High-dimensional Generalized Robust Regression and Outlier Detection

by

Yibo Wang

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

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

A great deal of statistical research has been done in high- and ultrahigh-dimensional settings in recent years. Regularized approaches have been extensively used in dealing with high-dimensional datasets. It is widely acknowledged that robust procedures are important to deal with the influence of outliers in high- and ultrahigh-dimensional regression problems. The methods based on the least squares regression produce satisfactory performance only when data have symmetric and light-tailed distributions. Quantile regression and least absolute deviation regression methods have been widely used to address data with heavy-tailed errors. However, quantile regression and least absolute deviation regression are less efficient. To this end, in this thesis, we aim to solve two problems: (i) Estimating the regression vector when both outliers and leverage points are present; (ii) Identifying the locations of outliers when the observations are contaminated and performing robust parameter estimation.

To handle the first problem, we propose two different procedures: the generalized adaptive robust regression (GAR) and l_q robust regression. To achieve this goal, a two-step procedure with adaptive weights in the l_1-penalty function is developed. We exhibit that both GAR regression and l_q robust regression estimators possess the oracle properties.

To address the second problem, we develop a new procedure that can perform outlier detection and robust estimation simultaneously. We demonstrate that the new methodology under the multivariate regression model enjoys robust estimation.

Extensive simulation results and real data examples are used to illustrate that the proposed new methods can handle the situation where outliers occur in the response and covariates with success.

I dedicate this to my grandfather.

Acknowledgements

I would like to begin by expressing my tremendous gratitude to my supervisor Dr. Rohana J. Karunamuni for his patient help and continuous guidance throughout my research program. His advice and encouragement have always been my huge inspiration.

I would like to thank Dr. Linglong Kong, Dr. Adam Kashlak, Dr. Feng Dai and Dr. Cristina Anton for serving in my graduate committee. I also wish to thank the external examiner Dr. Xuewen Lu, Department of Mathematics and Statistics, University of Calgary.

Moreover, I would like to express my thanks to all professors who have given me huge support during my studies at the University of Alberta and the Department of Mathematical and Statistical Sciences for offering me this opportunity.

Last but not least, I am especially grateful to my family for all of their love through what seemed like endless years of studying and research.

Contents

List of Tables

Nomenclature

Notation

$\lim_{n\to\infty} a_n = a$	a is the limit of the sequence a_n
$\|\mathbf{x}\|_1$	the l_1 norm of vector $\mathbf{x}$ defined as $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} \|x_i\|$
$\|\mathbf{x}\|_2$	the l_2 norm of vector $\mathbf{x}$ defined as $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
$\|\mathbf{x}\|_\infty$	the infinity norm of vector $\mathbf{x}$ defined as $\|\mathbf{x}\|_\infty = \max_i \|x_i\|$
$\|\mathbf{x}\|_0$	the number of nonzero components of vector $\mathbf{x}$
$\log a$	the natural logarithm (base e) of a
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^+$	the set of positive real numbers
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$N(0, 1)$	standard normal distribution in $\mathbb{R}$
$N(0, I_{n\times n})$	standard normal distribution in $\mathbb{R}^n$
$U(a, b)$	uniform distribution on the interval (a, b)
$\circ$	the Hadamard product (i.e., the componentwise product of two vectors)
$\mathcal{C}^1$	the continuously differentiable functions

$\phi(\cdot)$	pdf of the standard normal distribution
$\Phi(\cdot)$	cdf of the standard normal distribution
$\mathbb{I}(A)$	the indicator function of set A
A^c	the complement of set A
$\{a, b\}$	the set consisting of the elements a and b
$f(x) = O(g(x))$ for $x \to a$	"big O" means: $f(x)/g(x) \to C$ as $x \to a$, constant $C \neq 0$
$f(x) = o(g(x))$ for $x \to a$	"small o" means: $f(x)/g(x) \to 0$ as $x \to a$
$\xrightarrow{\text{P}}$	convergence in probability
$\xrightarrow{\text{D}}$	convergence in distribution
$X_{(k)}$	the kth order statistic
$\lvert x \rvert$	the absolute value of x
$f^{(1)}$	the first derivative of function f
$f^{(2)}$	the second derivative of function f
z_+	the positive part of z
$\lfloor x \rfloor$	the largest integer no larger than x

Chapter 1

Background

Many widely known parametric statistical models, including certain linear multivariate regression models, generalized linear models, single-index models, and the mean-shift model, are models carrying with covariates. Often many covariates are included in studies, but only a part of these observed variables is believed to be truly relevant to the response due to sparsity. For instance, in medical experiments particular models relating covariates to treatment effects are often adopted more for convenience and simplicity of interpretation than for validity. Variable selection methods are useful for identifying a subset of covariate variables associated with a response variable and for parameter estimation simultaneously. Effective variable selection can also lead to parsimonious models with better prediction accuracy and clearer interpretation. In recent years, a considerable amount of research has been devoted to this area, and many methods have since been developed. These methods have had varying degrees of success in dealing with contaminated data, however.

A common problem in applied statistics is the presence of outliers in the data. Outliers often occur in high-dimensional datasets. Furthermore, statistical models are just approximations to reality and real data never come from the specified model exactly. Therefore, the need for robust procedures in statistical inference has been widely recognized now. (Here the word 'robust' refers to the ability of a procedure to retain its validity under a model misspecification and/or when outliers are present.) The importance of robust

procedures has also been stressed for the variable selection methods, also known as the regularization (penalized) methods. Most well-known regularization methods are based on solving an optimization problem formed by the sum of a 'loss function' and a 'penalty function' - the resulting estimators are referred to as regularized M-estimators (Negahban et al. 2012). For instance, in application to linear models, the least absolute shrinkage and selection operator (LASSO) is based on a combination of the squared error loss with an l_1-penalty function, and so involves solving a quadratic program. Similar approaches have been applied to generalized linear models, resulting in more general (nonquadratic) convex programs with l_1-constraints.

A penalty function generally encourages variable selection in regression models, and various penalized regression methods have been proposed in the literature. Among them, bridge regression (Frank and Friedman 1993), Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), adaptive Lasso (Zou 2006), elastic-net (Zou and Hastie 2005), adaptive elastic-net (Zou and Zhang 2009), and MCP (Zhang 2010) are well-known. Efron et al. (2004) proposed the LARS algorithm for computing the entire LASSO solution path. Knight and Fu (2000) studied the asymptotic properties of the Lasso. Fan and Li (2001) showed that the SCAD enjoys the oracle property; that is, the SCAD estimator can perform as well as the oracle if the penalization parameter is appropriately chosen. The oracle property is also satisfied by the adaptive Lasso estimator (Zou 2006) and the adaptive elastic net estimator (Zou and Zhang 2009). Nonetheless, most well-known regularized methods such as penalized least squares and penalized likelihood are not designed for heavy-tailed distributions and are not robust in the presence of the outliers and model misspecification, thus prompting for methods that are more robust against outliers, in particular.

1.1 Multivariate linear regression model

The multivariate linear regression model is given by

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i, \ i = 1, ..., n, \tag{1.1}$$

where $\mathbf{X}_i \in \mathbb{R}^p$ is a vector of p predictor (design) variables, Y_i is the univariate response variable, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown regression parameter, and the error terms e_i's are independent of the $\mathbf{X}_i$'s. The lack of robustness of regularization estimators based on the squared error loss or the l_2-loss is widely known. Specifically, the outlying values of $\mathbf{X}_i$ (leverage points) or extreme values of $(\mathbf{X}_i, Y_i)$ (influence points) jointly can have an arbitrarily large influence on the l_2-loss based estimators. For model (1.1), a number of robust regularized methods have been proposed in the literature. They include, among others, the following notable works. Fan and Li (2001) established a general class of penalized robust regression estimators based on the Huber function (Huber 1981). Wang et al. (2007) proposed the LAD-LASSO with l_1-loss and $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj} |\beta_j|$, and Arslan (2012) provided a weighted version of the LAD-LASSO estimator that is more robust to leverage points. Johnson and Peng (2008) studied rank-based variable selection, and Wang and Li (2009) proposed a weighted Wilcoxon-type SCAD method for robust variable selection. Wu and Liu (2009) and Wang et al. (2012) investigated penalized quantile regression where $p_{\lambda_{nj}}(|\beta_j|)$ is either the SCAD or the adaptive Lasso penalty. Leng (2010) investigated variable selection via regularized rank regression, and Chen et al. (2010) proposed weighted l_2- and l_1-loss functions. Kai et al. (2011) examined variable selection in the semiparametric varying-coefficient partially linear model via a penalized composite quantile loss (Zou and Yuan 2008). Lambert-Lacroix and Zwald (2011) proposed to use the Huber loss together with the adaptive lasso penalty for robust estimation. Wang et al. (2013) implemented a bounded loss function of the form $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$ with γ a tuning parameter that controls the degree of robustness and efficiency of the estimator for the fixed-dimensional case. Alfons et al. (2013) and Öllerer et al. (2015) introduced

a sparse least trimmed squares estimator. Smucler and Yohai (2017) developed a robust l_1-penalized MM-estimator (Yohai 1987) with an adaptive l_1-penalty. Karunamuni et al. (2019) proposed an adaptive efficient robust regularized procedure.

Most of the above mentioned articles on robust estimation are for the fixed-dimensional case (i.e., p is fixed). On the other hand, regularized robust procedures for high- or ultrahigh-dimensional cases (i.e., p grows as a function of n) are rather sparse. Notable works include the following contributions. Belloni and Chernozhukov (2011) considered l_1-penalized quantile regression procedure in a high-dimensional setting. Li et al. (2011) examined a nonconcave penalized robust M-estimator, again in a high-dimensional setting. Bradic et al. (2011) studied a penalized composite likelihood method for ultrahigh-dimensional robust variable selection. van de Geer and Müller (2012) obtained bounds on the prediction error of a large class of l_1-penalized estimators that includes quantile regression. Wang et al. (2012) considered the nonconvex penalized quantile regression in an ultrahigh-dimensional setting. Wang (2013) studied an l_1-penalized LAD estimator for high-dimensional regression. In an ultrahigh-dimensional setting, Fan et al. (2014) investigated a penalized quantile regression procedure with a weighted l_1-penalty for robust regularization, as in Bradic et al. (2011). Fan et al. (2017) proposed an l_1-penalized procedure based on the Huber loss ρ_α with diverging parameter α in an ultrahigh-dimensional setting. Loh (2017) studied a class of generalized robust M-estimators with an "amenable (nonconvex) regularizer" for high-dimensional data. Loh (2018) explored an adaptive scale estimation using the Huber function. All these estimators differ greatly in terms of outlier resistance and efficiency under the model. For instance, Loh (2017)'s regularized estimators are stable against both the influence and leverage points, whereas the quantile regression based methods are robust only with respect to outliers in the response variable.

Other notable works include She and Owen (2011), where a novel procedure called IPOD procedure was established for outlier detection upon a class of penalty functions. They found that the version based on hard thresholding excels in correctly identifying outliers on some hard test problems. Nguyen and Trac (2012) developed a procedure

named Extended-LASSO via using two l_1-penalty functions on the linear regression vector $\boldsymbol{\beta}$ and the mean-shift vector γ, respectively. They provided error bounds and signed support recovery results for both the regression and corrupt vectors. Lee et al. (2012) studied a general loss function where they introduced a case-specific parameter vector $\boldsymbol{e} \in \mathbb{R}^p$ for the observation vectors and took $\boldsymbol{e}$ into account while minimizing the objective function. Alain et al. (2017) developed a new procedure for simultaneous estimation of the linear regression and mean-shift vectors via using two dedicated sorted l_1-penalty functions, called SLOPE. Kong et al. (2018) used an adaptive penalty on γ depending on the residuals from some robust initial fit and the l_1-penalty on $\boldsymbol{\beta}$ to achieve variable selection. Differing from She and Owen (2011), their procedure can attain high breakdown point by judiciously choosing the penalty functions.

1.2 Outline of the thesis

Despite considerable progress on variable selection in various models in high-dimensional problems, robustness issues of regularization methods have not been thoroughly studied and well understood. Efficiency and robustness are extremely important in the practice of statistics. Most well-known existing variable selection methods aforementioned, however, fail to achieve both of these goals simultaneously. To address this problem, we study and propose new methods in high-dimensional settings that attain desirable properties such as full oracle efficiency, maximum robustness, and computational feasibility. Our contributions are listed below:

1. In Chapter 2, we provide a generalized robust regression method with a weighted penalty function. The model consistency and related oracle properties of this new method are thoroughly studied. The proposed method when compared with other existing methods excels in (1) highly robust with respect to outliers in both the responses and covariates and (2) robust against heavy-tailed error distributions. In addition, the asymptotic results of the new method under a high-dimensional setting

are explored in this chapter. We believe that the proposed procedure can also be extended to address ultrahigh-dimensional data.

2. In Chapter 3, we investigate robust regression estimation using l_q-loss functions. Again in a high-dimensional setting, we study the asymptotic properties of the new estimator, including model consistency and oracle properties with only sub-Gaussian assumption on the model error distribution. In addition, to circumvent the non-smoothness of l_q-loss function, we utilize a half-quadratic approximation of convex functions to facilitate computation.

3. In Chapter 4, we consider a setting in which the observations are contaminated. Specifically, we develop a new method to tackle the situations where errors have either sub-Gaussian or sub-exponential distributions. We show that, under these two circumstances, the proposed procedure enjoys robust estimation and simultaneously performs outlier detection. We obtain bounds on the error of estimation. These bounds are non-asymptotic. Extensive simulation experiments are used to demonstrate the theoretical properties of the proposed procedure.

The proofs of the main results presented in Chapters 2, 3 and 4 are given in Appendices A, B and C, respectively.

Chapter 2

High-dimensional generalized robust regression

2.1 Introduction

Consider the linear regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i, \ \ i = 1, ..., n, \tag{2.1}$$

where $Y_i \in \mathbb{R}$ is a response variable, $\mathbf{X}_i \in \mathbb{R}^{p_n}$ is a covariate vector, $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ is a regression coefficient vector, and $e_i \in \mathbb{R}$ is an error term. We assume that $\{e_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) with the unknown distribution $\mathbb{F}$. We also assume that $\{\mathbf{X}_i\}_{i=1}^n$ are i.i.d. random vectors and that the e_i's are independent of the $\mathbf{X}_i$'s. We are interested in robust estimation of the p_n-dimensional coefficient vector $\boldsymbol{\beta}$ where p_n may increase with n and $\boldsymbol{\beta}$ is sparse in the sense that many of the elements are zero. We assume that the data are centered, so the intercept term is zero.

Robust procedures are important because outliers are often present in data coming from the model (2.1). Regularization methods based on the squared error loss lack robustness. Specifically, the outlying values of $\mathbf{X}_i$ (leverage points) or the extreme values of $(\mathbf{X}_i, Y_i)$ (influence points) can have an arbitrarily large influence on l_2-loss based estimators. A number of robust regularized methods have therefore been proposed. Fan and

Li (2001) examined a general class of penalized robust regression estimators of the form

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|), \tag{2.2}$$

where $p_{\lambda_{nj}}(|\beta_j|)$ is a penalty function (regularizer) on β_j, and ρ_α is the Huber function with tuning parameter α [Huber (1964)]: $\rho_\alpha(t) = \frac{1}{2}t^2 I[|t| \leq \alpha] + \alpha(|t| - \frac{1}{2}\alpha)I[|t| > \alpha]$ for $\alpha > 0$. Since then, many penalized robust regression estimators have been proposed based on various loss and penalty functions for the fixed-dimensional case. For example, Wang et al. (2007) proposed the LAD-LASSO with l_1-loss and $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}\,|\beta_j|$. A weighted version of the LAD-LASSO estimator was introduced by Arslan (2012). In comparison with LAD-LASSO, weighted LAD-LASSO is more robust to leverage points. In addition, Johnson and Peng (2008) studied rank-based variable selection, and Wang and Li (2009) proposed a weighted Wilcoxon-type SCAD method for robust variable selection. Wu and Liu (2009) and Wang et al. (2012) investigated penalized quantile regression with $p_{\lambda_{nj}}(|\beta_j|)$ either the SCAD or the adaptive LASSO penalty. Leng (2010) investigated variable selection via regularized rank regression, and Chen et al. (2010) proposed weighted l_2- and l_1-loss functions. Kai et al. (2011) examined variable selection in the semiparametric varying-coefficient partially linear model via a penalized composite quantile loss proposed by Zou and Yuan (2008). A combination of the Huber loss and the adaptive LASSO penalty was established by Lambert-Lacroix and Zwald (2011). With fixed dimensions, Wang et al. (2013) considered a bounded loss function of the form $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$ with γ a tuning parameter used to control the degree of robustness and efficiency of the estimator. Other noted contributions include Yohai (1987), Alfons et al. (2013), Öllerer et al. (2015), Smucler and Yohai (2017), and Karunamuni et al. (2019).

Robust regularized procedures for high- and ultrahigh-dimensional cases are relatively scarce. Notable contributions include the following. Fan and Lv (2008) investigated the sure independence screening method in the setting of light-tailed distributions. Belloni and Chernozhukov (2011) considered an l_1-penalized quantile regression

in a high-dimensional setting. Li et al. (2011) examined a nonconcave penalized robust M-estimator, again in a high-dimensional setting. Bradic et al. (2011) studied a penalized composite quasi-likelihood method for ultrahigh-dimensional robust variable selection. van de Geer and Müller (2012) obtained bounds on the prediction error of a large class of l_1-penalized estimators that includes quantile regression. Wang et al. (2012) considered a nonconvex penalized quantile regression in an ultrahigh-dimensional setting. Wang (2013) studied an l_1-penalized LAD estimator for high-dimensional regression. In an ultrahigh-dimensional setting, Fan et al. (2014) investigated a penalized quantile regression procedure with a weighted l_1-penalty for robust regularization, as in Bradic et al. (2011). She and Chen (2017) proposed a robust reduced-rank regression approach for joint modeling and outlier detection. Fan et al. (2017) proposed an l_1-penalized procedure based on the Huber loss ρ_α with diverging parameter α (or converging to zero if its reciprocal is used) in an ultrahigh-dimensional setting and obtained nonasymptotic bounds on the l_2-error. In a recent article, Sun et al. (2020) also dealt with the Huber loss and obtained nonasymptotic bounds on the l_2-error in high-dimensional settings. Loh (2017) studied a generalized robust M-estimator with an "amenable (nonconvex) regularizer" for high-dimensional data. This estimator excels quantile-regression-based methods in that it is stable against both the influence and leverage points. On the other hand, quantile-based oracle estimators attain "semiparametric efficiency" (Fan et al. 2014).

Most robust regularized regression methods usually focus on getting a handle on the outliers as one wants it – the price for that being a loss in efficiency. It has gradually gotten through to researchers that this efficiency deflation is important and that high-dimensional robust methods possessing high efficiency are worth investigating. On the other hand, the downweighting of outliers is often used to achieve robustness in a robust M-estimation context and many authors mentioned above have successfully implemented this technique for regularized regression. However, the approaches involving only downweighting of outliers without regard to model fit are always associated with a larger variance and hence less efficient. Specifically, these methods usually suffer from a loss of efficiency if

there are no outliers, since the observations with large covariate values are downweighted even if they are well-fitted. To achieve high efficiency and high robustness simultaneously, it is necessary to downweight the outliers adaptively (e.g., Gervini and Yohai 2002, Wang et al. 2013, Karunamuni et al. 2019). Our present work can be viewed as an improvement of a particularly sensible class of methods that can attain high efficiency while keeping the robustness in check for various types of outliers. To this end, we investigate generalized adaptive robust regularized estimators of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \phi((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_i/\gamma) + np_{\lambda_n}(\boldsymbol{\beta}), \tag{2.3}$$

where ϕ is a loss function, W is a non-negative weight function on large values of $\mathbf{X}_i$, $p_{\lambda_n}(\boldsymbol{\beta})$ is a coordinate-separable regularizer (penalty function) on $\boldsymbol{\beta}$ with regularization parameter $\lambda_n \geq 0$, and $\gamma > 0$ is a tuning parameter. The parameter γ controls the degree of efficiency/robustness of the estimator. We assume that ϕ is convex, nondecreasing on $[0, \infty)$ and has a bounded gradient or subgradient. This includes well-known loss functions such as the Huber loss, the quantile loss functions, and the l_1-loss, among others. It is well-known that loss functions with a bounded gradient lead to bounded influence functions in the fixed-covariate setting. For the weight function $W(\mathbf{x})$, we use a nonnegative function so that for large values of $\mathbf{X}_i$, $W(\mathbf{X}_i)$ would place a small weight on $\mathbf{X}_i$ and therefore have a small impact on $\hat{\boldsymbol{\beta}}$. The role of tuning parameter γ is to increase the efficiency of $\hat{\boldsymbol{\beta}}$ while holding on to high robustness. In general, for small values of γ, $\hat{\boldsymbol{\beta}}$ can be expected to be robust against both the influence and leverage data points, whereas $\hat{\boldsymbol{\beta}}$ would be more efficient for large values of γ. The proposed estimators are also robust against heavy-tailed error distributions. For a given pair of loss and penalty functions, γ may be chosen by minimizing the asymptotic variance or as a measure reflecting the goodness-of-fit of the model. In practice, the parameter γ is generally determined by a data-driven method.

