounded away from zero on D^+ and D^-. Thus (B2*) is satisfied. Denote the density and cumulative distribution function of standard normal distribution $N(0,1)$ as ϕ and Φ, respectively. Then we have $S(\boldsymbol{\beta}^*) = 0$, where $S(\cdot)$ is defined in (**??**), that is

$$E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)) = E_g(I(1 + \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)) \tag{C.1}$$

and

$$E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) = E_g(I(1 + \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) \tag{C.2}$$

For left hand of equation (C.1), we have $\mathbf{X}_-^T\boldsymbol{\beta}_-^* \sim N(\boldsymbol{\mu}^T\boldsymbol{\beta}_-^*, \boldsymbol{\beta}_-^{*T}\boldsymbol{\Sigma}\boldsymbol{\beta}_-^*)$, thus

$$E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)) = P_f(1 - \beta_0^* - \mathbf{X}_-^T\boldsymbol{\beta}_-^* \geq 0) = \Phi(c_f), \tag{C.3}$$

where $c_f = \frac{1-\beta_0^* - \boldsymbol{\mu}^T\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$. Similarly, $E_g(I(1+\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)) = \Phi(c_g)$. where $c_g = \frac{1+\beta_0^* + \boldsymbol{\nu}^T\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$.

To obtain an analytic expression of $E_f(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0))$, we consider an orthogonal matrix $\mathbf{P}$ that satisfies $\frac{\mathbf{P}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2} = (1,0,0,\ldots,0)^T$. Such a matrix $\mathbf{P}$ can always be constructed. Actually, let $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_p)^T$ and $\mathbf{P}_1 = \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$. By using Gram-Schmidt process, we can generate other orthogonal vectors $\mathbf{P}_i$ based on $\mathbf{P}_1$ with $i = 2,3,\ldots,p$. Since $\mathbf{P}\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_- - \boldsymbol{\mu}) = \mathbf{Z}$, a standard multivariate normal random vector, we have $I - \mathbf{X}^T\boldsymbol{\beta}^* = c_f||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2 - Z^T\mathbf{P}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*$. Thus

$$\begin{aligned} E_f(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) &= E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^T\mathbf{Z} + \boldsymbol{\mu})) \\ &= E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)\boldsymbol{\mu}) + E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)\boldsymbol{\Sigma}^{1/2}\mathbf{P}^T\mathbf{Z}). \end{aligned}$$

where $\boldsymbol{\phi}$ is the joint probability density function of a p-dimensional standard multivariate normal distribution. We will compute the above expectation componentwise. Without loss of generality, we consider $E_f(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-)$. As we discussed before, we just need to solve one entry of $E_f(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-)$. Without loss of generality, we calculate Let $\Sigma^{1/2} = \Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_p)^T$. For $k = 1,\ldots,p$, we have by (C.4),

$$\begin{aligned} E_f(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)X_k) &= E_\phi(I(c_f - Z_1 \geq 0)\mu_1) + E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)\boldsymbol{\Sigma}^{1/2}\mathbf{P}^T\mathbf{Z}) \\ &= \mu_k\Phi(c_f) + E_{\boldsymbol{\phi}}\big(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^p P_iZ_i\big) \\ &= \mu_1\Phi(c_f) + E_\phi(I(c_f - Z_1 \geq 0)\Lambda_1^TP_1Z_1) \end{aligned}$$

where. Since $Z_2,\ldots,Z_p$ have mean zero and are independent of Z_1,

$$\begin{aligned} E_{\boldsymbol{\phi}}\big(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^p(P_iZ_i)\big) &= E_\phi\big(I(c_f - Z_1 \geq 0)\Lambda_k^TP_1Z_1\big) \\ &= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}E_\phi\big(I(c_f - Z_1 \geq 0)Z_1\big) \\ &= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}\int_{-\infty}^{+\infty} I\big(c_f - x \geq 0\big)x\phi(x)dx \\ &= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}\int_{-\infty}^{c_f} x\phi(x)dx \end{aligned}$$

Since $x\phi(x)$ is an odd function and $\phi(x)$ is symmetric, we have

$$\int_{-\infty}^{c_f} x\phi(x)dx \;=\; \int_{-\infty}^{|c_f|} x\phi(x)dx = -\frac{1}{\sqrt{2\pi}}\int_{c_f^2}^{+\infty}\frac{1}{2}\exp(-\frac{z}{2})dz = -\phi(c_f).$$