We study the asymptotic properties of $\hat{\boldsymbol{\beta}}$ under two forms of the penalty function $p_{\lambda_n}(\boldsymbol{\beta})$. First, we consider a weighted l_1-penalty of the form $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \left\|\mathbf{d} \circ \boldsymbol{\beta}\right\|_1$, where $\|\mathbf{x}\|_1$ denotes the l_1-norm of any vector $\mathbf{x} = (x_1, ..., x_{p_n})^T$, $\mathbf{d} = (d_1, ..., d_{p_n})^T$ is a vector

of nonnegative weights, and $\circ$ denotes the Hadamard product (i.e., the componentwise product of two vectors). Next, we consider an adaptive weighted l_1-penalty function of the form $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \|\hat{\mathbf{d}} \circ \boldsymbol{\beta}\|_1$, where $\hat{\mathbf{d}} = (\hat{d}_1, ..., \hat{d}_{p_n})^T$ is a vector of non-negative stochastic weights constructed using a folded-concave penalty function and an initial estimator. The stochastic weights are introduced to reduce the bias induced by the l_1-penalty. In each case, we prove the model selection oracle property and establish the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Furthermore, we show that $\hat{\boldsymbol{\beta}}$ is consistent at a near-oracle rate under mild conditions on the error distribution and the design matrix. We impose no conditions on the heaviness of the tail probability of the error terms e_i. We establish asymptotic results in a high-dimensional setting, and they are similar to those in Fan et al. (2014) for quantile regression estimation obtained in an ultrahigh-dimensional setting. We carry out extensive simulation studies to compare the proposed method with other regularized robust estimators, including some quantile regression and least-absolute-deviation regression estimators. Our numerical studies demonstrate the favorable finite-sample performance of the proposed procedure for various shapes and tails of the error distribution. This advantage is most pronounced in the presence of high leverage points.

The rest of this chapter is organized as: in Section 2.2, we study the estimator $\hat{\boldsymbol{\beta}}$ using a weighted l_1-penalty function and investigate its asymptotic properties, including the estimation consistency and oracle properties. In Section 2.3, we investigate asymptotic properties using an adaptive penalty function. In Section 2.4, we present numerical studies and compare our method with other existing methods, and in Section 2.5, we illustrate the performance of our method using a real dataset. Section 2.6 provides concluding remarks. The proofs of the main results and lemmas are given in Appendix A.

2.2 Robust estimation with weighted penalty

In this section, we develop the asymptotic properties of the estimator $\hat{\boldsymbol{\beta}}$ defined by (2.3) with a weighted l_1-penalty of the form $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \|\mathbf{d} \circ \boldsymbol{\beta}\|_1$. The weights are designed

to reduce the bias induced by the l_1-penalty. Such penalty functions have also been recommended for computational expediency in high-dimensional settings (e.g., Bradic et al. (2011), Fan et al. (2014)), motivated by the fact that any penalty function can be approximated as $p_{\lambda_n}(|\beta_i|) \approx p_{\lambda_n}(|\tilde{\beta}_i|) + p^{(1)}_{\lambda_n}(|\tilde{\beta}_i|)(|\beta_i| - |\tilde{\beta}_i|)$, where $\tilde{\beta}_i$ is an initial estimate of β_i and $p^{(1)}_{\lambda_n}$ is the first derivative of p_{λ_n}. The choice $\mathbf{d} = (1, ..., 1)^T$ corresponds to the Lasso penalty.

We begin by introducing some notations. We let $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*T},\ \boldsymbol{\beta}_2^{*T}\right)^T$ denote the true parameter vector of $\boldsymbol{\beta}$, where each element in $\boldsymbol{\beta}_1^* \in \mathbb{R}^{k_n}$ is nonzero and $\boldsymbol{\beta}_2^* = \mathbf{0} \in \mathbb{R}^{p_n - k_n}$. For any weight vector $\mathbf{d}$, let $\mathbf{d}_0$ denote the first k_n elements of $\mathbf{d}$, and $\mathbf{d}_1$ denote the remaining part of $\mathbf{d}$, i.e., $\mathbf{d} = (\mathbf{d}_0^T,\ \mathbf{d}_1^T)^T$. For any vector $\mathbf{x} = (x_1, ..., x_p)^T \in \mathbb{R}^p$, let $\|\mathbf{x}\|_2$ denote the l_2-norm and $\|\mathbf{x}\|_\infty = \max(|x_1|, ..., |x_p|)$. We allow both p_n and k_n to diverge with the sample size n and assume k_n to be $o(n)$; that is, only a small number of true coefficients are nonzero. Let $\mathfrak{X} = (X_{ij}; 1 \le i \le n,\ 1 \le j \le p_n)$ denote the design matrix. We write $\mathfrak{X} = (\mathbf{S},\ \mathbf{Q})$ with the submatrices $\mathbf{S}$ and $\mathbf{Q}$ corresponding to the covariates whose coefficients are nonvanishing and vanishing, respectively. Thus, $\mathbf{S}$ and $\mathbf{Q}$ are $n \times k_n$ and $n \times (p_n - k_n)$, respectively. We will refer to the set of columns in $\mathbf{S}$ as the signal covariates, while those in $\mathbf{Q}$ will be called noise covariates. Further, for $i = 1, ..., n$, let $\mathbf{X}_i = (\mathbf{X}_{1i}^T,\ \mathbf{X}_{2i}^T)^T$, where $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ denote the covariates corresponding to $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, respectively. We assume without loss of generality that the X_{ij}'s are normalized, so $\mathbf{E}(X_{ij}^2) = 1$ for all i and j.

For our proofs, we require the following regularity conditions:

(C1) $\phi : \mathbb{R} \to [0, \infty)$ is a symmetric convex function on $\mathbb{R}$ satisfying $\phi(0) = 0$. Let $\phi^{(1)}$ denote a gradient or subgradient of ϕ. Assume $\phi^{(1)}$ is bounded and satisfies the "local δ-Lipschitz condition" for $\delta \in [0, 1]$, i.e., there exist positive constants ξ and M such that $\left|\phi^{(1)}(x + t) - \phi^{(1)}(x)\right| \le M\,|t|^\delta$ for all $x \in \mathbb{R}$ and $|t| \le \xi$.

(C2) The common distribution function $\mathbb{F}$ of the errors is sufficiently smooth and satisfies $\mathbb{F}(\mathcal{D}) = 0$, where $\mathcal{D}$ denotes the set of discontinuity points of $\phi^{(1)}$. The

function $\phi^{(1)}$ satisfies $\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\right) = 0$, $\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\right)^2 = \sigma_\gamma^2 < \infty$, and as $t \to 0$, $\mathbf{E}(\phi^{(1)}((e_1+t)/\gamma) - \phi^{(1)}(e_1/\gamma))^2 \to 0$,

$$\mathbf{E}\left(\phi^{(1)}((e_1+t)/\gamma)\right) \equiv g(\gamma)t + o(|t|q(\gamma))$$

for some positive functions g and q of γ.

(C3) $W : \mathbb{R}^{p_n} \to (0,1]$ is a weight function on $\mathbf{x} \in \mathbb{R}^{p_n}$ such that $|x_i|^{2+\varepsilon} W(\mathbf{x})$ is bounded for all $x_i \in \mathbf{x}$, $i = 1, ..., p_n$, where $0 < \varepsilon < 1$ is some constant.

(C4) The eigenvalues of $\frac{1}{n}\mathbf{E}\left(\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)$ are bounded from below and above by some positive constants.

In addition, we introduce the following three notations:

$$\mathbf{V}_n = \left(\mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\right)^{-1/2},$$

$\mathbf{Z}_n = \mathbf{S}\mathbf{V}_n$, and

$$\mathbf{Z}_n^* = \left(\frac{1}{\gamma}\mathbf{X}_{11}W\left(\mathbf{X}_1/\gamma\right), \ldots, \frac{1}{\gamma}\mathbf{X}_{1n}W\left(\mathbf{X}_n/\gamma\right)\right)^T \mathbf{V}_n,$$

where function $g(\cdot)$ is defined in condition (C2) above.

It is appropriate to make a few comments here about conditions (C1)–(C4). Conditions (C1) and (C2) are similar to standard assumptions imposed in the classical M-estimation theory of linear regression models; see, e.g., Bai et al. (1992) and Wu (2007). It can be shown that the Lipschitz condition holds for all δ if it is satisfied for some δ. When (C1) holds for $\delta = 0$, then the loss function ϕ is said to have a "locally uniformly bounded increment." This covers the important case of the l_1-loss function. The Huber loss satisfies condition (C2) for symmetric error distributions, for example. Weight functions satisfying condition (C3) can be easily constructed: e.g., $W(\mathbf{x}) = \min\{1, (\|\mathbf{x}\|_\infty)^{-b}\}$ for some $b \geq 3$. Condition (C4) warrants the inverse of eigenvalues of $\frac{1}{n}\mathbf{E}\left(\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)$ exists.

We consider the following objective function:

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) + n\lambda_n \|\mathbf{d} \circ \boldsymbol{\beta}\|_1. \tag{2.4}$$

Note that $Q_n(\boldsymbol{\beta})$ is a convex function of $\boldsymbol{\beta}$ under condition (C1) on the loss function ϕ and for a given weight vector $\mathbf{d}$. Thus, we define the regularized estimator of $\boldsymbol{\beta}^*$, again denoted as $\hat{\boldsymbol{\beta}}$ for notational convenience, by the (global) minimizer of (2.4):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}). \tag{2.5}$$

The next theorem establishes the consistency of $\hat{\boldsymbol{\beta}}$ defined by (2.5), given the oracle information on the location of the signal covariates, i.e., $\boldsymbol{\beta}_2 = \mathbf{0}$, with $\mathbf{0}$ being the vector of all zeros. For this purpose, let $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \mathbf{0}^T\right)^T$ with $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^{k_n}$ denoting the oracle-regularized estimator minimizing $Q_n(\boldsymbol{\beta})$ over the space $\Xi = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbb{R}^{p_n - k_n}\}$. Thus, $\hat{\boldsymbol{\beta}}_1$ represents the estimator of signal covariates. The next theorem gives the consistency of $\hat{\boldsymbol{\beta}}_1$. All limits are taken as $n \to \infty$ unless otherwise stated.

Theorem 2.1. *Assume that conditions* (C1) *to* (C4) *hold. Let* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \mathbf{0}^T\right)^T$ *be defined by* (2.5) *over* $\boldsymbol{\beta} \in \Xi$. *In addition, assume that* $\lambda_n \|\mathbf{d}_0\|_2 \sqrt{k_n/n} \to 0$ *and* $s_n = c\left(\sqrt{k_n(\log n)/n} + \lambda_n \|\mathbf{d}_0\|_2\right)$ *for some constant* $c > 0$. *Then, with probability tending to one, we have*

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le s_n \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \le s_n\sqrt{k_n}.$$

Further, if $s_n^{-1} \min_{1 \le j \le k_n} |\beta_j^*| \to \infty$, *then with probability tending to one, we have*

$$\operatorname{sgn}(\hat{\boldsymbol{\beta}}_1) = \operatorname{sgn}(\boldsymbol{\beta}_1^*),$$

where the above equation should be understood componentwise.

Theorem 2.1 shows that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator of $\boldsymbol{\beta}_1^*$ with an upper bound s_n for the rate of consistency. This result does not depend on p_n as $Q_n(\boldsymbol{\beta})$ is minimized

over the space Ξ to obtain $\hat{\boldsymbol{\beta}}$. We also observe that $\hat{\boldsymbol{\beta}}_1$ estimates the correct sign of the true coefficient vector $\boldsymbol{\beta}_1^*$ with probability tending to one. The first term in s_n, $\sqrt{k_n(\log n)/n}$, is the oracle rate within a factor of $\log n$, and the second term, $\lambda_n\|\mathbf{d}_0\|_2$, is the extra bias term due to regularization. The weights $\mathbf{d}_0 = (1, 1, ..., 1)^T$ are assumed if no prior information is available for the weights. Then we obtain an upper bound rate of $c\left(\sqrt{k_n(\log n)/n} + \lambda_n\sqrt{k_n}\right)$. If $\lambda_n = O(\sqrt{\log n/n})$ then the preceding rate reduces to $c\sqrt{k_n(\log n)/n}$, a near-oracle rate of consistency. The rate result obtained in Theorem 2.1 is in line with, for example, a similar result obtained in notable work by Fan et al. (2014) on ultrahigh-dimensional regularized regression estimation for a quantile regression estimator. They assumed that X_{ij}'s are fixed, however.

The next theorem establishes the model selection oracle property of the proposed regularized robust estimator $\hat{\boldsymbol{\beta}}$ defined by (2.5) without the oracle information.

Theorem 2.2. *Assume that conditions* (C1) *to* (C4) *hold and* $\min_{j\geq k_n+1} d_j$ *is strictly positive. In addition, assume that* $\lambda_n\|\mathbf{d}_0\|_2\sqrt{k_n/n} \to 0$, $\lambda_n > 2\sqrt{(1+c)(\log p_n)/n}$, $\lambda_n > c_0 s_n\sqrt{k_n}$, *and* $\sqrt{(1+k_n^{1+\delta/2}s_n)}(\log_2 n) = o(\sqrt{n}\lambda_n)$, *where* c *and* c_0 *are some positive constants,* δ *is defined in condition* (C1) *, and* s_n *is defined in Theorem* 2.1. *Then, there exists a global minimizer* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T\right)^T$ *defined by* (2.5) *such that* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$, $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq s_n$, *and* $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq s_n\sqrt{k_n}$ *with probability tending to one.*

Theorem 2.2 is in some sense an extension of Theorem 2.1 in which the asymptotic property of Theorem 2.1 is established without the oracle information. However, it imposes each coordinate of the noise covariates to be constrained by a condition that the corresponding weights have $\min_{j\geq k_n+1} d_j$ strictly positive, as in Fan et al. (2014). We also put some conditions on the signal covariates using $\|\mathbf{d}_0\|_2$. Theorem 2.2 holds with heavy-tailed errors provided condition (C2) is satisfied and the regularization parameter λ_n is selected larger than $\sqrt{\log p_n/n}$, the order of optimal choice of λ_n for Gaussian errors (Bickel et al. 2009). Other model selection oracle results have been established in this context for a variety of robust high- and ultrahigh-dimensional regularized M-estimators.

Notably, Loh (2017) investigated a generalized robust M-estimator and established a stronger consistency result. Loh's main result, which gives bounds on both the l_1- and l_2-errors under some assumptions on the regularizer and loss function, is entirely deterministic and guarantees the statistical consistency of stationary points within a local region of restricted strong convexity. Whereas, Theorem 2.2 concerns with the model selection oracle property of a global minimizer in an ultrahigh-dimensional setting.

Theorem 2.2 shows that $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_1$ is consistent for $\boldsymbol{\beta}_1^*$ with probability tending to one. The next theorem presents the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$.

Theorem 2.3. *Assume that the conditions of Theorem* 2.2 *hold. In addition, assume that* $\sqrt{n/k_n}\min_{1\le j\le k_n}\left|\beta_j^*\right| \to \infty$, $k_n = o(n^{\delta/(3+2\delta)})$, $\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2 = O(\sqrt{k_n})$, *and* $k_n = o(n^{\varepsilon/(4+2\varepsilon)})$, *where* δ *and* ε *are as defined in* (C1) *and* (C3), *respectively. Then there exists a global minimizer* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T\right)^T$ *defined by* (2.5) *such that* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *with probability tending to* 1 *and*

$$\mathbf{c}^T\left(\left(\mathbf{Z}_n^*\right)^T\mathbf{Z}_n^*\right)^{-1/2}\left(2\mathbf{V}_n^{-1}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right) + n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0\right) \xrightarrow{\mathcal{D}} N\left(0, \sigma_\gamma^2\right),$$

where $\mathbf{c}$ *is an arbitrary* k_n*-dimensional vector satisfying* $\mathbf{c}^T\mathbf{c} = 1$, *and* $\tilde{\mathbf{d}}_0$ *is a* k_n*-dimensional vector with the* j*th element as* $d_j\,\mathrm{sgn}(\beta_j^*)$, $1 \le j \le k_n$.

Note that moment assumptions such as sub-Gaussian or sub-exponential on the covariates $\{\mathbf{X}_i\}_{i=1}^n$ are not required in Theorems 2.1–2.3. This is an extra advantage of implementing a weight function $W(\cdot)$; it lessens the effect of large leverage points as well as reduces the assumptions on the covariates simultaneously. The theorems stated above are proved in Appendix A. The basic ideas of the proofs adopt the techniques developed in Fan et al. (2014) for regularized quantile regression estimation. However, we extended the arguments of Fan et al. (2014) here to cover a broader class of convex loss functions when combined with adaptive weight functions on the covariates. The present project in this sense complements Fan et al. (2014)'s paper by showing the wider applicability of their techniques. They investigated their estimator in a very challenging ultrahigh-dimensional

setting where $\log(p_n) = o(n^b)$ for some constant $b > 0$. The results in Theorems 2.1–2.3 are similar to those in Theorems 2.1–2.3, respectively, of Fan et al. (2014). But the regularity conditions employed in Theorems 2.1–2.3 above are different from those used in Fan et al. (2014). For instance, their condition (C2) is not necessary here due to the presence of weight function $W(\cdot)$.

Other authors have obtained asymptotic normality results of similar nature for various high-dimensional regularized robust regression M-estimators, e.g., Fan and Peng (2004), Li et al. (2011), Bradic et al. (2011), Fan et al. (2017), and Loh (2017), among others. We make the following observations. First, Theorem 2.3 studies an ultrahigh-dimensional setting, whereas Li et al. (2011) assumed $p_n^2/n \to 0$ and fixed-covariates. Second, the estimator in Theorem 2.3 is robust against outliers in both the response and covariates, whereas the high-dimensional regularized robust estimators studied in Fan and Peng (2004), Li et al. (2011), Bradic et al. (2011), Wagener and Dette (2013), and Fan et al. (2017) are affected by the outliers of the covariates. The latter outliers are known to severely influence the performance of estimators (Huber 1981). It has been observed that the Huber loss function cannot handle even moderate leverage points well (Huber 1981, p. 182), and the breakdown point of estimators based on the Huber loss in such cases is 0 (She and Owen 2011). In addition, we have introduced a new tuning parameter for efficiency improvement. It can considerably improve the efficiency of estimators in the presence of outliers, particularly the outliers of the covariates.

In practice, the tuning parameter γ is determined by a data-driven method. From a nonasymptotic viewpoint, γ can be chosen to balance the bias and robustness; see, e.g., Fan et al. (2017) and Sun et al. (2020). From an asymptotic standpoint on the other hand, γ may be chosen to minimize a sample estimate of the asymptotic variance of $\hat{\boldsymbol{\beta}}_1$, see Proposal 3 of Huber (1964, p. 98). For instance, in light of Theorem 2.3, a data-driven value for γ may be obtained as follows:

$$\hat{\gamma} = \arg\min_{\gamma} \mathbf{c}^T \hat{\mathbf{V}}_n \hat{\mathbf{V}}_n^T \left(\sum_{i=1}^{n} \left(\mathbf{X}_{1i} \mathbf{X}_{1i}^T / \gamma^2 \right) W^2 \left(\mathbf{X}_i / \gamma \right) \right) \hat{\mathbf{V}}_n \hat{\mathbf{V}}_n^T \mathbf{c} \hat{\sigma}_{\gamma}^2,$$

where $\mathbf{c}$ is a vector satisfying $\mathbf{c}^T\mathbf{c} = 1$, and $\hat{\mathbf{V}}_n$ and $\hat{\sigma}_\gamma^2$ denote consistent estimators of $\mathbf{V}_n$ and σ_γ^2, respectively. We employed this method in our numerical studies, see Section 2.4. One may also select γ so that $\sum_{i=1}^n [\phi^{(1)}((Y_i - \mathbf{X}_{1i}^T\hat{\boldsymbol{\beta}}_1)/\gamma)]^2$ has a predetermined value, see Proposal 2 of Huber (1964, p. 97). Huber favored the latter approach because it "fits best into the framework of conventional least squares techniques."

2.3 Robust estimation with adaptive penalty

For the high-dimensional oracle properties established in Theorems 2.2 and 2.3, the weight vector $\mathbf{d}$ plays a pivotal role. A good choice of $\mathbf{d}$ will enable the estimator $\hat{\boldsymbol{\beta}}$ to possess the oracle properties. However, restrictive constraints must be imposed on $\mathbf{d}$. Specifically, for the model selection result we require the components in $\mathbf{d}$ satisfy $\min_{j \geq k_n+1} d_j > 0$, and the norm $\|\mathbf{d}_0\|_2$ must not diverge too quickly. The extra bias term, $n\lambda_n \mathbf{V}_n \tilde{\mathbf{d}}_0$, is in fact caused by the penalty function. Although these types of conditions are standard for high-dimensional oracle results, they are somewhat restrictive, as Fan et al. (2014) have pointed out in the context of quantile regression estimation. In this section, we examine a procedure to deal with overestimation.

To deal with overestimation in high-dimensional settings, two-step procedures have been implemented; see, for instance, Bradic et al. (2011), Fan et al. (2014), and Wang et al. (2012). These authors proposed two-step procedures in the context of quantile regression estimation, while keeping the convex optimization problem intact. The idea is to first obtain an initial robust estimate of $\boldsymbol{\beta}^*$ and then to compute the weight vector $\mathbf{d}$ of the weighted l_1-penalty using a concave penalty function. In the second step, the penalized quantile regression is implemented with the computed weighted l_1-penalty. Another novel two-step procedure is proposed in Loh (2017) for symmetric loss functions.

Following their ideas, we propose the following two-step procedure for the bias reduction of our regularized robust estimator defined in the previous section. We first obtain an initial robust estimator of $\boldsymbol{\beta}^*$ and label it $\hat{\boldsymbol{\beta}}^*$. Next we let $\hat{\mathbf{d}} = (\hat{d}_1,\ \hat{d}_2, \ldots, \hat{d}_{p_n})^T$ with

$\hat{d}_j = p^{(1)}_{\lambda_n}(|\hat{\beta}^*_j|)$, where $p_{\lambda_n}(\cdot)$ is a folded-concave penalty function. Then we replace the weight vector $\mathbf{d}$ in (2.4) with $\hat{\mathbf{d}}$ and solve the regularization problem by minimizing the resulting objective function:

$$\hat{Q}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) + n\lambda_n \|\hat{\mathbf{d}} \circ \boldsymbol{\beta}\|_1. \tag{2.6}$$

We refer to the estimator obtained by minimizing (2.6) as the *generalized adaptive robust Lasso* (GAR-Lasso) *estimator.*

Let vector $\mathbf{d}^* = \left(d^*_1, \ldots, d^*_{p_n}\right)^T$, where $d^*_j = p^{(1)}_{\lambda_n}\left(|\beta^*_j|\right)$. We impose the following condition on the penalty function so that $\hat{\mathbf{d}}_0 = (\hat{d}_1,\ \hat{d}_2, \ldots, \hat{d}_{k_n})^T$ is close to $\mathbf{d}^*_0 = \left(d^*_1, \ldots, d^*_{k_n}\right)^T$ under the l_2-norm:

(C5) $p^{(1)}_{\lambda_n}(\cdot)$ is nonincreasing on $(0, \infty)$ satisfying $p^{(1)}_{\lambda_n}(c^*_1\sqrt{k_n \log(p_n)/n}) > \frac{1}{2}p^{(1)}_{\lambda_n}(0+)$ for sufficiently large n and some $c^*_1 > 0$, and the Lipschitz condition holds, i.e., there exists some constant $M > 0$ such that

$$\left|p^{(1)}_{\lambda_n}(t_1) - p^{(1)}_{\lambda_n}(t_2)\right| \leq M|t_1 - t_2|,$$

for any t_1, $t_2 \in \mathbb{R}$.

A well-known folded-concave penalty function is the SCAD penalty (Fan and Li 2001), and its first derivative $p^{(1)}_{\lambda_n}$ is

$$p^{(1)}_{\lambda_n}(t) = \lambda_n\{\mathbb{I}\left(t \leq \lambda_n\right) + \frac{(a\lambda_n - t)_+}{(a-1)\lambda_n}\mathbb{I}\left(t > \lambda_n\right)\}$$

for some $a > 2$ and $t > 0$, where z_+ stands for the positive part of z. It can be shown that condition (C5) is satisfied by the SCAD penalty function provided $\lambda_n > 2(a + 1)^{-1}c\sqrt{k_n(\log p_n)/n}$ for some constant $c > 0$. When the stochastic weights $\hat{d}_j$ are defined using the SCAD penalty function, the resulting $\hat{d}_j$ will be close, or even equal, to zero for $1 \leq j \leq k_n$, and hence the magnitude of the weight vector $\|\hat{\mathbf{d}}_0\|_2$ will be close to zero.

Theorem 2.4. *Assume that condition* (C5) *holds. In addition,* $\min_{j \geq k_n+1} d^*_j$ *is strictly positive. Suppose that the initial estimate* $\hat{\boldsymbol{\beta}}^*$ *satisfies* $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\|_2 \leq c^*_1\sqrt{k_n \log p_n/n}$ *where*

$c_1^* > 0$ *is defined in* (C5). *In addition, assume that conditions of Theorem* 2.2 *hold with* $\mathbf{d} = \mathbf{d}^*$ *and* $s_n = b_n$ *where* b_n *is defined as*

$$b_n = c_2^* \left(\sqrt{k_n \log n/n} + \lambda_n \left(\|\mathbf{d}_0^*\|_2 + c_3^* c_1^* \sqrt{k_n \log p_n/n} \right) \right)$$

with some positive constants c_2^* *and* c_3^*. *Then, with probability tending to one, there exists a global minimizer* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \, \hat{\boldsymbol{\beta}}_2^T \right)^T$ *of* $\hat{Q}_n(\boldsymbol{\beta})$ *defined by* (2.6) *such that* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *and* $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq b_n$.

The results in Theorem 2.4 are similar to those in Theorem 2.2. Theorem 2.4 shows that the regularized robust estimator $\hat{\boldsymbol{\beta}}$ possesses the model selection oracle property, and $\hat{\boldsymbol{\beta}}_1$ converges to $\boldsymbol{\beta}_1^*$ with a rate of convergence of b_n. In comparison with the results in Theorem 2.2, the rate of consistency b_n contains an extra term of the order of $\lambda_n \sqrt{k_n (\log p_n)/n}$. This is in fact caused by the bias of the initial estimate $\hat{\boldsymbol{\beta}}^*$, and it seems to be the price to pay for using adaptive penalty weights (Fan et al. 2014). However, since $\lambda_n \to 0$ as $n \to \infty$, the second term in b_n will be negligible. An initial robust estimator $\hat{\boldsymbol{\beta}}^*$ can easily be constructed using the objective function (2.4) with the loss function ϕ as the Huber function, the weight function $W(\mathbf{x}) = 1$, the penalty weight vector $\mathbf{d}_0 = (1, ..., 1)^T$, and the tuning parameter $\gamma = 1$. Then, it can be shown that the condition on the initial estimator is satisfied, i.e., $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\|_2 = O\left(\sqrt{k_n (\log p_n)/n} \right)$. Other possible choices for $\hat{\boldsymbol{\beta}}^*$ having the preceding consistency property are high-dimensional quantile regression (Belloni and Chernozhukov 2011) and LAD regression (Wang 2013) estimators.

2.4 Simulation studies

In this section, we evaluate the finite-sample performance of our method and compare it with some popular existing methods. We used the high-dimensional linear regression model studied in Bradic et al. (2011) for our simulation. That is, we simulated data from the regression model $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i$ with sample size $n = 100$ and $p_n = 400$. The true

regression coefficient vector was set as

$$\boldsymbol{\beta} = (2, 0, 1.5, 0, 0.80, 0, 0, 1, 0, 1.75, 0, 0, 0.75, 0, 0, 0.3, 0, \ldots, 0)\,.$$

We considered six symmetric and asymmetric error distributions for e_i:

(a) A normal error with mean 0 and variance 4, $N(0, 4)$;

(b) Twice the t-distribution with 3 degrees of freedom, denoted $2t_3$;

(c) A mixture of normal distributions, MN, $0.6N(1, 1) + 0.4N(3, 25)$;

(d) A Laplace distribution with location parameter 1 and scale parameter 3;

(e) A log-normal distribution, LogNormal, defined as $e = \exp(1 + 1.2Z)$ with Z having the standard normal distribution;

(f) A Weibull distribution with shape parameter 0.3 and scale parameter 1.

Table 2.1 summarizes the six error scenarios according to the shapes and tails of the error distributions. For each distribution listed above, we generated the predictor variables $\mathbf{X}_i$ from two settings:

(I) $\mathbf{X}_i$'s follow a multivariate normal distribution $N(0, \Omega_1)$ with covariance matrix $\Omega_1 = 0.5^{|i-j|}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$, where $n = 100$ and $p_n = 400$;

(II) $\mathbf{X}_i$'s follow a mixture normal distribution $0.8N(0, \Omega_2) + 0.2N(\mu, \Omega_1)$, where $\Omega_2 = I_{p_n \times p_n}$, $\mu = 3\mathbf{1}_{p_n}$, Ω_1 is as in setting I, and $p_n = 400$. This setting produces covariates with some influential points.

We assessed the finite-sample performance of our GAR-Lasso estimator with the following specifications. We minimized (2.6) by setting ϕ as the Huber function ρ_α with $\alpha = 1.345$, and the weight function $W(\mathbf{x}) = \min\{1, 1/(\|\mathbf{x}\|_\infty)^3\}$, which satisfies condition (C3). The loss function tuning parameter γ in (2.6) was adaptively chosen to minimize

the asymptotic variance. Further, to obtain an optimal value for λ_n, we implemented a five-fold cross-validation. Adaptive weights on the penalty function were computed based on the SCAD penalty. Finally, we iteratively applied the gradient descent algorithm to obtain our estimator.

To evaluate the performance of our estimator and others, we used the following six performance measures:

(1) l_2 loss, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$;

(2) l_1 loss, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$;

(3) FP: the number of false positives, i.e., the number of noises included in the model;

(4) FN: the number of false negatives, i.e., the number of signal covariates that are not included;

(5) SE_1: the standard error of l_1 loss;

(6) SE_2: the standard error of l_2 loss.

In our simulation, we compared our estimator with the following four methods:

(i) *Lasso*, the penalized least squares estimator with the l_1-penalty;

(ii) *R-Lasso*, the regularized LAD estimator with the l_1-penalty as in Wang (2013), which is the same as the R-Lasso estimator in Fan et al. (2014);

(iii) *Huber-Lasso*, the Huber function with $\alpha = \text{IQR}(y)/10$ plus the l_1-penalty, where IQR stands for the inter-quartile range;

(iv) *Adaptive Robust Lasso* (AR-Lasso), the AR-Lasso estimator of Fan et al. (2014).

Tables 2.2 and 2.3 summarize our simulation results quantified by measures (1) to (6) according to the shapes and tails of the error distributions (a) to (f), under settings I

and II, respectively. The results are the average of each performance measure over 200 repetitions.

From the results in Tables 2.2 and 2.3, we make the following observations. First, the quantile and Huber-loss-based estimators were more robust in dealing with outliers. When the errors were asymmetric and heavy-tailed, the performance of Lasso deteriorated. For instance, with the Weibull distribution, which is asymmetric and heavy-tailed, the performance of Lasso was extremely unstable: the average l_1 and l_2 losses were significantly larger than those of the other methods. GAR-Lasso had a clear advantage over Lasso, R-Lasso, and Huber-Lasso: it performed better for all the cases considered, particularly heavy-tailed distributions.

We now compare GAR-Lasso with AR-Lasso. In Table 2.2, we observe that GAR-Lasso and AR-Lasso performed similarly when the errors followed distributions with light and symmetric tails. However, GAR-Lasso generally performed better than AR-Lasso in terms of l_2 loss under almost all settings. It is worth noting that GAR-Lasso began to excel AR-Lasso when the tails of errors became heavier. A similar phenomenon is observed in Table 2.3. It is evident that under setting II, where outliers are more likely to occur in the covariates, GAR-Lasso performed better than AR-Lasso in terms of l_1 and l_2 losses when the errors were distributed with heavy and asymmetric tails. This demonstrates the superiority of GAR-Lasso when dealing with errors having heavy-tailed distributions.

Then we compare the estimators in terms of the false positives (FP) and false negatives (FN). The AR-Lasso and GAR-Lasso estimators consistently selected fewer false positives and false negatives than other estimators in both settings I and II. For example, in setting I with $N(0, 4)$ errors, Lasso, R-Lasso, and Huber-Lasso had 18.34, 25.6, and 22.95 false positives, respectively, while AR-Lasso and GAR-Lasso only had 20.52 and 18.83, accordingly. In general, GAR-Lasso selected fewer false positives than other estimators under setting I and setting II. In addition, we observe that for the errors having logNormal or Weibull distributions, GAR-Lasso performed uniformly better than AR-Lasso under both

settings.

For the errors having heavy-tailed distributions, GAR-Lasso tended to yield larger values for the tuning parameter γ in setting II than in setting I. This indicates that γ automatically adapts to errors with different shapes and provides greater flexibility for the handling of high leverage points.

To further examine the performance of these estimators, we also ran a simulation study based on a heteroscedastic model given by

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \left(\sqrt{3}\|\boldsymbol{\beta}\|_2^2\right)^{-1}\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)e_i,\ i = 1, \ldots, n, \tag{2.7}$$

with $n = 100$ and $p_n = 400$. In model (2.7), $\mathbf{X}_i$'s were generated from the same setups as in the original model $Y_i = \mathbf{X}_i^T\boldsymbol{\beta}+e_i$. For model (2.7), the performance of each estimator is reported in Tables 2.4 and 2.5. The results are again based on 200 repetitions.

The results in Tables 2.4 and 2.5 show that methods using adaptive weights outperformed other estimating procedures under all the scenarios considered. The two adaptive procedures, namely AR-Lasso and GAR-Lasso, again exhibited superior performance overall when compared with others in terms of all six measures studied in this simulation study. In particular, we observe that two adaptive procedures consistently selected a smaller set of variables than other non-adaptive procedures. In addition, two adaptive procedures produced estimators with smaller standard errors when compared with other procedures. The values in Tables 2.4 and 2.5 indicate that GAR-Lasso attained the lowest values of l_1 and l_2 losses in most cases of error distributions. Indeed, GAR-Lasso had the best performance in Table 2.5 in terms of l_1 and l_2 losses. Furthermore, GAR-Lasso and AR-Lasso had comparable standard errors. However, GAR-Lasso outperformed AR-Lasso in terms of all six measures when error distributions were heavy-tailed. By comparing Tables 2.2 and 2.3 with Tables 2.4 and 2.5, we find that the advantage of GAR-Lasso procedure over others was more pronounced in the heteroscedastic model than in the homoscedastic model. Furthermore, GAR-Lasso seemed to have the biggest advantage overall when errors were asymmetric and heavy-tailed (LogNormal and Weibull).

In summary, GAR-Lasso is more flexible and efficient than Lasso, R-Lasso, and Huber-Lasso when dealing with errors with different shapes and tails. It is competitive with AR-Lasso for errors with light tails and has a clear advantage over others for heavy-tailed error distributions. It has good efficiency for clean-data situations and exhibits excellent robustness to outliers in a range of situations where outliers are likely to be present in either the response or covariates or both. In addition, GAR-Lasso selects fewer false positives and false negatives in most cases.

Table 2.1: Summary of the shapes and tails of error distributions.

	Light tails	**Heavy tails**
Symmetric	N(0,4) Laplace	$2t_3$
Asymmetric	MN	LogNormal Weibull

2.5 Real data application

Atherosclerosis is a disease of inflammation characterized by interactions among platelets, leukocytes, and endothelial cells; see, e.g., Libby (2002). Atherosclerosis is the leading cause of cardiovascular disease (CVD). By narrowing the coronary arteries responsible for bringing oxygenated blood to the heart, it can result in symptoms such as sweating, nausea, dizziness, or breathlessness and possibly lead to a cardiac arrest.

There have been various studies of atherosclerosis at the genetic level. Huang et al. (2011) studied the role of the innate immune system on the development of atherosclerosis by examining gene profiles from the peripheral blood of 119 patients. There were 119 subjects and 494 variables. The data were collected using an Illumina HumanRef8 V2.0 Bead Chip and are available from the gene expression omnibus. To test the influence of gene expression profiles on atherosclerosis, Huang et al. (2011) conducted a microarray

Table 2.2: Simulation results of the estimators under setting I.

Scenario		**Simulation results**				
		Lasso	R-Lasso	Huber-Lasso	AR-Lasso	GAR-Lasso
N(0,4)	L_2 loss	1.326	1.589	1.407	1.279	1.284
	L_1 loss	5.409	5.646	5.448	4.518	4.776
	FP,FN	18.34, 0.76	25.60, 0.92	22.95, 0.90	20.52, 0.92	18.83, 0.93
	SE_1,SE_2	0.423, 0.071	0.477, 0.108	0.411, 0.072	0.144, 0.027	0.144, 0.025
	γ	-	-	-	-	0.544
$2t_3$	l_2 loss	1.286	1.269	1.299	1.205	1.201
	l_1 loss	5.523	4.373	4.582	3.645	3.802
	FP,FN	28.79, 0.76	32.36, 0.69	23.57, 0.71	20.66, 0.60	18.75, 0.87
	SE_1,SE_2	0.723, 0.238	0.701, 0.231	0.683, 0.271	0.155, 0.061	0.187, 0.058
	γ	-	-	-	-	0.263
MN	l_2 loss	2.050	1.689	1.762	1.349	1.333
	l_1 loss	7.869	7.214	7.286	4.139	4.818
	FP,FN	19.57, 1.47	27.70, 1.23	26.74, 1.40	15.22, 1.01	16.50, 1.55
	SE_1,SE_2	0.722, 0.210	0.668, 0.177	0.574, 0.180	0.192, 0.051	0.184, 0.048
	γ	-	-	-	-	0.586
Laplace	l_2 loss	2.425	2.237	2.222	2.092	2.078
	l_1 loss	9.774	6.482	6.757	5.427	5.371
	FP,FN	23.82, 2.28	17.52, 2.13	17.59, 2.37	13.86, 1.35	15.11, 1.58
	SE_1,SE_2	0.616, 0.278	0.871, 0.291	0.799, 0.284	0.179, 0.088	0.162, 0.076
	γ	-	-	-	-	0.710
LogNormal	l_2 loss	2.955	2.312	2.319	2.185	2.012
	l_1 loss	7.979	7.227	7.111	6.167	5.812
	FP,FN	25.21, 3.46	17.52, 2.42	15.13, 2.14	10.93, 1.67	10.11, 1.25
	SE_1,SE_2	1.105, 0.988	0.986, 0.572	1.072, 0.471	0.174, 0.050	0.168, 0.048
	γ	-	-	-	-	1.171
Weibull	l_2 loss	13.547	1.926	1.633	1.610	1.515
	l_1 loss	81.281	8.929	6.430	4.895	4.221
	FP,FN	26.7, 4.66	17.46, 1.30	16.79, 1.24	9.59, 1.83	8.54, 1.38
	SE_1,SE_2	1.201, 0.890	0.538, 0.341	0.414, 0.266	0.207, 0.101	0.196, 0.071
	γ	-	-	-	-	0.474

gene expression profiling of whole blood from a population of healthy women with low to intermediate Framingham risk scores (FRSs). In addition, they identified genes and pathways that are closely connected with a significant burden of atherosclerosis among the women predicted to be at a lower risk of CVD events. They demonstrated that the toll-like receptor (TLR) signaling that pathway plays a crucial role in triggering the innate immune system when atherosclerosis invades. Furthermore, they found that the gene

Table 2.3: Simulation results of the estimators under setting II.

Scenario		**Simulation results**				
		Lasso	R-Lasso	Huber-Lasso	AR-Lasso	GAR-Lasso
N(0,4)	l_2 loss	1.425	1.415	1.369	1.281	1.267
	l_1 loss	4.649	5.370	5.357	4.519	4.485
	FP,FN	18.93, 0.71	29.10, 0.76	27.10, 0.50	14.13, 0.52	16.67, 0.62
	SE_1,SE_2	0.372, 0.051	0.311, 0.048	0.324, 0.041	0.168, 0.019	0.142, 0.023
	γ	-	-	-	-	0.569
$2t_3$	l_2 loss	1.493	1.171	1.095	1.042	1.041
	l_1 loss	4.776	4.309	4.290	3.683	3.743
	FP,FN	19.42, 0.78	32.90, 0.68	23.54, 0.62	21.71, 0.59	19.09, 0.65
	SE_1,SE_2	0.217, 0.192	0.200, 0.102	0.203, 0.086	0.177, 0.056	0.154, 0.046
	γ	-	-	-	-	0.548
MN	l_2 loss	1.984	1.803	1.566	1.246	1.254
	l_1 loss	6.490	5.747	5.300	4.842	4.851
	FP,FN	12.93, 1.73	16.20, 2.04	15.43, 1.73	15.02, 0.80	14.75, 1.04
	SE_1,SE_2	1.107, 0.412	1.109, 0.399	0.781, 0.377	0.133, 0.031	0.137, 0.032
	γ	-	-	-	-	0.705
Laplace	l_2 loss	2.028	2.041	2.131	1.892	1.900
	l_1 loss	8.181	8.716	8.416	7.321	7.035
	FP,FN	23.30, 1.39	22.09, 2.10	21.95, 1.85	19.48, 1.29	19.50, 1.20
	SE_1,SE_2	0.511, 0.168	0.446, 0.176	0.478, 0.154	0.134, 0.043	0.126, 0.048
	γ	-	-	-	-	0.859
LogNormal	l_2 loss	3.430	2.114	2.140	2.087	2.066
	l_1 loss	17.132	8.511	8.127	6.498	6.266
	FP,FN	22.8, 5.38	30.48, 1.92	18.14, 2.16	13.70, 2.60	11.29, 3.35
	SE_1,SE_2	1.728, 0.776,	1.288, 0.480	1.310, 0.482	0.139, 0.093	0.138, 0.069
	γ	-	-	-	-	1.514
Weibull	l_2 loss	16.816	2.217	2.209	1.948	1.770
	l_1 loss	91.095	8.332	8.488	6.764	6.014
	FP,FN	25.53, 4.49	37.46, 4.96	21.12, 1.18	15.75, 1.82	14.75, 1.70
	SE_1,SE_2	3.199, 2.922	0.306, 0.135	0.322, 0.117	0.201, 0.066	0.167, 0.064
	γ	-	-	-	-	1.135

TLR8 is closely associated with atherosclerosis.

Fan et al. (2017) carried out a data analysis of this microarray dataset. To further examine the relationship between the crucial gene TLR8 and the other genes, they regressed it on 464 genes from 12 pathways (TLR, CCC, CIR, IFNG, MAPK, RAPO, EXAPO, INAPO, DRS, NOD, EPO, and CTR) that are related to the TLR pathway. Their RA-Lasso method found that 34 genes are associated with TLR8. They also noted

Table 2.4: Simulation results of estimators under setting I for heteroscedastic model.

Scenario		**Simulation results**				
		Lasso	R-Lasso	Huber-Lasso	AR-Lasso	GAR-Lasso
N(0,4)	l_2 loss	0.510	0.504	0.506	0.266	0.266
	l_1 loss	1.425	1.516	1.403	0.812	0.769
	FP,FN	24.56, 1.13	21.25, 1.05	24.36, 0.95	12.67, 0.09	8.68, 0.07
	SE_1,SE_2	0.048, 0.021	0.040, 0.019	0.045, 0.019	0.019, 0.007	0.018, 0.007
	γ	-	-	-	-	0.863
$2t_3$	l_2 loss	0.624	0.606	0.617	0.325	0.295
	l_1 loss	1.580	1.859	1.688	1.023	0.946
	FP,FN	26.72, 1.33	25.23, 1.05	27.43, 0.85	12.63, 0.61	10.45, 0.60
	SE_1,SE_2	0.058, 0.015	0.067, 0.028	0.073, 0.015	0.027, 0.008	0.024, 0.007
	γ	-	-	-	-	1.163
MN	l_2 loss	1.305	0.434	0.397	0.184	0.204
	l_1 loss	5.348	1.148	1.207	0.513	0.531
	FP,FN	33.40, 1.70	28.52, 1.73	28.92, 1.62	11.85, 0.47	10.54, 0.26
	SE_1,SE_2	0.174, 0.042	0.101, 0.041	0.087, 0.031	0.019, 0.007	0.019, 0.006
	γ	-	-	-	-	1.042
Laplace	l_2 loss	0.614	0.934	0.665	0.346	0.342
	l_1 loss	1.928	2.333	1.629	1.057	0.996
	FP,FN	25.39, 1.45	22.17, 1.38	22.02, 1.28	11.63, 0.70	8.68, 0.62
	SE_1,SE_2	0.076, 0.057	0.167, 0.081	0.102, 0.061	0.041, 0.014	0.038, 0.012
	γ	-	-	-	-	1.278
LogNormal	l_2 loss	2.209	0.680	0.665	0.302	0.288
	l_1 loss	4.221	1.877	1.983	0.718	0.642
	FP,FN	35.48, 2.76	26.03, 1.78	26.52, 1.64	14.14, 0.97	11.52, 0.75
	SE_1,SE_2	0.172, 0.135	0.079, 0.032	0.089, 0.035	0.022, 0.009	0.021, 0.008
	γ	-	-	-	-	1.543
Weibull	l_2 loss	1.254	0.495	0.506	0.189	0.176
	l_1 loss	3.027	1.642	1.511	0.484	0.459
	FP,FN	21.16, 1.53	18.54, 1.26	19.11, 1.27	8.18, 0.40	6.67, 0.42
	SE_1,SE_2	0.082, 0.026	0.064, 0.014	0.092, 0.029	0.017, 0.006	0.015, 0.006
	γ	-	-	-	-	0.688

that Lasso and R-Lasso found 1 and 17 associated genes, respectively.