Therefore, for $k = 1, \ldots, p$, $E_f\big(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)X_k\big) = \mu_k\Phi(c_f) - \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}\phi(c_f)$. By following the same procedure, we can obtain Hence

$$E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) = \boldsymbol{\mu}\Phi(c_f) - \phi(c_f)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1.$$

Similarly,

$$E_g(I(1 + \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) = \boldsymbol{\nu}\Phi(c_g) + \phi(c_g)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1$$

Then, we have

$$\Phi(c_f) = \Phi_(c_g) \tag{C.4}$$

and

$$\boldsymbol{\mu}\Phi(c_f) - \phi(c_f)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1 = \boldsymbol{\nu}\Phi(c_g) + \phi(c_g)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1 \tag{C.5}$$

From (C.4), we have $\tilde{c} = c_f = c_g$, which implies

$$\boldsymbol{\beta}_-^{*T}(\boldsymbol{\mu} + \boldsymbol{\nu}) = -2\beta_0^* \tag{C.6}$$

From (C.5),

$$\frac{\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2} = \frac{\Phi(\tilde{c})}{2\phi(\tilde{c})}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}) \tag{C.7}$$

Let $d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) = ((\boldsymbol{\mu} - \boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}))^{1/2}$ be the Mahalanobis distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $R(x) = \frac{\phi(x)}{\Phi(x)}$. As $\boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$ has l_2 norm equal to 1, we have $||\frac{\Phi(\tilde{c})}{2\phi(\tilde{c})}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \boldsymbol{\nu})||_2 = 1$, *i.e.*, $R(\tilde{c}) = \frac{d_\Sigma(\boldsymbol{\mu},\boldsymbol{\nu})}{2}$. $R(x)$ is a monotonically decreasing function, thus we have $\tilde{c} = R^{-1}\big(\frac{d_\Sigma(\boldsymbol{\mu},\boldsymbol{\nu})}{2}\big)$. Meanwhile, $\tilde{c} = c_f = \frac{1-\beta_0^* - \boldsymbol{\mu}^T\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$, we can solve the problem based on (C.6) and (C.7),

$$\beta_0^* = -\frac{(\boldsymbol{\mu} - \boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} + \boldsymbol{\nu})}{2\tilde{c}d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) + d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})} \tag{C.8}$$

From (C.5),

$$\boldsymbol{\beta}_-^* = \frac{2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})}{2\tilde{c}d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) + d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})} \tag{C.9}$$

By plugging (C.8) and (C.9) into (**??**), we can calculate $H(\boldsymbol{\beta}^*)$ as

$$H(\boldsymbol{\beta}^*) = \frac{\phi(\tilde{c})}{4}(2\tilde{c} + d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu})) \begin{pmatrix} 2 & (\boldsymbol{\mu}+\boldsymbol{\nu})^T \\ \boldsymbol{\mu}+\boldsymbol{\nu} & H_{22}(\boldsymbol{\beta}^*) \end{pmatrix} \tag{C.10}$$

where

$$H_{22}(\boldsymbol{\beta}^*) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\nu}\boldsymbol{\nu}^T + 2\boldsymbol{\Sigma} + 2\left(\left(\frac{\tilde{c}}{d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu})}\right)^2 + \frac{\tilde{c}}{d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu})} - \frac{1}{d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})}\right)(\boldsymbol{\mu}-\boldsymbol{\nu})(\boldsymbol{\mu}-\boldsymbol{\nu})^T$$

As we have obtained the analytic form of $H(\boldsymbol{\beta}^*)$, we consider Model 1 in Section 5.1 as an example. In Model 1, $q = 5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T$ and $\boldsymbol{\nu} = (-0.1, -0.2, -0.3, -0.4, -0.5, 0, \ldots, 0)^T \in \mathbb{R}^p$ and $\pi^+ = \pi^- = 1/2$. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ consists of nonzero entries $\sigma_{ij} = -0.2$ for $1 \leq i \neq j \leq q$ and other entries equal to 0. From (C.8) and (C.9), we have $\boldsymbol{\beta}^* = (0, 1.39, 1.47, 1.56, 1.65, 1.74, 0, \ldots, 0)^T$. Based on (C.10), we can derive the $H(\boldsymbol{\beta}^*)$ and numerically validate its positive-definiteness.