We employ the same microarray dataset to illustrate the performance of three estimators: Huber-Lasso, AR-Lasso, and GAR-Lasso. Our aim is also to find genes connected with TLR8, which is believed to play a significant role in the development of atherosclerosis. We used five-fold cross-validation to choose the tuning parameters for the three methods. We then applied the three methods to select significant genes from the 464

Table 2.5: Simulation results of estimators under setting II for heteroscedastic model.

Scenario		**Simulation results**				
		Lasso	R-Lasso	Huber-Lasso	AR-Lasso	GAR-Lasso
N(0,4)	l_2 loss	0.814	0.872	0.816	0.421	0.408
	l_1 loss	2.975	2.224	2.414	1.292	1.189
	FP,FN	25.41, 1.23	26.26, 1.34	25.31, 1.09	11.32, 0.87	9.85, 0.85
	SE_1,SE_2	0.097, 0.059	0.099, 0.072	0.063, 0.053	0.037, 0.011	0.038, 0.013
	γ	-	-	-	-	1.765
$2t_3$	l_2 loss	1.114	1.291	1.208	0.490	0.478
	l_1 loss	4.417	3.579	3.928	1.573	1.435
	FP,FN	26.56, 1.61	26.67, 1.47	28.96, 2.08	15.03, 1.03	11.55, 0.92
	SE_1,SE_2	0.163, 0.052	0.196, 0.102	0.173, 0.063	0.054, 0.017	0.049, 0.016
	γ	-	-	-	-	1.462
MN	l_2 loss	1.014	0.859	0.686	0.325	0.314
	l_1 loss	4.272	1.909	2.080	1.008	0.894
	FP,FN	19.64, 1.47	14.17, 1.05	15.05, 1.12	11.43, 0.75	10.14, 0.68
	SE_1,SE_2	0.164, 0.054	0.189, 0.074	0.099, 0.056	0.051, 0.014	0.042, 0.013
	γ	-	-	-	-	1.676
Laplace	l_2 loss	1.313	1.385	1.025	0.553	0.516
	l_1 loss	5.295	3.178	2.606	1.817	1.548
	FP,FN	18.97, 1.95	15.19, 1.56	13.45, 1.85	8.12, 0.80	6.36, 0.98
	SE_1,SE_2	0.175, 0.053	0.171, 0.054	0.154, 0.058	0.079, 0.022	0.062, 0.019
	γ	-	-	-	-	1.116
LogNormal	l_2 loss	3.262	0.938	0.976	0.321	0.315
	l_1 loss	5.354	1.867	2.850	0.843	0.779
	FP,FN	37.16, 3.17	22.23, 1.65	19.60, 1.81	12.95, 1.07	9.533, 0.82
	SE_1,SE_2	0.219, 0.146	0.079, 0.043	0.082, 0.036	0.037, 0.011	0.029, 0.009
	γ	-	-	-	-	2.764
Weibull	l_2 loss	1.935	0.519	0.712	0.271	0.228
	l_1 loss	5.185	1.5718	1.690	0.760	0.615
	FP,FN	43.13, 3.93	20.44, 2.02	19.60, 2.30	8.30, 1.19	5.68, 1.05
	SE_1,SE_2	0.202, 0.076	0.065, 0.020	0.062, 0.025	0.037, 0.012	0.028, 0.010
	γ	-	-	-	-	1.368

samples. Table 2.6 reports the results: it displays the genes selected by the three methods. We observe that Huber-Lasso, AR-Lasso, and GAR-Lasso selected 33, 65, and 75 genes, respectively. Therefore, the adaptive methods (AR-Lasso and GAR-Lasso) selected more genes than the non-adaptive method.

As in the preceding article, we further compared the methods as follows. We randomly selected 24 patients as the test set, and then we applied Huber-Lasso, AR-Lasso, and

GAR-Lasso to the remaining data to obtain estimates of the coefficients. We then used them to predict the responses of the test set. We repeated this process 100 times and then obtained the average of the mean squared errors and mean absolute errors. Table 2.7 summarizes our results. GAR-Lasso performed the best, obtaining the lowest values for the mean squared and mean absolute errors.

Table 2.6: Genes selected by Huber-Lasso, AR-Lasso and GAR-Lasso.

Summary of three methods				
Huber Lasso	**AR-Lasso**		**GAR-Lasso**	
CR2: 0.1005	SERPINF2: 0.0263	RIPK1: -0.0392	PLG: 0.0416	CD86: -0.0677
IL2: 0.0364	PLAUR: -0.0005	TLR1: 0.0692	PLAU: -0.0216	ARHGAP10: -0.0312
IFNG: 0.0028	KNG1: -0.0061	RELA: 0.0326	KLKB1: -0.0101	DAPK2: -0.0194
CSF3: 0.0079	CR2: 0.0720	PIK3R1: -0.1332	F7: 0.0185	PSMD6: 0.0109
SPI1: 0.0498	CPB2: 0.0064	PIK3CG: -0.0563	CR1: 0.0496	PSME1: -0.0022
IRF4: -0.0585	C8G: -0.0307	IL12A: 0.0007	C9: 0.0132	PSMD3: -0.0157
IFI6: 0.01065	C5AR1: 0.0539	CD40: -0.0588	C7: -0.0274	PSMD2: -0.0908
MAP2K7: -0.0109	BDKRB1: 0.0158	CASP8: -0.0030	C3AR1: 0.0803	PSMC2: 0.0270
PRKCH: -0.0318	TNF: -0.0031	PSME3: 0.0187	SERPING1: -0.0131	PSMC1: 0.0844
PPP3CB: -0.0452	IL2: 0.0598	PSMD6: -0.0011	SERPINC1: 0.0005	PSMB3: -0.0161
NF1: -0.0538	SPI1: 0.0336	PSMD4: -0.0026	IL1B: 0.0218	ACIN1: 0.0607
MAP3K4: -0.0606	REG1A: 0.0356	PSMD3: -0.0639	REG1A: 0.0239	STK24: -0.0045
DUSP6: -0.0161	CXCL9: -0.0455	PSMC3: 0.0516	PTPN11: 0.0521	VIM: 0.0943
AKT1: 0.0110	IRF4: -0.0636	PSMC2: 0.0880	CIITA: -0.0629	ROCK1: -0.0078
RIPK1: -0.0797	IRF8: -0.0566	PSMB4: -0.0025	IRF2: -0.0536	HMGB2: 0.0404
TLR1: 0.1135	IFI6: 0.0201	CLSPN: 0.0618	ICAM1: -0.1116	HIST1H1C: -0.0226
RELA: 0.0156	AKT3: 0.0165	STK24: 0.0742	CYBB: 0.0271	GAS2: -0.0114
PIK3R1: -0.0349	TRAF2: -0.0093	KPNB1: 0.0254	MAP4K4: 0.0152	DSG1: -0.0261
CD40: -0.0001	SRF: -0.0044	HIST1H1D: -0.0106	TRAF2: -0.0156	DFFA: -0.0359
PSMD6: -0.0059	RASA1: -0.0119	DSG2: -0.0232	TP53: -0.0145	CASP7: -0.0429
PSMD3: -0.0125	MAP2K7: -0.1027	DFFB: -0.0175	RASA1: 0.0003	YWHAB: -0.2035
PSMC2: 0.0256	PRKCH: -0.0831	DFFA: 0.0042	RAP1A: -0.0635	TFDP1: -0.0170
DCC: -0.0220	PPP5C: -0.0265		MAP2K6: -0.1062	BAD: 0.0711
CLSPN: 0.0570	PPP3CB: -0.0696		PRKCH: 0.0044	APAF1: -0.0010
CDH1: -0.0247	PPP3CA: -0.0103		PRKCG: -0.1099	RIPK2: 0.0200
CASP6: -0.0279	NF1: -0.0213		PPP3CC: -0.0386	EPO: 0.0128
BCL2L11: -0.1683	MAP3K4: -0.1208		PPP3CA: -0.0575	ICOS: 0.0062
BAX: 0.0369	MEF2C: 0.0085		PPM1B: -0.0106	WAS: -0.0346
BAK1: 0.0365	GCK: -0.0691		MYC: -0.0236	CREB1: -0.0143
AIM2: 0.0309	DUSP6: -0.0212		MEF2C: -0.1256	CDC42: -0.0692
LHB: -0.01852	DUSP4: -0.0546		MAX: 0.0134	CD3E: -0.0007
CREB1: -0.0337	AKT1: 0.0714		HSPA5: 0.0002	CD3D: 0.0063
CD3E: 0.0318	FADD: 0.0665		FOS: -0.0798	
	CD3E: 0.0092		DUSP5: -0.0264	
	CREB1: -0.0761		DUSP1: -0.0922	
	ZAP70: -0.0494		ACVR1B: 0.0752	
	EPOR: 0.0164		RIPK1: 0.0781	
	AIM2: 0.0027		PIK3R3: -0.0072	
	BAK1: 0.0698		SPP1: 0.0447	
	YWHAB: -0.0278		MAP2K3: 0.0155	
	BCL2L11: -0.2199		PIK3CG: -0.1612	
	CASP6: -0.0051		PIK3CD: -0.0703	
	CDH1: -0.0382		IL8: 0.0172	

Table 2.7: Prediction errors of Huber-Lasso, AR-Lasso and GAR-Lasso.

Summary of three methods				
		Huber Lasso	**AR-Lasso**	**GAR-Lasso**
Mean Squared Error	Mean	6.286	4.207	4.071
	Median	2.273	2.008	2.137
Mean Absolute Error	Mean	0.774	0.682	0.652
	Median	0.702	0.601	0.560

2.6 Conclusions

In this chapter, we investigate the regularized high-dimensional robust estimation for sparse linear regression. Specifically, we develop regularized estimators that are highly robust with respect to outliers in both the responses and covariates under two forms of the penalty function. The proposed estimators are also robust against heavy-tailed error distributions. Our estimators are generalized and adaptive in the sense that they downweight both the influence and leverage observations adaptively using a data-driven method in practical applications. The proposed estimators satisfy the oracle properties with a near-optimal rate of consistency under mild assumptions on the error distribution and covariates. We establish the asymptotic results in a high-dimensional setting. Our numerical studies and data analysis show that the finite-sample performance of our doubly adaptive two-step procedure, GAR-Lasso, is very promising. The numerical results also demonstrate that the performance of GAR-Lasso is satisfactory even for heteroscedastic linear models with asymmetric noise distributions. It appears that GAR-Lasso is fairly adaptive to various shapes and tails of the error distribution. Further, the simulations show that the leverage points can make a significant impact on the performance of regularized estimators. Thus, protection against such outliers is highly recommended.

Chapter 3

High-dimensional robust regression estimation with l_q-loss function

3.1 Introduction

High-dimensional data are now common in studies in many areas, including medicine, biomedical, machine learning, bioinformatics, astronomy, etc. The analysis of high-dimensional data is extremely challenging. In particular, separating the useful information from the noise is generally difficult in high-dimensional settings. There has been a considerable development of statistical methods for analyzing high-dimensional data, and regularization methods are among the most widely used tools for analyzing such data. Variable selection is fundamentally important for knowledge discovery with fixed- and high-dimensional data and it could greatly enhance the prediction performance of the fitted model. Regularization techniques play an important role in identifying covariates that truly affect the outcome of a response in models containing covariates and a response variable. A vast amount of research has been done in this area, and some robust procedures have also been developed. In this chapter, we still consider the following the situation where we have a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the linear model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i, \ \ i = 1, ..., n, \tag{3.1}$$

where $Y_i \in \mathbb{R}$ is a response variable, $\mathbf{X}_i \in \mathbb{R}^{p_n}$ is a covariate vector, $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ is a regression coefficient vector, and $e_i \in \mathbb{R}$ is an error term. We assume that $\{e_i\}_{i=1}^n$ are independent

and identically distributed (i.i.d.) with some unknown distribution $\mathbb{F}$. We also assume that $\{\mathbf{X}_i\}_{i=1}^n$ are i.i.d. random vectors and that the e_i's are independent of the $\mathbf{X}_i$'s. We assume that the distributions of $\mathbf{X}_i$ and e_i both have mean zero. We are interested in high-dimensional robust estimation of the p_n-dimensional coefficient vector $\boldsymbol{\beta}$ where p_n may increase with n and $\boldsymbol{\beta}$ is sparse in the sense that many of the elements are zero.

Robustness has been a core issue in the statistical learning of high-dimensional problems now. Heavy-tailed errors are stylized features of high-dimensional data and errors with heavy tails result in outliers. These heavy-tailed errors impair the consistency of many high-dimensional regression methods. For robust estimation of $\boldsymbol{\beta}$ in high-dimensional settings, a number of robust regularization methods have been proposed in the literature to deal with the outliers in $(\mathbf{X}_i, Y_i)$'s as well as heavy-tailed errors. Belloni and Chernozhukov (2011) considered l_1-penalized quantile regression in a high-dimensional setting. Li et al. (2011) examined a nonconcave penalized robust M-estimator, again in a high-dimensional setting. Bradic et al. (2011) studied a penalized composite quasi-likelihood method for ultrahigh-dimensional robust variable selection. van de Geer and Müller (2012) obtained bounds on the prediction error of a large class of l_1-penalized estimators that includes quantile regression. Wang et al. (2012) considered nonconvex penalized quantile regression in an ultrahigh-dimensional setting. Wang (2013) studied an l_1-penalized LAD estimator for high-dimensional regression. In an ultrahigh-dimensional setting, Fan et al. (2014) investigated a penalized quantile regression procedure with a weighted l_1-penalty for robust regularization, as in Bradic et al. (2011). She and Chen (2017) proposed a robust reduced-rank regression approach for joint modeling and outlier detection. Fan et al. (2017) proposed an l_1-penalized procedure based on the Huber loss (Huber 1964) with diverging parameter α (or converging to zero if its reciprocal is used) in an ultrahigh-dimensional setting and obtained nonasymptotic bounds on the l_2-error. Sun et al. (2020) also dealt with the Huber loss and obtained nonasymptotic bounds on the l_2-error in high-dimensional settings. Loh (2017) studied a generalized robust M-estimator with an "amenable (nonconvex) regularizer" for high-dimensional data.

These methods essentially fall into one group: they all consider a loss function having a bounded gradient or subgradient in order to develop robust estimators. In the classical robust statistics, a loss function with a bounded derivative leads to a bounded influence function and, hence, the corresponding estimator would be stable locally to outliers in the response variable (Huber 1981). This phenomenon, however, is not a necessary condition for an estimator to be robust, as noted in Yohai and Zamar (1993) and Gervini and Yohai (2002), among others, for robust regression estimation and in Karunamuni et al. (2019) for the regularized robust regression for the fixed-dimensional case. Some classes of estimators, such as GM-estimators, have bounded influence but their breakdown points go to zero when the dimension of $\boldsymbol{\beta}$ increases. Hence, bounded influence is neither a necessary nor a sufficient condition for robust regression estimation (Gervini and Yohai 2002).

In this chapter our main interest is investigating high-dimensional robust sparse estimation under l_q-loss functions, $1 \leq q < 2$. For $1 < q < 2$, the l_q-loss functions have unbounded subgradients. However, the l_q-loss functions produce robust estimators when q is close to 1, while the estimators are more efficient when q is close to 2. In other words, q is a tuning parameter that controls the degree of robustness and efficiency. To the best of our knowledge, high-dimensional regularized estimation in model (3.1) under l_q-loss functions ($1 < q < 2$) has not been fully studied in the literature. An interesting aspect of l_q-loss functions ($1 \leq q < 2$) is that they do not satisfy *restricted strong convexity* conditions near the origin. The preceding conditions have been shown to play a critical role in the study of statistical consistency of high-dimensional M-estimators with smooth loss functions (Negahban et al. 2012). This makes high-dimensional estimation under l_q-loss functions ($1 \leq q < 2$) a challenging problem.

Due to non-smoothness of l_q-loss functions ($1 \leq q < 2$) near the origin, the corresponding regularized optimization problems are computationally challenging and expensive. To circumvent this issue, we utilize a half-quadratic approximation of convex functions (Geman and Reynolds 1992, Geman and Yang 1995) in order to accelerate computation.

The alternate optimization problem of the resultant (augmented) objective function has a simple explicit form and is easy to implement in practice.

We investigate high-dimensional robust sparse estimation under a class of convex loss functions, including the l_q-loss functions ($1 \leq q < 2$) and the Huber loss function, among others. For the regularization, we employ a weighted l_1-penalty function of the form $\lambda_n \|\mathbf{d} \circ \boldsymbol{\beta}\|_1$, where $\lambda_n > 0$ is a regularization parameter, $\|\mathbf{x}\|_1$ denotes the l_1-norm of any vector $\mathbf{x} = (x_1, ..., x_p)^T$, $\mathbf{d} = (d_1, ..., d_{p_n})^T$ is a vector of nonnegative weights, and $\circ$ denotes the Hadamard product (i.e., the componentwise product of two vectors). We establish the model selection oracle property, rate of consistency, and the asymptotic normality of our estimator in a high-dimensional setting. Our method of proofs relies on a variation of techniques developed in Fan et al. (2014) for ultrahigh-dimensional quantile regression estimation. To facilitate the proofs, we prove a new concentration type inequality similar in spirit to those in Bühlmann and van de Geer (2011) but carefully tailored for our purposes. We carry out extensive simulation studies to compare the proposed method with other regularized robust estimators, including some quantile regression and least-absolute-deviation regression estimators. Our numerical studies demonstrate the favorable finite-sample performance of the proposed procedure for various shapes and tails of the error distribution. This advantage is most pronounced in the presence of high leverage points.

The rest of this chapter is structured as: in Section 3.2 and 3.3, we study the proposed estimator using a weighted l_1-penalty function and investigate its asymptotic properties, including the model consistency and oracle properties. Section 3.4 provides the computational algorithm. In Section 3.5, we present numerical studies and compare our method with other existing methods, and in Section 3.6, we illustrate the performance of our method using a real dataset. Section 3.7 provides concluding remarks. The proofs of our main results and important lemmas are provided in Appendix B.

3.2 Robust estimation with l_q-loss

We start by introducing some useful notations. First, we let $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*T},\ \boldsymbol{\beta}_2^{*T}\right)^T$ denote the true coefficient vector of $\boldsymbol{\beta}$, where each element in $\boldsymbol{\beta}_1^* \in \mathbb{R}^{k_n}$ is nonzero and $\boldsymbol{\beta}_2^* = \mathbf{0} \in \mathbb{R}^{p_n - k_n}$. Then, for $i = 1, ..., n$, let $\mathbf{X}_i = (\mathbf{X}_{1i}^T,\ \mathbf{X}_{2i}^T)^T$, where $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ are the covariates corresponding to $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^* = \mathbf{0}$, respectively. For any weight vector $\mathbf{d}$, let $\mathbf{d}_0$ denote the first k_n elements of $\mathbf{d}$, and $\mathbf{d}_1$ denote the remaining part of $\mathbf{d}$, i.e., $\mathbf{d} = \left(\mathbf{d}_0^T,\ \mathbf{d}_1^T\right)^T$. For any vector $\mathbf{x} = (x_1, ..., x_p)^T \in \mathbb{R}^p$, let $\|\mathbf{x}\|_2$ denote the l_2-norm and $\|\mathbf{x}\|_\infty = \max(|x_1|, ..., |x_p|)$. We allow both p_n and k_n to diverge with the sample size n and assume k_n to be $o(n)$; that is, only a small number of true coefficients are nonzero. Let $\mathfrak{X} = (X_{ij}; 1 \le i \le n,\ 1 \le j \le p_n)$ denote the design matrix. We write $\mathfrak{X} = (\mathbf{S},\ \mathbf{Q})$ with the submatrices $\mathbf{S}$ and $\mathbf{Q}$ corresponding to the covariates whose coefficients are nonvanishing and vanishing, respectively. Thus, $\mathbf{S}$ and $\mathbf{Q}$ are $n \times k_n$ and $n \times (p_n - k_n)$, respectively. We will refer to the set of columns in $\mathbf{S}$ as the signal covariates, while those in $\mathbf{Q}$ will be called noise covariates. We assume without loss of generality that the X_{ij}'s are normalized, so $\mathbf{E}(X_{ij}^2) = 1$ for all i and j.

We consider regularized estimators of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) + np_{\lambda_n}(\boldsymbol{\beta}), \tag{3.2}$$

where ϕ is a symmetric convex loss function on $\mathbb{R}$ satisfying $\phi(0) = 0$, W is a non-negative weight function on large values of $\mathbf{X}_i$, $p_{\lambda_n}(\boldsymbol{\beta})$ is a regularizer (penalty function) on $\boldsymbol{\beta}$ with regularization parameter $\lambda_n > 0$, and $\gamma > 0$ is a tuning parameter. The parameter γ controls the degree of efficiency/robustness of the estimator. For the weight function $W(\mathbf{x})$, we use a non-negative function such that $x_i W(\mathbf{x})$ is bounded for all $x_i \in \mathbf{x}$. Then for large values of $\mathbf{X}_i$, $W(\mathbf{X}_i)$ would place a small weight on $\mathbf{X}_i$ and hence would have a small impact on $\hat{\boldsymbol{\beta}}$. The outliers of the covariates $\mathbf{X}_i$ are known to severely influence the performance of estimators (Huber 1981). For the penalty function $p_{\lambda_n}(\boldsymbol{\beta})$, we consider a weighted l_1-penalty function. But the results can be obtained for the adaptively weighted

l_1-penalty function case as well. We will prove the model selection oracle property and establish the asymptotic normality of $\hat{\boldsymbol{\beta}}$.

We require the following conditions for our asymptotic results.

D1. The loss function ϕ satisfies the "local δ-Lipschitz condition." That is, if $\phi^{(1)}$ denotes a subgradient of ϕ then for some $\delta \in [0,1]$ there exist positive constants ξ_q and C_q such that $\left|\phi^{(1)}(x+t) - \phi^{(1)}(x)\right| \leq C_q \left|t\right|^{\delta}$ for all $x \in \mathbb{R}$ and $|t| \leq \xi_q$. Let $\mathcal{D}$ denote the set of discontinuity points of $\phi^{(1)}$, which is the same for all choices of $\phi^{(1)}$. The common distribution function $\mathbb{F}$ of the errors satisfies $\mathbb{F}(\mathcal{D}) = 0$, $\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\right) = 0$, $\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\right)^2 = \sigma_\gamma^2 < \infty$, and as $t \to 0$, $\mathbf{E}(\phi^{(1)}\left((e_1+t)/\gamma\right) - \phi^{(1)}(e_1/\gamma))^2 \to 0$,

$$\mathbf{E}\left(\phi^{(1)}\left((e_1+t)/\gamma\right)\right) \equiv g(\gamma)t + o(|t|q(\gamma))$$

for some positive functions g and q of γ.

D2. $W : \mathbb{R}^{p_n} \to (0,1]$ is a weight function on $\mathbf{x} \in \mathbb{R}^{p_n}$ such that $\left|x_i\right|^{2+\varepsilon} W(\mathbf{x})$ is bounded for all $x_i \in \mathbf{x}$, $i = 1, ..., p_n$, where $0 < \varepsilon < 1$ is some constant.

D3. The eigenvalues of $\frac{1}{n}\mathbf{E}\left(\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)$ are bounded from below and above by some positive constants.

D4. Random variables $\phi^{(1)}\left(e_i/\gamma\right)$, $i = 1, ..., n$, are sub-Gaussian; i.e., for some constants K_0 and σ_0, $K_0^2\left(\mathbf{E}\exp(|\phi^{(1)}(e_i/\gamma)|^2/K_0^2) - 1\right) \leq \sigma_0^2$ for all i, where $\phi^{(1)}$ is defined in condition (D1).

D5. $p_{\lambda_n}^{(1)}(\cdot)$ is non-increasing on $(0,\infty)$ satisfying $p_{\lambda_n}^{(1)}(\sqrt{k_n \log(p_n)/n}) > \frac{1}{2}p_{\lambda_n}^{(1)}(0+)$ for sufficiently large n and the Lipschitz condition holds, i.e., there exists some constant $M > 0$ such that

$$\left|p_{\lambda_n}^{(1)}(t_1) - p_{\lambda_n}^{(1)}(t_2)\right| \leq M|t_1 - t_2|,$$

for any t_1, $t_2 \in \mathbb{R}$.

In addition, we define matrices $\mathbf{V}_n = \left(\mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\right)^{-1/2}$, $\mathbf{Z}_n =$

$\mathbf{SV}_n$, and $\mathbf{Z}_n^* = \left(\frac{1}{\gamma}\mathbf{X}_{11}W\left(\mathbf{X}_1/\gamma\right), \ldots, \frac{1}{\gamma}\mathbf{X}_{1n}W\left(\mathbf{X}_n/\gamma\right)\right)^T \mathbf{V}_n$, where $W(\cdot)$ is a weight function on the covariates $\mathbf{X}_i$, $i \geq 1$, and function g is defined in condition (D1).

Condition (D1) is similar to standard assumptions imposed in the classical M-estimation theory of linear regression models; see, e.g., Bai et al. (1992) and Wu (2007). Weight functions satisfying condition (D2) can be easily constructed, e.g. $W(\mathbf{x}) = \min\{1, (\|\mathbf{x}\|_\infty)^{-3}\}$. Condition (D3) warrants the inverse of eigenvalues of $\frac{1}{n}\mathbf{E}\left(\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)$ exists. A condition like (D4) is generally placed directly on the errors; see, e.g., Fan et al. (2017). Condition (D5) is a standard assumption imposed on penalty functions, e.g., Fan et al. (2014). The SCAD penalty function satisfies (D5) under some conditions on λ_n (Fan and Li 2001).

3.3 Estimator with weighted penalty

In this section, we assume that the penalty function $p_{\lambda_n}(\boldsymbol{\beta})$ is a weighted l_1-penalty of the form $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \|\mathbf{d} \circ \boldsymbol{\beta}\|_1$, where $\mathbf{d} = (d_1, ..., d_{p_n})^T$ is a vector of nonnegative weights, and $\circ$ denotes the Hadamard product (i.e., the componentwise product of two vectors). Thus, our estimator $\hat{\boldsymbol{\beta}}$ is obtained by minimizing the objective function

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^n \phi((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_i/\gamma) + n\lambda_n\|\mathbf{d} \circ \boldsymbol{\beta}\|_1 \tag{3.3}$$

with the pair $(\phi,\ W)$ as in (3.2).

Let $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \mathbf{0}^T\right)^T$ with $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^{k_n}$ denoting the oracle-regularized estimator minimizing $Q_n(\boldsymbol{\beta})$ defined by (3.3) over the space $\mathbb{A} = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbb{R}^{p_n - k_n}\}$. That is, $\hat{\boldsymbol{\beta}}_1$ represents the estimator of signal covariates. The next theorem establishes the consistency of $\hat{\boldsymbol{\beta}}_1$. All limits in the theorems below are taken as $n \to \infty$ unless otherwise stated.

Theorem 3.1. *Assume that conditions* (D1) *to* (D4) *hold. Suppose* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \mathbf{0}^T\right)^T$ *is obtained by minimizing function* (3.3) *over the space* $\mathbb{A}$. *Additionally, suppose that* $\lambda_n\|\mathbf{d}_0\|_2\sqrt{k_n/n} \to 0$, $\lambda_n \geq \sqrt{(\log n)/n}$, *and let* $\alpha_n = c\left(\sqrt{k_n(\log n)/n} + \lambda_n\|\mathbf{d}_0\|_2\right)$ *for*

some constant $c > 0$. Then, with probability tending to one,

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \alpha_n \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq \alpha_n \sqrt{k_n}.$$

Further, if $\alpha_n^{-1} \min_{1 \leq j \leq k_n} |\beta_j^*| \to \infty$, then with probability tending to one, we have

$$\operatorname{sgn}(\hat{\boldsymbol{\beta}}_1) = \operatorname{sgn}(\boldsymbol{\beta}_1^*),$$

where the above equation should be understood componentwise.

Theorem 3.1 shows the rate of consistency for estimator $\hat{\boldsymbol{\beta}}$ in terms of l_2- and l_1-norms. Note that α_n is not a function of p_n as $Q_n(\boldsymbol{\beta})$ is minimized over the space $\mathbb{A}$ to obtain $\hat{\boldsymbol{\beta}}_1$. Also note that $\hat{\boldsymbol{\beta}}_1$ estimates the correct sign of the true coefficient vector $\boldsymbol{\beta}_1^*$ with probability tending to one. The first component of α_n, $\sqrt{k_n(\log n)/n}$, is the oracle rate within a factor of $\log n$, and the second term, $\lambda_n \|\mathbf{d}_0\|_2$, represents the extra bias caused by regularization. If no prior information is available then one can set $\mathbf{d}_0 = (1, 1, ..., 1)^T$, in which case the l_2 rate of consistency of $\hat{\boldsymbol{\beta}}_1$ satisfies $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq c(\sqrt{k_n \log n/n} + \lambda_n \sqrt{k_n})$ with probability tending to one.

Theorem 3.2. *Assume that conditions* (D1) *to* (D4) *hold and* $\min_{j \geq k_n+1} d_j > 0$. *Let* α_n *be as in Theorem* 3.1. *Further, assume that* $\lambda_n \|\mathbf{d}_0\|_2 \sqrt{k_n/n} \to 0$, $\lambda_n > C^* \alpha_n \sqrt{k_n}$, $\lambda_n > 2\sqrt{(1+c)\log p_n/n}$, *and* $k_n^{1/2} \alpha_n^{\delta} \sqrt{(1 + k_n^{3/2} \alpha_n) \log_2 n} = o(\sqrt{n}\lambda_n)$, *where* $c > 0$ *and* C^* *are some constants. Then there exists a global minimizer* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T\right)^T$ *of* $Q_n(\boldsymbol{\beta})$ *defined by* (3.3) *satisfying* $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \alpha_n$, $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq \alpha_n \sqrt{k_n}$, *and* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *with probability tending to one.*

Theorem 3.2 presents the rate of consistency and the model selection oracle property of the proposed regularized estimator $\hat{\boldsymbol{\beta}}$ when the oracle information is not available. Theorem 3.2, however, requires that the weights in the weight vector $\mathbf{d}$ corresponding to the noise covariates to be strictly positive, i.e., $\min_{j \geq k_n+1} d_j > 0$, such that each coordinate is penalized. The results in Theorem 3.2 also hold for heavy-tailed errors

provided that conditions (D1) and (D4) are satisfied and that the regularization parameter $\lambda_n \geq \sqrt{\log p_n/n}$, which is the optimal choice for Gaussian errors (Bickel et al. 2009).

If subgradient $\phi^{(1)}$ is bounded, such as for the l_1-loss and the Huber loss functions, then Theorem 3.2 holds with the condition $\sqrt{(1+k_n^{1+\delta/2}\alpha_n)}\log_2 n = o\left(\sqrt{n}\lambda_n\right)$ replaced by $\sqrt{(1+k_n\alpha_n)}\log_2 n = o(\sqrt{n}\lambda_n)$. For $\mathbf{d}_0 = (1,1,...,1)^T$ and $\lambda_n = c_1'\sqrt{\log p_n/n}$, α_n reduces to $\alpha_n = c_2'\left(\sqrt{k_n(\log n)/n} + \sqrt{k_n(\log p_n)/n}\right)$, and hence the upper bound on $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2$ becomes $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq 2c_2'\sqrt{k_n(\log p_n)/n}$, provided $k_n = o(n^{2/3}(\log n)^{1/3})$ and $k_n = o(n(\log n)^{-1/2})$, where c_1' and c_2' are positive constants. The preceding upper bound convergence rate is the same as the rate for the LAD estimator in Wang (2013) and the Huber loss estimator in Loh (2017, 2018).

The proofs of Theorems 3.1–3.2 are based on Lemmas 3.2–3.3 given in Appendix B. Lemma 3.2 is a novel concentration type inequality, established using the symmetrization theorem (see, e.g., Theorem 14.3 in Bühlmann and van de Geer 2011), Massart's concentration theorem (see, e.g., Theorem 14.2 in Bühlmann and van de Geer 2011), and Theorem 2.7.11 of van der Vaart and Wellner (1996) on the bracketing/covering numbers. A key condition used in the proofs is condition (D4), which enables us to establish an important large deviation type result using concentration inequalities for sub-Gaussian random variables (see, e.g., Corollary 14.6 in Bühlmann and van de Geer 2011).

Theorem 3.3. *Suppose the conditions of Theorem* 3.2 *,* $\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2 = O(\sqrt{k_n})$*, and* $\sqrt{n/k_n}\min_{1\leq j\leq k_n}|\beta_j^*| \to \infty$ *hold. In addition, assume* k_n *satisfies* $k_n = o(n^{\delta/(3+2\delta)})$ *and* $k_n = o(n^{\varepsilon/(4+2\varepsilon)})$ *simultaneously where* δ *and* ε *are defined in conditions* (D1) *and* (D2)*. Then with probability tending to one, there exists a global minimizer* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T,\ \hat{\boldsymbol{\beta}}_2^T\right)^T$ *of* (3.3) *such that* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *and*

$$\mathbf{c}^T\left(\left(\mathbf{Z}_n^*\right)^T\mathbf{Z}_n^*\right)^{-1/2}\left(2\mathbf{V}_n^{-1}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right) + n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0\right) \xrightarrow{\mathcal{D}} N\left(0,\sigma_\gamma^2\right),$$

where $\mathbf{c}$ *is an arbitrary* k_n*-dimensional vector satisfying* $\mathbf{c}^T\mathbf{c} = 1$ *and* $\tilde{\mathbf{d}}_0$ *is a* k_n*-dimensional vector with the* j*th element* $d_j\,\mathrm{sgn}(\beta_j^*)$*.*

The framework of proving Theorem 3.3 is similar to that of the proof on the asymptotic normality theorem for the quantile regression estimator in Fan et al. (2014), in which their theorem is proved under a very challenging ultrahigh-dimensional setting where $\log(p_n) = o(n^b)$ for some constant $b > 0$. Many authors have obtained results of similar nature as in Theorems 3.2–3.3 on consistency, model selection oracle property, and asymptotic normality for various high- and ultrahigh-dimensional regularized robust regression M-estimators, e.g., Fan and Peng (2004), Li et al. (2011), Bradic et al. (2011), Fan et al. (2017), Loh (2017), and Sun et al. (2020), among others. Notwithstanding these advances, only Li et al. (2011) have investigated such results for l_q-loss functions ($1 \leq q < 2$), as far as we are aware. They considered the fixed-covariates case and assumed that $p_n^2/n \to 0$. Thus, our results are rather interesting and original.

It is easy to verify that the l_q-loss functions with $1 \leq q < 2$ satisfy the local δ-Lipschitz condition. Several loss functions with bounded gradients/subgradients, including the Huber loss, also satisfy this condition. If $\phi^{(1)}$ is bounded then condition (D4) easily follows, so it is not necessary.

The outliers of covariates are known to severely influence the performance of estimators (Huber 1981). We note that our estimator is not affected by the outliers of covariates due to the weight $W(\cdot)$ placed on the covariates. This feature, however, results in some loss in efficiency, as the observations with large covariate values are downweighted even if they are well-fitted. To mitigate this problem, we have introduced a tuning parameter γ for efficiency improvement. It can considerably improve the efficiency of estimators in the presence of outliers, particularly the outliers of the covariates.

In practice, the tuning parameter γ is determined by a data-driven method. From a nonasymptotic viewpoint, γ can be chosen to balance the bias and robustness; see, e.g., Fan et al. (2017) and Sun et al. (2020). From an asymptotic standpoint on the other hand, γ may be chosen to minimize a sample estimate of the asymptotic variance of $\hat{\boldsymbol{\beta}}_1$, see Proposal 3 of Huber (1964, p. 98). For instance, in light of Theorem 3.3, a data-driven

value for γ may be obtained as follows:

$$\hat{\gamma} = \arg\min_{\gamma} \mathbf{c}^T \hat{\mathbf{V}}_n \hat{\mathbf{V}}_n^T \left(\sum_{i=1}^{n} \left(\mathbf{X}_{1i} \mathbf{X}_{1i}^T / \gamma^2 \right) W^2 \left(\mathbf{X}_i / \gamma \right) \right) \hat{\mathbf{V}}_n \hat{\mathbf{V}}_n^T \mathbf{c} \hat{\sigma}_{\gamma}^2,$$

where $\mathbf{c}$ is a vector satisfying $\mathbf{c}^T\mathbf{c} = 1$, and $\hat{\mathbf{V}}_n$ and $\hat{\sigma}_{\gamma}^2$ denote consistent estimators of $\mathbf{V}_n$ and σ_{γ}^2, respectively. We employed this method in our numerical studies, see Section 3.5. One may also select γ so that $\sum_{i=1}^{n} [\phi^{(1)}((Y_i - \mathbf{X}_{1i}^T \hat{\boldsymbol{\beta}}_1)/\gamma)]^2$ has a predetermined value, see Proposal 2 of Huber (1964, p. 97). Huber favored the latter approach because it "fits best into the framework of conventional least squares techniques."

3.4 Computational algorithm

For some loss functions, such as the l_q-loss functions with $1 < q < 2$, the objective function (3.3) is not very smooth and, hence, the corresponding regularized optimization problem of (3.3) is computationally challenging and costly. In order to accelerate the computation of $\hat{\boldsymbol{\beta}}$ defined based on (3.3), here we implement a half-quadratic (HQ) reformulation of the original objective function pioneered by Geman and Reynolds (1992) and Geman and Yang (1995). A real-valued function $K(\cdot, \cdot)$ of two variables u and v is said to be HQ if K is the quadratic function of u, where u and v may be vectors.

To motivate, let ϕ be a symmetric loss function satisfying the following conditions:

(i) ϕ is convex and even;

(ii) $\phi(\sqrt{\cdot})$ is concave on $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$;

(iii) ϕ is continuous near zero and $\mathcal{C}^1$ on $\mathbb{R} \setminus \{0\}$, where $\mathcal{C}^1$ denotes continuously differentiable functions;

(iv) $\lim_{x \to \infty} \phi(x)/x^2 = 0$.

That is, ϕ grows slowly compared to the squared error loss function, and hence we expect ϕ to be more robust against outliers. Most loss functions used in the robust estimation literature satisfy the above conditions (i) to (iv), including the l_q-loss functions $(0 < q < 2)$ and the Huber loss function.

The dual function of ϕ is given by $\psi(t) = \sup_{x\in\mathbb{R}}(\phi(x) - tx^2)$. Reciprocally, we have $\phi(x) = \inf_{t\in\mathbb{R}}(tx^2 + \psi(t))$ using the theory of convex conjugate functions; see, e.g., Boyd and Vandenberger (2004). That is, ϕ is the infimum of a family of quadratic functions. This is known as a half-quadratic approximation of a convex function (Geman and Reynolds 1992). The relationship between ϕ and ψ under different assumptions on ϕ is analyzed in Geman and Reynolds (1992), Charbonnier et al. (1997), and Idier (2001), among others. Using a one-to-one transformation $\phi(x)$ can be reformulated as $\phi(x) = \inf_{s\geq 0}(\psi^{-1}(s)x^2 + s)$. Then the function $\psi^{-1}(s)x^2 + s$ is convex in (x, s) (see Idier 2001). Furthermore, $\psi^{-1}(s)$ is a decreasing function on $[0, \infty)$.

Based on the above construction we now consider the objective function

$$Q_n(\boldsymbol{\beta},\, \mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n}[\psi^{-1}(s_i)(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + s_i] + \lambda_n\|\mathbf{d}\circ\boldsymbol{\beta}\|_1 \tag{3.4}$$

where $s_i \in I_\phi,\ i = 1, ..., n;\ \mathbf{s} = (s_1, ..., s_n)^T \in I_\phi^n = I_\phi \times ... \times I_\phi$ (n terms); and $I_\phi = [0, \infty)$. Then $Q_n(\boldsymbol{\beta},\ \mathbf{s})$ is a convex function in $(\boldsymbol{\beta},\ \mathbf{s})$, as the first term in (3.4) is a sum of convex functions. Then we define an estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \arg_{\boldsymbol{\beta}} \min_{(\boldsymbol{\beta},\mathbf{s})\in\mathbb{R}^{p_n}\times I_\phi^n} Q_n(\boldsymbol{\beta},\ \mathbf{s}). \tag{3.5}$$

Observe that $\hat{\boldsymbol{\beta}}$ is robust against outlying observations, as $\psi^{-1}(\cdot)$ assigns a small value for outliers. The objective function (3.4) can also be used for outlier detection: a small value for $\psi^{-1}(s_{i_0})$ indicates that $(Y_{i_0}, \mathbf{X}_{i_0})$ is a potential outlier. The objective function (3.4) also exhibits how a robust loss function ϕ operates in the presence of outliers. The optimization problem (3.5) can be solved iteratively. That is, at each iteration one minimizes $Q_n(\boldsymbol{\beta},\ \mathbf{s})$ w.r.t. $\boldsymbol{\beta}$ for fixed $\mathbf{s}$ and then minimizes w.r.t. $\mathbf{s}$ for fixed $\boldsymbol{\beta}$. More specifically, if $\boldsymbol{\beta}^{(i)}$ and $\mathbf{s}^{(i)}$ denote the values after i^{th} iteration, then $(i+1)^{st}$ step of the

algorithm is given by

$$\boldsymbol{\beta}^{(i+1)} = \arg\min_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta},\ \mathbf{s}^{(i)})$$

$$\mathbf{s}^{(i+1)} = \arg\min_{\mathbf{s}\in I_\phi^n} Q_n(\boldsymbol{\beta}^{(i+1)},\ \mathbf{s}).$$

For the l_q-loss functions $\phi_q(t) = |t|^q$ with $1 \leq q < 2$, we have $\psi_q^{-1}(d) = \frac{q}{2}(\frac{2d}{2-q})^{(q-2)/q}$ for $d > 0$ (Idier 2001).

3.5 Simulation studies

A simulation study was conducted to evaluate the performance of the estimator $\hat{\boldsymbol{\beta}}$ along with other well-known estimators. As seen in the previous section, we can compute our estimator $\hat{\boldsymbol{\beta}}$ by minimizing an objective function of the form

$$\sum_{i=1}^{n}\left(\frac{q}{2}\left(\frac{2d_i}{2-q}\right)^{(q-2)/q}\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)^2 + d_i\right) W(\mathbf{X}_i/\gamma) + n\lambda_n\|\hat{\mathbf{d}}\circ\boldsymbol{\beta}\|, \tag{3.6}$$

with an adaptive penalty function as in Chapter 2 simulation.

We now state the setups used in this simulation study. We set $q = 1.5$ in (3.6) throughout this experiment. In addition, we fixed the sample size n to be 100 and the dimension p_n to be 400. In addition, the true regression vector was set to be

$$\boldsymbol{\beta} = (3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 0, \ldots, 0)\,,$$

where the number of signals was fixed to be 10.

Moreover, we generated the covariates from two scenarios. That is, we generated the covariates $\mathbf{X}_i$ from the following two settings:

(I) $\mathbf{X}_i$'s follow a multivariate normal distribution $N(0, \Omega_1)$ with covariance matrix $\Omega_1 = 0.5^{|i-j|}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$, where $n = 100$ and $p_n = 400$;

(II) $\mathbf{X}_i$'s follow a mixture normal distribution $0.8N(0, \Omega_2) + 0.2N(\mu, \Omega_1)$, where $\Omega_2 = I_{p_n\times p_n}$, $\mu = 3\mathbf{1}_{p_n}$, Ω_1 is as in setting I, and $p_n = 400$. This setup produces covariates

with some influential points.

Scenario I is widely used in simulation studies, and scenario II is used to generate covariates with very large values, i.e., high leverage points. Also, we assessed the performance of our estimator with different errors. Five different distributions of the noise were investigated in this simulation study:

1 Normal distribution with mean 0 and standard deviation 2.5, denoted as $N(0, 2.5)$;

2 Cauchy distribution with location 0 and scale 0.2;

3 MN1: A scale mixture of normal distributions for which $\sigma = 0.1$ with probability of 0.9 and $\sigma = 10$ with probability of 0.1;

4 MN2: Normal distribution $N(0, \sigma)$ where the standard deviation σ is sampled from a uniform distribution $Unif(1, 6)$;

5 t_2: Student's t distribution with degrees of freedom 2, denoted as t_2.

Then we implemented two regression models to obtain the response variable Y_i:

1. A high-dimensional linear regression model given by $Y_i = \mathbf{X}_i^T\boldsymbol{\beta}+e_i$ with $i = 1, \ldots, n$. This model is denoted as model 1;

2. A high-dimensional heteroscedastic model given by

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \left(\|\boldsymbol{\beta}\|_2^2\right)^{-1}\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)e_i,\ \ i = 1, \ldots, n.$$

This model is denoted as model 2.

To demonstrate the advantages of the proposed procedure in the presence of the outliers and leverage points, we implemented other three methods for comparison:

(i) *Lasso*, the penalized least squares estimator with the l_1-penalty;

(ii) *R-Lasso*, the regularized LAD estimator with the l_1-penalty as in Wang (2013), which is the same as the R-Lasso estimator in Fan et al. (2014);

(iii) *Huber-Lasso*, the Huber function with $\alpha = \text{IQR}(y)/10$ plus the l_1-penalty, where IQR stands for the inter-quartile range.

The performance of the estimators of interest was assessed using the following measures:

(1) l_2 loss, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$;

(2) l_1 loss, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$;

(3) FP: the number of false positives, i.e., the number of noises included in the model;

(4) FN: the number of false negatives, i.e., the number of signal covariates that are not included;

(5) SD_1: the standard deviation of l_2 loss;

(6) SD_2: the standard deviation of l_1 loss.

For l_2 and l_1, we report both mean and median values in the tables presented below.

The minimization problem of function (3.6) was solved by using the block coordinate gradient descent (BCGD) algorithm. To simplify the simulations, in this simulation study, we set the regularization parameter λ_n to be $0.3\sqrt{\log p_n/n}$. Our simulation demonstrates that this choice of the regularization parameter could produce satisfactory results without complicated computation that is demanded by applying cross-validation to select the optimal λ_n. The weight vector $\hat{\mathbf{d}}$ in (3.6) was calculated based on the SCAD penalty using R-Lasso as the initial estimator, and the tuning parameter γ in (3.6) was chosen as described in Section 3.3. The results were obtained from 100 randomly generated datasets. The results of the simulation study are summarized in Tables 3.1–3.4.

We now comment on the simulation results. Tables 3.1 and 3.3 present the simulation results under model 1 with covariates from settings I and II, accordingly. It can be seen from Tables 3.1 and 3.3 that the new method, l_q Adaptive-Lasso, performed overwhelmingly better than other procedures under almost all the scenarios. For example,

when errors followed normal distributions, the l_2 losses of Lasso, R-Lasso and Huber-Lasso were 2.329, 1.875, and 1.914, respectively, while the l_q Adaptive-Lasso gave only 1.553 for the l_2 loss. A bigger advantage of using the new estimator is evident with the l_1 loss and FN. For example, again when errors followed normal distributions, our estimator's l_1 loss was only 4.792, which was far less than the l_1 losses of other estimators. A similar phenomenon is also revealed under other setups of error distributions. However, we shall note that our estimator yielded larger l_2 losses than R-Lasso when the errors followed Cauchy distributions. Next, we comment on the simulation results in Tables 3.2 and 3.4, where the simulation results are based on model 2 with settings I and II, respectively. Overall, a similar trend as in Tables 3.1 and 3.3 is observed; however, model 2 tended to produce smaller estimation errors when compared to those of model 1. For example, Table 3.2 reveals that if the noise e_i followed the MN2 distribution, the l_2 losses of Lasso, R-Lasso, and Huber-Lasso were 0.440, 0.306, and 0.384, respectively, whereas the l_2 loss of l_q Adaptive-Lasso was 0.268, which again demonstrates that our estimator can well handle large values of the covariates, i.e., high leverage points.

Under most scenarios studied in this simulation study, we observe that R-Lasso outperformed the other two methods without adaptive weights in the penalties in terms of l_1 and l_2 losses. In addition, the simulation results reveal that as tails got heavier, Lasso's performance began to deteriorate quickly. This observation affirms again that R-Lasso and Huber-Lasso are more capable of handling errors with heavy tails than Lasso method. On the other hand, in the presence of outliers and high leverage points, the effectiveness of our estimator is self-evident.

3.6 Real data application

In this example, we used the ovarian dataset 8-7-02 provided by the National Cancer Institute (NCI) to evaluate the performance of our estimator. The dataset has also been

Table 3.1: Simulation results under model 1 with covariates from setting I.

Scenario	**Method**	$\boldsymbol{l}_2$	$\boldsymbol{l}_1$	**SD1**, **SD2**	**FN**, **PN**
Normal	Lasso	2.329, 2.293	15.454, 15.359	0.332, 1.916	75.13, 0
	R-Lasso	1.875, 1.835	8.268, 8.191	0.358, 1.481	33.99, 0
	Huber-Lasso	1.914, 1.879	10.564, 10.483	0.289, 1.682	49.71, 0
	l_q Adaptive-Lasso	1.553, 1.540	4.792, 4.569	0.302, 1.037	13.75, 0
Cauchy	Lasso	6.994, 2.214	55.937, 14.089	15.836, 150.252	75.91, 0.54
	R-Lasso	0.295, 0.280	1.281, 1.219	0.081, 0.337	33.65, 0
	Huber-Lasso	0.523, 0.496	1.832, 1.690	0.174, 0.773	14.11, 0
	l_q Adaptive-Lasso	0.344, 0.269	1.081, 0.723	0.353, 0.998	13.50, 0
MN1	Lasso	3.044, 2.986	20.919, 20.949	1.000, 7.389	78.46, 0
	R-Lasso	0.930, 0.917	4.053, 3.924	0.210, 0.863	33.95, 0
	Huber-Lasso	1.134, 1.074	5.429, 5.092	0.349, 2.013	34.27, 0
	l_q Adaptive-Lasso	0.820, 0.839	2.331, 2.381	0.216, 0.700	7.45, 0
MN2	Lasso	3.669, 3.634	25.495, 25.410	0.474, 3.140	79.93, 0
	R-Lasso	2.554, 2.432	11.051, 10.714	0.561, 2.232	33.45, 0
	Huber-Lasso	2.951, 2.902	17.271, 17.024	0.476, 2.839	56.55, 0
	l_q Adaptive-Lasso	2.332, 2.304	8.863, 8.879	0.411, 1.783	13.63, 0
$\boldsymbol{t}_2$	Lasso	3.096, 2.154	21.851, 14.147	3.861, 32.849	73.62, 0.17
	R-Lasso	1.056, 1.087	4.611, 4.678	0.249, 0.991	34.72, 0
	Huber-Lasso	1.133, 1.104	5.671, 5.592	0.265, 1.395	37.72, 0
	l_q Adaptive-Lasso	0.916, 0.891	2.709, 2.529	0.242, 0.871	10.42, 0

Table 3.2: Simulation results under model 2 with covariates from setting I.

Scenario	**Method**	l_2	l_1	**SD1**, **SD2**	**FN**, **PN**
Normal	Lasso	0.258, 0.254	0.928, 0.899	0.057, 0.251	16.20, 0
	R-Lasso	0.234, 0.226	0.989, 0.971	0.054, 0.220	33.03, 0
	Huber-Lasso	0.303, 0.302	0.822, 0.807	0.062, 0.173	1.23, 0
	l_q Adaptive-Lasso	0.197, 0.198	0.511, 0.509	0.046, 0.122	2.41, 0
Cauchy	Lasso	0.705, 0.190	4.461, 0.548	2.985, 24.099	19.03, 0.05
	R-Lasso	0.035, 0.032	0.145, 0.135	0.012, 0.043	30.96, 0
	Huber-Lasso	0.275, 0.266	0.741, 0.717	0.054, 0.146	0.28, 0
	l_q Adaptive-Lasso	0.052, 0.029	0.134, 0.077	0.068, 0.174	5.80, 0
MN1	Lasso	0.300, 0.275	1.230, 0.994	0.148, 0.867	18.96, 0
	R-Lasso	0.110, 0.109	0.470, 0.463	0.114, 0.076	31.10, 0
	Huber-Lasso	0.298, 0.284	0.805, 0.767	0.076, 0.224	1.02, 0
	l_q Adaptive-Lasso	0.099, 0.094	0.257, 0.237	0.029, 0.076	3.46, 0
MN2	Lasso	0.440, 0.434	2.033, 1.958	0.107, 0.638	33.10, 0
	R-Lasso	0.306, 0.303	1.301, 1.274	0.078, 0.303	34.00, 0
	Huber-Lasso	0.384, 0.365	1.129, 1.075	0.089, 0.292	5.71, 0
	l_q Adaptive-Lasso	0.268, 0.256	0.714, 0.698	0.079, 0.220	3.40, 0
t_2	Lasso	0.246, 0.213	0.923, 0.672	0.117, 0.754	13.46, 0
	R-Lasso	0.134, 0.124	0.559, 0.555	0.037, 0.155	32.26, 0
	Huber-Lasso	0.294, 0.279	0.802, 0.753	0.075, 0.212	0.82, 0
	l_q Adaptive-Lasso	0.126, 0.109	0.334, 0.281	0.066, 0.182	3.55, 0

Table 3.3: Simulation results under model 1 with covariates from setting II.

Scenario	**Method**	$\boldsymbol{l}_2$	$\boldsymbol{l}_1$	**SD1**, **SD2**	**FN**, **PN**
Normal	Lasso	2.699, 2.649	17.579, 17.376	0.456, 2.805	78.31, 0
	R-Lasso	2.211, 2.177	9.505, 9.218	0.446, 2.215	31.60, 0
	Huber-Lasso	2.077, 2.036	8.801, 8.663	0.360, 2.078	33.62, 0
	l_q Adaptive-Lasso	2.001, 2.006	7.365, 6.978	0.438, 2.275	12.50, 0
Cauchy	Lasso	6.447, 2.595	50.975, 18.200	13.744, 130.232	79.66, 0.46
	R-Lasso	0.393, 0.369	1.702, 1.589	0.160, 0.780	31.53, 0
	Huber-Lasso	0.678, 0.623	2.829, 2.628	0.240, 1.050	24.24, 0
	l_q Adaptive-Lasso	0.392, 0.357	1.156, 1.442	0.177, 0.636	19.50, 0
MN1	Lasso	3.170, 3.082	21.208, 20.793	1.050, 7.541	81.01, 0.01
	R-Lasso	1.093, 1.041	4.772, 4.492	0.314, 1.581	32.41, 0
	Huber-Lasso	1.167, 1.153	5.110, 5.066	0.348, 1.635	25.08, 0
	l_q Adaptive-Lasso	1.001, 0.940	3.182, 2.940	0.338, 1.401	11.42, 0
MN2	Lasso	3.900, 3.803	26.798, 26.356	0.654, 4.289	87.05, 0
	R-Lasso	3.066, 3.015	13.320, 12.746	0.667, 3.417	32.58, 0
	Huber-Lasso	2.935, 2.866	14.056, 13.942	0.626, 3.588	39.69, 0
	l_q Adaptive-Lasso	2.837, 2.877	11.291, 11.413	0.558, 2.455	17.70, 0
$\boldsymbol{t}_2$	Lasso	3.017, 2.575	20.353, 16.408	2.313, 18.823	76.12, 0.24
	R-Lasso	1.279, 1.244	5.314, 5.329	0.252, 1.232	31.74, 0
	Huber-Lasso	1.304, 1.342	5.766, 5.963	0.316, 1.581	30.06, 0
	l_q Adaptive-Lasso	1.160, 1.113	3.671, 3.460	0.247, 0.982	9.88, 0

Table 3.4: Simulation results under model 2 with covariates from setting II.

Scenario	**Method**	l_2	l_1	**SD1**, **SD2**	**FN**, **PN**
Normal	Lasso	0.954, 0.827	5.220, 4.356	0.465, 2.968	49.09, 0
	R-Lasso	0.357, 0.324	1.567, 1.453	0.126, 0.606	35.81, 0
	Huber-Lasso	0.666, 0.631	2.826, 2.654	0.224, 1.052	25.10, 0
	l_q Adaptive-Lasso	0.313, 0.284	0.904, 0.826	0.105, 0.323	9.85, 0
Cauchy	Lasso	1.410, 0.332	9.453, 1.387	2.813, 21.861	38.96, 0.08
	R-Lasso	0.073, 0.063	0.327, 0.291	0.034, 0.155	35.78, 0
	Huber-Lasso	0.349, 0.300	1.451, 1.263	0.181, 0.787	23.11, 0
	l_q Adaptive-Lasso	0.096, 0.052	0.230, 0.145	0.144, 0.444	12.80, 0
MN1	Lasso	1.238, 0.972	7.421, 5.365	1.008, 6.708	49.67, 0
	R-Lasso	0.193, 0.175	0.868, 0.775	0.091, 0.437	35.55, 0
	Huber-Lasso	0.387, 0.318	1.609, 1.301	0.238, 1.041	22.97, 0
	l_q Adaptive-Lasso	0.221, 0.177	0.636, 0.508	0.148, 0.447	12.65, 0
MN2	Lasso	1.845, 1.826	11.421, 11.458	0.683, 4.547	67.11, 0
	R-Lasso	0.469, 0.436	2.097, 1.954	0.203, 0.907	36.36, 0
	Huber-Lasso	0.849, 0.811	3.685, 3.600	0.276, 1.330	26.48, 0
	l_q Adaptive-Lasso	0.416, 0.405	1.291, 1.149	0.178, 0.556	18.17, 0
$\boldsymbol{t}_2$	Lasso	0.918, 0.682	4.878, 3.254	0.706, 4.588	40.62, 0
	R-Lasso	0.219, 0.204	0.937, 0.854	0.086, 0.421	35.80, 0
	Huber-Lasso	0.481, 0.431	2.050, 1.826	0.196, 0.890	24.58, 0
	l_q Adaptive-Lasso	0.209, 0.193	0.610, 0.528	0.098, 0.299	10.76, 0

investigated by Wu et al. (2003) where they compared the performance of several classes of statistical methods for the classification of cancer based MS spectra that have been widely used for biomarker identification and genome-wide protein profiling. Yu et al. (2005) proposed a novel method for dimensionality reduction for ovarian cancer identification based on mass spectrometry data. Li et al. (2011) examined the performance of the nonconcave penalized M-estimation method in an ultrahigh-dimensional setup with using the ovarian dataset. We compared our method with Lasso, R-Lasso and Huber Lasso using this dataset.

The data is written as $(\mathbf{x}_i, Y_i)$ where $\mathbf{x}_i \in \mathbb{R}^{p_n}$ represents the intensity vector and $Y_i = 0/1$ denotes the sample cancer status (0 for control, 1 for cancer). This dataset consists of 15154 features and 253 spectra samples: 162 ovarian cancer samples and 91 control samples. The tuning parameter λ_n was determined by five-fold cross-validation. We randomly selected 113 cases (73 ovarian cancer samples and 40 control samples) and applied Lasso, R-Lasso, Huber-Lasso and l_q Adaptive-Lasso respectively to select the significant genes with the above tuning parameter. For further investigation, we selected 24 samples as the training data to fit the regression and predicted the responses of the test set. We repeated this process 100 times and then obtained the average of the mean squared errors of the four methods. Table 3.5 summarizes our numerical findings. According to Table 3.5, we observe that the new method tended to select more variables when compared with the other three methods. Moreover, similar to what we have observed from the simulation experiment, the new method produced a small mean squared error, thus providing an accurate estimate.

Table 3.5: The summary of real data analysis.

	Lasso	R-Lasso	Huber-Lasso	l_q Adaptive-Lasso
Number of selected variables	21	42	47	63
Mean squared error	0.0395	0.0238	0.2086	0.1773

3.7 Conclusions

In this chapter, we study the high-dimensional l_q-loss robust estimator for sparse linear regression. We construct an l_q-loss regularized estimator that is highly robust with respect to the outliers in both the responses and covariates. We introduce a novel computational algorithm to deal with computationally challenging l_q-loss estimation problems. We prove that the new estimator possesses the parameter consistency and oracle properties provided certain conditions are satisfied.

Compared with the estimators proposed in Chapter 2, l_q-loss regularized estimator based on a data-driven penalty function also excels other well-known methods in reducing the effect of outlying observations. Specifically, the proposed estimator enjoys advantages in lessening the influence of outlying values of the response and predictors. The data analysis part also demonstrates the practical use of the proposed method.

Future research along with this path may include choosing q by a data-driven method. It is expected that for errors with heavy-tailed distributions, q can be chosen in a way close to 1, while for errors with light-tailed and symmetric distributions, q can be chosen close to 2.

Chapter 4

Outlier detection and robust estimation via penalized regression

4.1 Introduction

One of the most challenging tasks in statistics is to accurately estimate the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ from some corrupted data. Recent progresses have largely centered around investigating large-scale datasets in which the number of variables greatly exceeds the number of observations. Datasets with a large number of variables but a relatively small sample size, i.e. high-dimensional datasets, pose unprecedented difficulties in statistical research. There have been various lines of research on high-dimensional statistical inference. One of the fundamental work is Lasso that uses an l_1-penalty (Tibshirani 1996) to enforce sparsity:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_n \|\boldsymbol{\beta}\|_1.$$

Another pioneering work is Fan and Li (2001), where they examined a general class of penalized robust regression estimators of the form

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|).$$

Other noted works include Wright et al. (2009), Wainwright (2009), Lv and Fan (2009), Zhang (2010), Meinshausen and Bühlmann (2010), and Loh (2018).

In practical applications, outliers are bound to occur and without proper methods to

handle them, those outliers may cause severe statistical problems in parameter estimation, inference, and model selection. Hampel et al. (1986) noted that a real dataset usually contains 1 to 10 percent outliers. Most of the existing techniques for parameter estimation, however, are highly fragile in the presence of outliers and high leverage points. Lasso, for example, is sensitive to outliers - even a single unusual observation may make it impossible to faithfully recover the regression vector $\boldsymbol{\beta}$. Quantile regression is robust to outliers, yet sensitive to the leverage points. Weighted LAD (WLAD) regression was then proposed by Giloni et al. (2006) to overcome the weakness of LAD regression thus improving the robustness. However, the performance of the WLAD estimator deteriorates when the percentage of outliers increases. Researchers have since managed to develop a variety of methods to identify the outliers so that robust estimation can be guaranteed; the leave-one-out approach (Weisberg 1985) is among them. However, these methods are restricted to tackle the situation where only one outlier occurs.

Thus, developing an estimator that is able to resolve outlier detection and robust estimation simultaneously has attracted increasing attention from statistical researchers. Gannaz (2006), McCann and Welsch (2007), and She and Owen (2011) investigated the *mean-shift model*: $Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \gamma_i + e_i$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$ acts as a vector indicating the locations of outliers. Andrea (2010) developed multivariate outlier tests based on the high-breakdown minimum covariance determinant estimator. She and Owen (2011) introduced an estimator by solving the squared loss of the mean-shift model with an l_1-penalty on the mean-shift vector $\boldsymbol{\gamma}$. This method enjoys good performance on identifying outliers. However, it fails to perform variable selection and also has small breakdown points. Alfons et al. (2013) developed a procedure that is called the *sparse least trimmed squares method*. This procedure is designed to solve $\min_{i=1}^{h} r_{(i)}^2$ where $r_{(i)}^2$'s are the order statistics of $r_i^2 = (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2$ and h is a truncation number. Kong et al. (2018) proposed a method based on the squared loss of the mean-shift model with two penalty functions on the mean-shift vector and the parameter vector. Their estimator can achieve both high breakdown points and high efficiency. Gao and Feng (2018) proposed the PWLAD method

in which the weights quantifying the outlying effects for each observation are introduced to the objective function. Jiang et al. (2020) inherited the idea of Gao and Feng (2018) proposing the penalized weighted LAD-LASSO (PWLAD-LASSO) estimator such that the new estimator possesses both robust estimation and outlier detection properties. She et al. (2021) established a general resistant learning framework to robustify an arbitrarily given loss. In their paper, they introduced an $l_0 + l_2$ form of regularization to address the situation in which gross outliers occur in the high-dimensional data.

In this chapter, our goal is also to develop a procedure that can perform outliers detection and robust parameter estimation simultaneously. Again we consider the following linear regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i, \ i = 1, ..., n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ is a p_n-dimensional coefficient vector, $\boldsymbol{X}_i \in \mathbb{R}^{p_n}$ is a random covariate vector and $e_i \in \mathbb{R}$ is a random error. We denote $\boldsymbol{\beta}^*$ as the true regression vector of $\boldsymbol{\beta}$.

To tackle the minimization of regularized functions, Geman and Reynolds (1992) and Geman and Yang (1995) developed the multiplicative and additive half-quadratic reformulation of the original functions. For the Huber loss (Huber 1964) defined as $\phi_\alpha(t) = \frac{1}{2}t^2 I[|t| \leq \alpha] + \alpha(|t| - \frac{1}{2}\alpha) I[|t| > \alpha]$, where $\alpha > 0$, the dual function of $\phi_\alpha(t)$ satisfying $\psi_\alpha(s) = \sup_{x \in \mathbb{R}} (\phi_\alpha(x) - \frac{1}{2}(x\sqrt{c} - \frac{s}{\sqrt{c}})^2)$ is given by $\psi_\alpha(s) = \frac{s^2}{2c(c-1)}$ if $|s| \leq (c-1)\alpha$, otherwise $\psi_\alpha(s) = \frac{\alpha|s|}{c} - \frac{\alpha^2(c-1)}{2c}$, for $c > 1$. For $c = 1$, $\psi_\alpha(s) = \alpha|s|$. Then we have $\phi_\alpha(x) = \inf_{s \in \mathbb{R}} (\frac{1}{2}(x\sqrt{c} - \frac{s}{\sqrt{c}})^2 + \psi_\alpha(s))$, see Nikolova and Ng (2005). Based on this construction, we obtain the estimator of $\boldsymbol{\beta}^*$ by minimizing the following objective function

$$Q_n(\boldsymbol{\beta}, \mathbf{s}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{c}{2} \left(Y_i - \mathbf{X}_i^T \boldsymbol{\beta} - \frac{s_i}{c} \right)^2 + \psi_\alpha(s_i) \right) W(\mathbf{X}_i) + p_{\lambda_n}(\boldsymbol{\beta}), \tag{4.1}$$

where $c \geq 1$, $\boldsymbol{s} = (s_1, ..., s_n)^T$ and $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^{p_n} |\beta_j|$. See Section 3.4 for motivation. For a vector $\mathbf{x} \in \mathbb{R}^{p_n}$ and an arbitrary set $T \subseteq \{1, \ldots, p_n\}$, we let $\mathbf{x}_T \in \mathbb{R}^T$ denote the vector $\mathbf{x}$ restricted to T. In this chapter, we denote $T := \text{supp}(\boldsymbol{\beta}^*)$, the support of $\boldsymbol{\beta}^*$. As

in the classical framework of regularization problems, $p_{\lambda_n}(\boldsymbol{\beta})$ encourages the sparsity of $\boldsymbol{\beta}$. We aim at minimizing $Q_n(\boldsymbol{\beta},\ \mathbf{s})$ w.r.t. $(\boldsymbol{\beta},\ \mathbf{s})$:

$$\left(\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{s}}\right) = \arg\min_{(\boldsymbol{\beta},\ \mathbf{s})} Q_n(\boldsymbol{\beta},\ \mathbf{s}). \tag{4.2}$$

The optimization problem (4.2) can be solved iteratively. That is, at each iteration one minimizes $Q_n(\boldsymbol{\beta},\ \mathbf{s})$ w.r.t. $\boldsymbol{\beta}$ for fixed $\mathbf{s}$ and then minimizes w.r.t. $\mathbf{s}$ for fixed $\boldsymbol{\beta}$. In (4.1), the auxiliary vector $\mathbf{s}$ can be viewed as the error vector, similar to $\boldsymbol{\gamma}$ in the mean-shift model. Therefore, the new procedure can be used for parameter estimation and outlier detection simultaneously.

To establish the desired theoretical results, we need a *restricted strong convexity* (RSC) condition (see Negahban et al. 2012), which has been shown to be critical in the study of high-dimensional frameworks. The loss function $L_n(\boldsymbol{\theta})$ satisfies the RSC condition over some set $\mathbb{A}$ if

$$L_n(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - L_n(\boldsymbol{\theta}^*) - \langle \nabla L_n(\boldsymbol{\theta}^*), \boldsymbol{\Delta} \rangle \geq \kappa_L \|\boldsymbol{\Delta}\|_2^2 - \tau_L^2$$

holds with some curvature $\kappa_L > 0$ and tolerance τ_L for all $\boldsymbol{\Delta} \in \mathbb{A}$ where $\boldsymbol{\theta}^*$ is the vector containing the true values of $\boldsymbol{\theta}$.

The organization of this chapter is as follows. In Section 4.2, we provide non-asymptotic bounds for the l_2 error: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}^*$. Section 4.2 also details the specific choices of the regularization parameter λ_n under two types of error distributions: sub-Gaussian and sub-exponential distributions. Sections 4.3 and 4.4 present the simulation and real data analysis results, respectively. The proofs related to our theoretical results are present in Appendix C.

4.2 Parameter estimation of $\boldsymbol{\beta}$

Theorem 4.1. *Suppose that* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *are i.i.d random variables satisfying* $\|\boldsymbol{X}_i\|_2^2 W(\boldsymbol{X}_i) \leq 1$ *and* $\lambda_{\min}\left(\mathbf{E}\left(\frac{c}{2}\boldsymbol{X}_1\boldsymbol{X}_1^T W(\boldsymbol{X}_1)\right)\right) > 0$ *for all* $i = 1, \ldots, n$. *Further, assume that* $\lambda_n \geq$

$2\|\nabla_{\boldsymbol{\beta}} L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s})\|_\infty$ *for any* $\boldsymbol{s}$, *where*

$$L_n^*(\boldsymbol{\beta}, \boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{c}{2}\left(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - \frac{s_i}{c}\right)^2 + \psi_\alpha(s_i)\right) W(\mathbf{X}_i).$$

Then with high probability, there exists some constant $\zeta > 0$ *such that the estimate* $\hat{\boldsymbol{\beta}}$ *from* (4.2) *satisfies*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 3\zeta^{-1}\lambda_n\sqrt{k_n},$$

where k_n *is the number of the signals of* $\boldsymbol{\beta}^*$.

In Theorems 4.2 and 4.3 exhibited below, we give the mechanisms for selecting the regularization parameter λ_n such that the constraint $\lambda_n \geq 2\|\nabla_{\boldsymbol{\beta}} L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s})\|_\infty$ holds. For this purpose, we assume the following condition on $\boldsymbol{s}$ which gives an upper bound on the number of contaminated observations.

Condition 1: $\boldsymbol{s}$ satisfies $\|\boldsymbol{s}\|_1 \leq c_1 a_n$, where $a_n \leq \sqrt{n\log(p_n)}$ and c_1 is some positive constant.

Theorem 4.2. *Suppose that condition* 1 *holds and* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *are i.i.d random variables satisfying* $\|\boldsymbol{X}_i\|_2^2 W(\boldsymbol{X}_i) \leq 1$. *Moreover, assume that the* e_i*'s are independent of the* $\mathbf{X}_i$*'s and follow a sub-Gaussian distribution with* $\mathbf{E}(e_i) = 0$. *In addition, if the regularization parameter* λ_n *satisfies*

$$\lambda_n = 2\left(c\sqrt{\frac{2\sigma^2\varepsilon\log p_n}{n}} + c_1\sqrt{\frac{\log p_n}{n}}\right)$$

for some constants $\varepsilon > 1$ *and* $c_1 > 0$, *then with high probability,* $\lambda_n \geq 2\|\nabla_{\boldsymbol{\beta}} L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s})\|_\infty$.

Theorem 4.3. *Suppose that condition* 1 *holds and* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *are i.i.d random variables satisfying* $\|\boldsymbol{X}_i\|_2^2 W(\boldsymbol{X}_i) \leq 1$. *Moreover, assume that the* e_i*'s are independent of the* $\mathbf{X}_i$*'s and follow a distribution that satisfies* $\mathbf{E}(e_i) = 0$ *and*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}|e_i|^m \leq \frac{m!}{2}K^{m-2}$$

for any $i = 1, \ldots, n$, $m = 2, 3, \ldots$, *and some constant* $K > 0$. *In addition, if the regularization parameter* λ_n *satisfies*

$$\lambda_n = 2\left(c_2\sqrt{\frac{\log(2p_n)}{n}} + c_1\sqrt{\frac{\log(p_n)}{n}}\right)$$

where c_1 *and* c_2 *are some positive constants, then with high probability,* $\lambda_n \geq 2\|\nabla_{\boldsymbol{\beta}} L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s})\|_\infty$.

The proofs of above theorems are provided in Appendix C.

4.3 Simulation study

In this section, we demonstrate the theoretical properties of our new method via a simulation study. First, let us introduce the mean-shift model defined by

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + u_i + \varepsilon_i$$

for $i = 1, \ldots, n$, where n is the sample size. Note that $Y_i \in \mathbb{R}$ is a response variable, $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ is a p_n-dimensional regression coefficient vector, $\mathbf{X}_i \in \mathbb{R}^{p_n}$ is a covariate vector, and $\varepsilon_i \in \mathbb{R}$ represents an error term that is generated from a standard normal distribution $N(0, 1)$. The mean-shift vector $\boldsymbol{u} = (u_1, \ldots, u_n)^T$ artificially injects outliers when u_i is nonzero. In other words, when $u_i = 0$ the ith observation is not an outlier, otherwise the ith observation is an outlier. It is expected that most of the components in $\boldsymbol{u}$ are zero as in general only a few outliers are present in applications. More details about the formulation of the mean-shift model can be found in Gannaz (2006) and McCann and Welsch (2007).

In our simulation study, we plan to verify that the new procedure is capable of detecting outliers and has good robustness properties. That is, it could effectively reduce *masking* and *swamping*, thus improving the efficiency and accuracy of the estimators. When multiple outliers mask each other and go undetected is known as masking, whereas swamping essentially means labeling good observations as outliers. These two effects are

more common in large datasets with multiple outliers.

To evaluate the performance of the new method, two different settings were investigated:

1. Setting I: We generated the design matrix $\mathbf{X}$ from $N(0, \mathbf{\Sigma})$ where $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. Then we modified the first O rows of $\mathbf{X}$ to be $L \times [1, \ldots, 1]$. Two different values of L were studied: 5/10. We set the shift vector $\boldsymbol{u}$ to be $\left(\{5\}^O, \{0\}^{n-O}\right)^T$. We investigated different O in this simulation experiment. The true coefficient vector was defined as

$$\boldsymbol{\beta}_1^* = (1, 0.3, 0.2, 0, 0, -0.2, -0.3, -1, 0, \ldots, 0),$$

where $p_n = 500$ and $n = 200$.

2. Setting II: We considered a setup where sample size $n = 200$ and p_n was set to be 1000 with the true vector being fixed as

$$\boldsymbol{\beta}_2^* = (1, 0.5, 0, 0, -0.5, -1, 0, \ldots, 0).$$

Note that the number of signals was fixed as 4. The design matrix $\boldsymbol{X}$ was generated from the same scheme employed in setting I.

Note again that we are interested in minimizing the objective function given below:

$$Q_n(\boldsymbol{\beta}, \mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{c}{2}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - \frac{s_i}{c})^2 + \psi_\alpha(s_i)\right) W(\mathbf{X}_i) + \lambda_n \sum_{j=1}^{p_n} |\beta_i|,$$

where $s_i \in I_\phi,\ i = 1, ..., n$, $\mathbf{s}$ is sparse defined as $\mathbf{s} = (s_1, ..., s_n)^T \in I_\phi^n = I_\phi \times ... \times I_\phi$ (n terms) and $W(\boldsymbol{x})$ is a weight function given as $W(\boldsymbol{x}) = \min\{1, \frac{b}{(\|\mathbf{B}\boldsymbol{x}\|_2)^2}\}$ with $b \in \mathbb{R}$ and $\mathbf{B} \in \mathbb{R}^{p_n \times p_n}$ fixed constants. Without loss of generality, we set $b = 1$ and $\mathbf{B}$ to be an identity matrix. We also set $c = 1$, thus resulting in $\psi_\alpha(s) = \alpha|s|$. The tuning parameter α was fixed to be 1.34 throughout the experiment. The tuning parameter λ_n was fixed as $4\sqrt{\sigma^2 \log(p_n)/n}$.

The performance of the proposed estimator was assessed using the following measures:

1. M: the fraction of undetected true outliers (masking);

2. S: the fraction of good points perceived as outliers (swamping);

3. JD: the number of simulations without masking;

4. The mean squared error defined as

$$Err = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

   where $\hat{\boldsymbol{\beta}}$ is the estimate of the parameter vector $\boldsymbol{\beta}^*$;

5. The false zero rate (FZR), that is the fraction of nonzero coefficients that are estimated as zero;

6. The false positive rate (FPR), that is the fraction of zero coefficients that are estimated as nonzero;

7. The correct selection rate (SR), the fraction of identifying both nonzeros and zeros of $\boldsymbol{\beta}^*$;

8. The correct coverage rate (CR), the fraction of identifying nonzeros of $\boldsymbol{\beta}^*$.

In addition, we compared our method with the sparse least trimmed squares method (S-LTS), PM method (Kong et al. 2018), and PIQ (She et al. 2021). They were served as benchmarks for the evaluation of our procedure's performance. To reduce the computational complexity, we fixed the regularization parameter as $0.1\sqrt{\log(p_n)/n}$ for both setting I and setting II. A similar approach can be seen in Loh (2017). For the S-LTS method, the truncation number was set to be the largest integer less than $0.75n$, $\lfloor 0.75n \rfloor$. Alfrons et al. (2013) suggested that taking a value of h equal to 75% of the sample size could guarantee a sufficiently high statistical efficiency. For PM method, we took an approach by using the sparse least trimmed squares method to obtain the initial estimator. Again the truncation parameter was fixed as the floor of $0.75n$ throughout the simulation experiment to warrant a consistent comparison. For each setting, we present the average

values of the performance measures considered in this simulations by 100 repetitions. The simulation results are summarized in Tables 4.1–4.4:

Table 4.1: The Summary of the simulation study with $n = 200$, $p_n = 500$ and $L = 5$ under setting I.

O	**Method**	**M**	**S**	**JD**	**Err**	**FZR**	**FPR**	**SR**	**CR**
10	New	0.008	0.197	92	0.218	0.183	0.016	0	0.26
	PM	1	0.0003	0	0.654	0.480	0.004	0	0
	S-LTS	0	0.211	100	0.204	0.130	0.043	0	0.40
	PIQ	0.148	0.034	64	0.239	0.243	0.009	0	0.06
20	New	0.010	0.191	89	0.201	0.160	0.021	0	0.30
	PM	1	0	0	0.723	0.470	0.006	0	0
	S-LTS	0.001	0.167	98	0.195	0.086	0.067	0.02	0.48
	PIQ	0.239	0.082	52	0.297	0.267	0.009	0	0.06
30	New	0.057	0.187	76	0.256	0.160	0.021	0	0.36
	PM	1	0	0	0.753	0.490	0.005	0	0
	S-LTS	0.057	0.131	90	0.249	0.120	0.054	0	0.46
	PIQ	0.213	0.120	48	0.385	0.317	0.010	0	0.02
40	New	0.301	0.251	46	0.625	0.323	0.018	0	0.14
	PM	1	0.0003	0	0.765	0.477	0.007	0	0
	S-LTS	0.465	0.179	48	0.603	0.283	0.039	0.02	0.20
	PIQ	0.179	0.170	54	0.529	0.343	0.010	0	0

Now let us remark on the numerical results. From Tables 4.1 and 4.2, we find that our method had its best performance when L was relatively small and O was relatively large. For example when $L = 5$ and $O = 30$, the M values of both the new estimator and the S-LTS estimator were 0.057, while the M value of PIQ was 0.213. In addition, when O increased to 40, our estimator began to outperform the S-LTS estimator in terms of masking. However, when compared with the S-LTS estimator, our estimator tended to produce larger values of swamping and errors. We also notice that PM method, though had the smallest swamping values, heavily suffered from masking.

Additionally, Tables 4.1 to 4.4 reveal that the performance of methods investigated was also significantly affected by the values of L. To be specific, as the values of L increased, PM, PIQ, S-LTS along with our estimator began to yield larger M and Err

Table 4.2: The Summary of the simulation study with $n = 200$, $p_n = 500$ and $L = 10$ under setting I.

O	**Method**	**M**	**S**	**JD**	**Err**	**FZR**	**FPR**	**SR**	**CR**
10	New	0.028	0.194	90	0.209	0.143	0.015	0	0.34
	PM	1	0	0	0.319	0.313	0.003	0	0.02
	S-LTS	0.002	0.211	98	0.193	0.100	0.039	0	0.48
	PIQ	0.356	0.045	46	0.222	0.233	0.009	0	0.12
20	New	0.070	0.186	76	0.207	0.143	0.017	0	0.38
	PM	1	0	0	0.332	0.323	0.003	0	0.02
	S-LTS	0.002	0.167	96	0.196	0.107	0.041	0	0.50
	PIQ	0.427	0.103	40	0.329	0.280	0.009	0	0.04
30	New	0.103	0.186	60	0.228	0.187	0.018	0	0.32
	PM	1	0	0	0.409	0.407	0.002	0	0
	S-LTS	0.004	0.118	93	0.218	0.130	0.041	0	0.48
	PIQ	0.514	0.173	30	0.496	0.347	0.010	0	0
40	New	0.108	0.202	64	0.379	0.243	0.021	0	0.26
	PM	1	0.0003	0	0.407	0.423	0.003	0	0
	S-LTS	0.096	0.087	86	0.375	0.190	0.046	0	0.36
	PIQ	0.395	0.224	40	0.663	0.377	0.010	0	0

Table 4.3: The Summary of the simulation study with $n = 200$ and $p_n = 1000$ with L=5 under setting II.

O	**method**	**M**	**S**	**JD**	**Err**	**FZR**	**FPR**	**SR**	**CR**
10	New	0.002	0.197	98	0.241	0	0.011	0	1
	PM	1	0	0	0.936	0.295	0.002	0.02	0.12
	S-LTS	0	0.211	100	0.225	0	0.028	0	1
	PIQ	0.088	0.031	75	0.151	0.060	0.002	0	0.78
20	New	0.011	0.182	85	0.272	0	0.013	0	1
	PM	1	0	0	1.124	0.390	0.002	0	0.08
	S-LTS	0	0.167	100	0.263	0	0.032	0	1
	PIQ	0.140	0.071	70	0.241	0.105	0.002	0	0.63
30	New	0.055	0.187	80	0.323	0.031	0.014	0	0.90
	PM	1	0	0	1.001	0.330	0.003	0	0.06
	S-LTS	0.067	0.129	93	0.315	0.013	0.031	0	0.95
	PIQ	0.128	0.105	67	0.423	0.185	0.003	0	0.41
40	New	0.223	0.255	42	0.756	0.175	0.011	0	0.58
	PM	1	0	0	1.163	0.405	0.003	0	0.08
	S-LTS	0.402	0.163	55	0.732	0.120	0.023	0	0.66
	PIQ	0.110	0.152	54	0.772	0.308	0.003	0	0.17

Table 4.4: The Summary of the simulation study with $n = 200$ and $p_n = 1000$ with L=10 under setting II.

O	method	M	S	JD	Err	FZR	FPR	SR	CR
10	New	0.082	0.207	80	0.265	0	0.009	0.04	1
	PM	0.998	0.0001	0	0.450	0.125	0.002	0.18	0.58
	S-LTS	0.004	0.211	95	0.243	0	0.025	0.02	1
	PIQ	0.208	0.041	63	0.172	0.078	0.002	0	0.70
20	New	0.058	0.186	78	0.255	0.010	0.010	0.02	0.96
	PM	1	0	0	0.641	0.150	0.001	0.08	0.46
	S-LTS	0.002	0.167	94	0.240	0.010	0.027	0	0.96
	PIQ	0.231	0.090	61	0.344	0.168	0.003	0	0.42
30	New	0.068	0.199	74	0.305	0.031	0.011	0	0.95
	PM	1	0	0	0.519	0.140	0.002	0.14	0.48
	S-LTS	0.025	0.122	90	0.294	0.025	0.027	0	0.95
	PIQ	0.178	0.135	57	0.489	0.258	0.003	0	0.23
40	New	0.133	0.195	65	0.430	0.087	0.014	0	0.88
	PM	1	0	0	0.527	0.140	0.002	0.12	0.56
	S-LTS	0.118	0.092	87	0.431	0.075	0.030	0	0.88
	PIQ	0.167	0.167	42	0.878	0.365	0.003	0	0.11

values. Furthermore, it is also revealed that the number of outliers cast a huge influence on the performance of these methods under investigation. For instance, when the number of outliers increased, the values of both masking and swamping increased accordingly. We also find that under both settings of interest, $L = 5$ and $L = 10$, PM recorded the worst performance.

Another finding is that the dimensions of the regression vector also affected the estimators' performance. For example, when $L = 5$ and p_n increased from 500 to 1000, we observe an improvement of PIQ, PM and the new estimator in terms of M, S and JD. On the other hand, the S-LTS estimator and the new estimator enjoyed very small masking when O was relatively small.

4.4 Real data application

In this section, we investigated the colon tumor dataset. Colon cancer is a substantial public health problem. Throughout the years, we have witnessed a quick population

growth of this cancer on a global level. The colon tumor dataset contains 62 samples of colon epithelial cells from colon cancer patients. This dataset consists of tumor bipsies collected from tumors, and normal biopsies collected from the healthy part of the colons of the same patient, which results in 40 tumors samples and 22 normal colon tissues samples, respectively. The gene expressions in the colon tumor dataset were analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes by Alon et al. (1999). They reduced the dimension to 2000.

Since no outliers were specified in this colon tumor dataset, we artificially injected outliers into the dataset. We began by creating the following two sets $O = \{10, 15, 20\}$ and $L = \{2000, 20000, 25000\}$, where set O contains the number of rows we planned to modify and L contains the values of the leverage points. We investigated different combinations of O and L, thus rendering in 9 different scenarios. Considering the computational complexity given the large dimension, the tuning parameter was set as $4\sqrt{\log p_n/n}$. Empirical results of the colon tumor dataset are presented in Table 4.5. Note that when presenting the results, we use a three-dimensional vector to represent the values of masking, swamping, and the number of nonzero components identified respectively. The $*$ denotes the values that are not applicable.

Table 4.5: The summary of real data application.

Scenarios	**New**	**S-LTS**	**PM**
(0, 0)	($*$, $*$, 0)	($*$, $*$, 15)	($*$, $*$, 0)
(10, 2000)	(0.5, 0, 5)	(0.5, 0.192, 15)	(1, 0, 0)
(15, 2000)	(0.467, 0, 8)	(0.467, 0.149, 15)	(0.467, 0, 8)
(20, 2000)	(0.5, 0, 10)	(0.5, 0.119, 15)	(0.5, 0, 10)
(10, 20000)	(0.5, 0, 5)	(0, 0.096, 15)	(1, 0, 0)
(15, 20000)	(0.467, 0, 8)	(0, 0, 15)	(0.533, 0, 7)
(20, 20000)	(0.5, 0, 10)	(0.5, 0.119, 15)	(1, 0, 0)
(10, 25000)	(0.5, 0, 5)	(0, 0.096, 15)	(1, 0, 0)
(15, 25000)	(0.467, 0, 8)	(0, 0, 15)	(1, 0, 0)
(20, 25000)	(0.5, 0, 10)	(0.85, 0.286, 15)	(0.5, 0, 10)

According to Table 4.5, we find that when the leverage value was relatively small, for example 2000, all the methods under investigation suffered from either masking or swamping. Although, the new method suffered from masking, yet enjoyed good performance with regard to swamping. We also observe that the new method tended to select fewer observations as outliers when compared with the S-LTS method.

4.5 Conclusions

This chapter studies the parameter estimation and outlier detection for the linear regression model in a high-dimensional setting. We investigate the multivariate linear regression model. In the multivariate linear regression model, outliers are mainly introduced by errors with heavy-tailed distributions. To tackle this problem, we propose a new method that can detect the presence of outliers and perform parameter estimation. The theoretical results show that our new method can accurately estimate the regression vector under certain conditions. To illustrate our methodology, we carried out extensive simulations and real data analysis. From the empirical results, we are able to show that our method is robust against outliers and estimates the parameter accurately by producing small l_2 errors.

Chapter 5

Conclusions

In this thesis, two important high- and ultrahigh-dimensional statistical topics are studied:

- Parameter estimation and model selection under the high-dimensional linear regression model;
- Outlier detection and parameter estimation under the high-dimensional linear regression model.

The following is an outline of the main results, the limitations that the proposed methods have, and the possible future directions in high- and ultrahigh-dimensional statistics.

In Chapters 2 and 3, a two-step procedure is proposed on a class of loss functions. Two estimators are derived based on this two-step procedure. First: the generalized adaptive robust regression estimator. We prove that the proposed estimator can handle the outliers and leverage points with high-dimensional data under certain conditions. Also, this procedure is applicable for ultrahigh-dimensional datasets. The second estimator proposed is l_q robust regression estimator. It is shown that this new estimator still possesses the estimation consistency and oracle properties. Simulation results demonstrate that methods using adaptive weights in the penalty functions can lead to more accurate estimation. However, the empirical results also reveal the limitations of our methods. The problems may be resolved by either applying our methods to datasets with different sample sizes and dimensions or choosing a different initial estimator.

In Chapter 4, we investigate the problem of robust estimation and outlier detection simultaneously in the linear regression model under a high-dimensional setting. We present the estimation consistency of the proposed method. Our simulation results show that our estimator enjoys good performance with regard to masking, swamping, and parameter estimation; however, it has limitations when compared with the S-LTS method. We believe that the new method is not sufficient in dealing with situations where outliers are present in certain high-dimensional settings and further research is needed in this area.

Due to the novelty of the topics investigated in this thesis, we hope that further research will follow from this thesis.

Bibliography

[1] Alain, V., Agathe, G., Stéphane, G., and Malgorzata, B. (2017). High-dimensional robust regression and outliers detection with slope. *arXiv:1712.02640*.

[2] Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics,* **7**, 226–248.

[3] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America,* **96**, 6745–6750.

[4] Andrea, C. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association,* **105**, 147–156.

[5] Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics and Data Analysis,* **56**, 1952–1965.

[6] Bai, Z. D., Rao, C. R., and Wu, Y. (1992). M-estimation of multivariate linear regression parameter under a convex discrepancy function. *Statistica Sinica,* **2**, 237–254.

[7] Belloni, A. and Chernozhukov, V. (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics,* **39**, 82–130.

[8] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics,* **37,** 1705–1732.

[9] Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *Journal of the Royal Statistical Society Series* **B**, **73**, 325–349.

[10] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization,* Cambridge University Press.

[11] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications,* Springer, Heidelberg.

[12] Charbonnier, P., Blanc-Féraud, L., Aubert, G., and Barlaud, M. (1997). Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing,* **6**, 298–311.

[13] Chen, X., Wang, J., and McKeown, A. (2010). Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transactions on Information Theory,* **56**, 5131–5149.

[14] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics,* **32**, 407–451.

[15] Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics,* **42**, 324–351.

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

[17] Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high-dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society Series* **B**, **79**, 247–265.

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

[19] Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics,* **35**, 109–135.

[20] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh-dimensional feature space. *Journal of the Royal Statistical Society Series* **B**, **70**, 849–911.

[21] Gannaz, I. (2006). Robust estimation and wavelet thresholding in partial linear models. Technical report, University Joseph Fourier, Grenoble, France.

[22] Gao, X. and Feng, Y. (2018). Penalized weighted least absolute deviation regression. *Statistics and its Interface,* **11**, 79–89.

[23] Geman, D. and Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing,* **4**, 932–946.

[24] Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics,* **30**, 583–616.

[25] Geman, S. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **14**, 367—383.

[26] Giloni, A., Simonoff, J. S., and Sengupta, B. (2006). Robust weighted LAD regression. *Computational Statistics and Data Analysis,* **50**, 3124–3140.

[27] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley, New York.

[28] Huang, C. C., Liu, K., Pope, R. M., Du, P., Lin, S., Rajamannan, N. M., Huang, Q. Q., Jafar, N., Burke, G. L., Post, W., Watson, K. E., Johnson, C., Daviglus, M. L., and Lloyd-Jones, D. M. (2011). Activated TLR signaling in atherosclerosis among

women with lower Framingham risk score: The multi-ethnic study of atherosclerosis. *PLoS One,* **6**, e21067.

[29] Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics,* **35**, 73–101.

[30] Huber, P. J. (1981). *Robust Statistics,* Wiley, New York.

[31] Idier, J. (2001). Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Transactions on Image Processing,* **10**, 1001–1009.

[32] Jiang, Y., Wang, Y., Zhang, J., Xie, B., Liao, J., and Liao, W. (2020). Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method. *Journal of Applied Statistics,* **48**, 234–246.

[33] Johnson, B. and Peng, L. (2008). Rank-based variable selection. *Journal of Nonparametrics Statistics,* **20**, 241–252.

[34] Kai, B., Li, R., and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics,* **39**, 305–332.

[35] Karunamuni, R. J., Kong, L., and Tu, W. (2019). Efficient robust doubly adaptive regularized regression with applications. *Statistical Methods in Medical Research,* **28**, 2210–2226.

[36] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics,* **28**, 1356–1378.

[37] Kong, D., Bondell, H., and Wu, Y. (2018). Fully efficient robust estimation, outlier detection, and variable selection via penalized regression. *Statistica Sinica,* **28**, 1031–1052.

[38] Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive Lasso penalty. *Electronic Journal of Statistics,* **5**, 1015–1053.

[39] Lee, Y., MacEachern, S. N., and Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science,* **27**, 350–372.

[40] Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica,* **20**, 167–181.

[41] Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica,* **21**, 391–419.

[42] Libby, P. (2002). Inflammation in atherosclerosis. *Nature,* **420**, 868–874.

[43] Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics,* **40**, 1637–1664.

[44] Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics,* **45**, 866–896.

[45] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics,* **37**, 3498–3528.

[46] Loh, P.-L. (2018). Scale calibration for high-dimensional robust regression. *arXiv:1811.02096.*

[47] McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics and Data Analysis,* **52**, 249–257.

[48] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series* **B**, **72**, 417–473.

[49] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science,* **27**, 538–557.

[50] Nikolova, M. and Ng, M. K. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific Computing,* **27**, 937–966.

[51] Nguyen, N. H. and Tran, T. D. (2012). Robust lasso with missing and grossly corrupted observations. *IEEE Transactions on Information Theory,* **59**, 2036–2058.

[52] Öllerer, V., Croux, C., and Alfons, A. (2015). The influence function of penalized regression estimators. *Statistics,* **49**, 741–765.

[53] She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika,* **104**, 633–647.

[54] She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association,* **106**, 626–639.

[55] She, Y., Wang, Z., and Shen, J. (2021). Gaining outlier resistance with progressive quantiles: fast algorithms and theoretical studies. *Journal of the American Statistical Association,* to appear.

[56] Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression model. *Computational Statistics and Data Analysis,* **111**, 116–130.

[57] Sun, Q., Zhou, W., and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association,* **115**, 254–265.

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

[59] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics,* Springer, New York.

[60] van der Geer, S. and Müller, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science,* **27**, 469–480.

[61] Wang, L. (2013). The l_1 penalized LAD estimator for high-dimensional linear regression. *Journal of Multivariate Analysis,* **120**, 135–151.

[62] Wang, L. and Li, R. (2009). Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics,* **65**, 564–571.

[63] Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics,* **25**, 347–355.

[64] Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association,* **107**, 214–222.

[65] Wagener, J. and Dette, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics,* **22**, 137–154.

[66] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory,* **55**, 2183–2202.

[67] Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association,* **108**, 632–643.

[68] Weisberg, S. (1985). *Applied Linear Regression (2nd ed.),* Wiley, New York.

[69] Wright, J., Yang, A. Y., Ganesh, A. S., Sastry, S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **31**, 210–227.

[70] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics,* **19**, 1636–1643.

[71] Wu, W. B. (2007). M-estimation of linear models with dependent errors. *The Annals of Statistics,* **35**, 495–521.

[72] Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica,* **19**, 801–817.

[73] Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics,* **15**, 642–656.

[74] Yohai, V. J. and Zamar, R. H. (1993). A minimax-bias property of the least α-quantile estimates. *The Annals of Statistics,* **21**, 1824–1842.

[75] Yu, J. S., Ongarello, S., Fiedler, R., Chen, X. W., Toffolo, G., Cobelli, C., and Trajanoski, Z. (2005). Ovarian cancer identification based on dimensionality reduction for high throughput mass spectrometry data. *Bioinformatics,* **21**, 2200–2209.

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

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

[78] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics,* **36**, 1108–1126.

[79] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series* **B**, **67**, 301–320.

[80] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics,* **37**, 1733–1751.

Appendix A: Supplementary material for Chapter 2

A.1: Proofs of theorems in Chapter 2

Proof of Theorem 2.1.

For a given $\mu > 0$, define the set

$$\mathbb{B}(\mu) = \{\boldsymbol{\beta} =(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \mu,\ \text{supp}(\boldsymbol{\beta}) \subseteq \text{supp}(\boldsymbol{\beta}^*)\},$$

where $\text{supp}(\boldsymbol{\beta}^*) = \{1, ..., k_n\}$. Denote

$$I = \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta}^*)/\gamma\right) W(\mathbf{X}_i/\gamma).$$

We first show that for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T \in \mathbb{B}(\mu)$,

$$\mathbf{E}(I) \geq cn \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2, \tag{A.1}$$

for sufficiently large n and when μ is properly chosen, where $c > 0$ is some constant.

Write $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + \boldsymbol{\kappa}$, where $\|\boldsymbol{\kappa}\|_2 = \mu$. Then note that

$$I = \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T(\boldsymbol{\beta}_1^* + \boldsymbol{\kappa})\right)/\gamma\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1^*\right)/\gamma\right) W(\mathbf{X}_i/\gamma)$$

$$= \sum_{i=1}^{n} \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right)\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right) dt$$

$$- \frac{1}{\gamma}\sum_{i=1}^{n} W\left(\mathbf{X}_i/\gamma\right)\phi^{(1)}\left(e_i/\gamma\right)\mathbf{X}_{1i}^T\boldsymbol{\kappa}.$$

As the errors e_i's are independent of the covariates $\mathbf{X}_i$'s and from condition (C2), we obtain

$$\mathbf{E}(I) = \mathbf{E}\left(\sum_{i=1}^{n} \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right)\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right) dt\right).$$

Again using (C2), we have

$$\begin{aligned}
\mathbf{E}\left(I\right) &= \sum_{i=1}^{n} \mathbf{E}\left(\frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right) \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \mathbf{E}\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right)\right) dt\right) \\
&= \sum_{i=1}^{n} \mathbf{E}\left(\frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right) \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \left(g(\gamma)t + o\left(|t|q(\gamma)\right)\right) dt\right) \qquad \text{(A.2)} \\
&= \frac{g(\gamma)}{2\gamma}\boldsymbol{\kappa}^T\mathbf{E}\left(\sum_{i=1}^{n} \mathbf{X}_{1i}\mathbf{X}_{1i}^T W\left(\mathbf{X}_i/\gamma\right)\right)\boldsymbol{\kappa} + o(1)\frac{q(\gamma)}{2\gamma} n\|\boldsymbol{\kappa}\|_2^2.
\end{aligned}$$

By condition (C4), we have $\boldsymbol{\kappa}^T\mathbf{E}\left(\sum_{i=1}^{n} \mathbf{X}_{1i}\mathbf{X}_{1i}^T W\left(\mathbf{X}_i/\gamma\right)\right)\boldsymbol{\kappa} \geq cn\|\boldsymbol{\kappa}\|_2^2$ for some constant $c > 0$. Further $\|\boldsymbol{\kappa}\|_2 = \mu$. Therefore, (A.1) now follows from (A.2).

Although (A.1) holds for any vector $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \mathbf{0}^T\right)^T \in \mathbb{B}(\mu)$, it is possible that $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T\right)^T$ may not be in the set $\mathbb{B}(\mu)$. Therefore, consider a new vector $\tilde{\boldsymbol{\beta}}^* = \left((\tilde{\boldsymbol{\beta}}_1^*)^T, \mathbf{0}^T\right)^T$, with

$$\tilde{\boldsymbol{\beta}}_1^* = M\hat{\boldsymbol{\beta}}_1 + (1 - M)\boldsymbol{\beta}_1^*,$$

where $M = \mu/(\mu + \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2)$. It is easy to verify that $\tilde{\boldsymbol{\beta}}^* \in \mathbb{B}(\mu)$. Then using the convexity of the objective function $Q_n(\boldsymbol{\beta})$ defined by (2.4) and the definition of $\hat{\boldsymbol{\beta}}_1$, we have

$$Q_n(\tilde{\boldsymbol{\beta}}^*) \leq MQ_n(\hat{\boldsymbol{\beta}}_1, \mathbf{0}) + (1 - M)Q_n(\boldsymbol{\beta}_1^*, \mathbf{0}) \leq Q_n(\boldsymbol{\beta}^*).$$

Denote $\upsilon_n(\boldsymbol{\beta}) = \sum_{i=1}^n \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma)$. Using the preceding inequality and the triangle inequality, we obtain

$$\begin{aligned}\mathbf{E}\left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*) - \upsilon_n(\boldsymbol{\beta}^*)\right) &= (\upsilon_n(\boldsymbol{\beta}^*) - \mathbf{E}\upsilon_n(\boldsymbol{\beta}^*)) - \left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*) - \mathbf{E}\upsilon_n(\tilde{\boldsymbol{\beta}}^*)\right) \\ &\quad + Q_n(\tilde{\boldsymbol{\beta}}^*) - Q_n(\boldsymbol{\beta}^*) + n\lambda_n\|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 - n\lambda_n\|\mathbf{d}_0 \circ \tilde{\boldsymbol{\beta}}_1^*\|_1 \\ &\leq nZ_n(\mu) + n\lambda_n\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* - \tilde{\boldsymbol{\beta}}_1^*)\|_1,\end{aligned}$$

where $Z_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)} \frac{1}{n}\left|\upsilon_n(\boldsymbol{\beta}) - \upsilon_n(\boldsymbol{\beta}^*) - \mathbf{E}(\upsilon_n(\boldsymbol{\beta}) - \upsilon_n(\boldsymbol{\beta}^*))\right|$. By the Cauchy-Schwarz inequality, the second term is bounded by $n\lambda_n\|\mathbf{d}_0\|_2\mu$, and hence we obtain

$$\mathbf{E}\left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*) - \upsilon_n(\boldsymbol{\beta}^*)\right) \leq nZ_n(\mu) + n\lambda_n\|\mathbf{d}_0\|_2\mu. \tag{A.3}$$

Define $A_n = \{Z_n(\mu) \leq 2\mu\sqrt{k_n(\log n)/n}\}$. By Lemma 2.2, it follows that $\mathbf{P}(A_n) \to 1$ as $n \to \infty$. Now combining (A.1) and (A.3), on A_n we have

$$cn\|\tilde{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1^*\|_2^2 \leq 2\mu\sqrt{k_n n\log n} + n\lambda_n\|\mathbf{d}_0\|_2\mu.$$

Take $\mu = O(\sqrt{k_n/n} + \lambda_n\|\mathbf{d}_0\|_2)$. Then by conditions in Theorem 2.1, $\mu \to 0$. The above inequality then yields that

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1^*\|_2^2 &\leq O\left(\left(\sqrt{k_n(\log n)/n} + \lambda_n\|\mathbf{d}_0\|_2\right)\left(\sqrt{k_n/n} + \lambda_n\|\mathbf{d}_0\|_2\right)\right), \\ &\leq O\left(\sqrt{k_n(\log n)/n} + \lambda_n\|\mathbf{d}_0\|_2\right)^2. \end{aligned}\tag{A.4}$$

Since $\|\boldsymbol{\beta}_1^* - \tilde{\boldsymbol{\beta}}_1^*\|_2 \leq O(\mu)$ implies $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq O(\mu)$, from (A.4) it follows that on the event A_n,

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq O\left(\sqrt{k_n(\log n)/n} + \lambda_n\|\mathbf{d}_0\|_2\right).$$

A bound for the l_1-loss follows from the inequality $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq \sqrt{k_n}\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2$. This completes the proof of the first part of Theorem 2.1. The second part follows trivially.

Proof of Theorem 2.2.

From Theorem 2.1, $\hat{\boldsymbol{\beta}}_1$ is a minimizer of $Q_n(\boldsymbol{\beta}_1, \mathbf{0})$, so it satisfies the Karush-Kuhn-Tuker (KKT) conditions in optimization theory. Thus, in order to prove that $\hat{\boldsymbol{\beta}} =$

$\left(\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T\right)^T$ is a global minimizer of $Q_n(\boldsymbol{\beta})$ in the space $\mathbb{R}^{p_n}$, we only need to show that the following condition is satisfied:

$$\left\|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T \Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\hat{\boldsymbol{\beta}}\right)\right\|_\infty < n\lambda_n, \tag{A.5}$$

where $\Psi(\mathbf{Y} - \mathfrak{X}\hat{\boldsymbol{\beta}}) = (\phi((Y_1 - \mathbf{X}_1^T\hat{\boldsymbol{\beta}})/\gamma)W(\mathbf{X}_1/\gamma), ..., \phi((Y_n - \mathbf{X}_n^T\hat{\boldsymbol{\beta}})/\gamma)W(\mathbf{X}_n/\gamma))^T$, $\mathbf{d}_1^{-1} = \left(d_{k_n+1}^{-1}, \ldots, d_{p_n}^{-1}\right)^T$, and $\mathbf{Q}$ is the submatrix containing the noise covariates. Note that $Y_i - \mathbf{X}_i^T\boldsymbol{\beta} = Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1$ when $\boldsymbol{\beta}_2 = \mathbf{0}$ for all i. Then the KKT conditions together with the convexity of $Q_n(\boldsymbol{\beta})$ verify that $\hat{\boldsymbol{\beta}}$ is a global minimizer of

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^n \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) + n\lambda_n\|\mathbf{d} \circ \boldsymbol{\beta}\|_1.$$

Consider the events

$$N_1 = \{\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq s_n\},$$

and

$$N_2 = \{\sup_{\boldsymbol{\beta}\in\mathbb{N}} \|\boldsymbol{d}_1^{-1} \circ \mathbf{Q}^T \Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right)\|_\infty < n\lambda_n\},$$

where $\Psi(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}) = (\phi((Y_1 - \mathbf{X}_1^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_1/\gamma), ..., \phi((Y_n - \mathbf{X}_n^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_n/\gamma))^T$ and $\Psi^{(1)}$ is the first derivative of Ψ. Also define the set

$$\mathbb{N} = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0},\ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq s_n\},$$

where s_n is as defined in Theorem 2.1. Then from Theorem 2.1 and Lemma 2.3, it follows that $P\left(N_1 \cap N_2\right) \to 1$. Since $\hat{\boldsymbol{\beta}} \in \mathbb{N}$ on the event N_1, (A.5) holds on $N_1 \cap N_2$. Again, a bound for the l_1-loss follows from the inequality $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq \sqrt{k_n}\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2$. This completes the proof.

Proof of Theorem 2.3.

Define $G_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\beta}_1, \mathbf{0}) - Q_n(\boldsymbol{\beta}_1^*, \mathbf{0})$, where $Q_n(\boldsymbol{\beta})$ is defined by (2.4) and $\boldsymbol{\theta} = \mathbf{V}_n^{-1}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)$ with $\mathbf{V}_n = \left(\mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\right)^{-1/2}$. Then $G_n(\boldsymbol{\theta})$ can be

written as

$$G_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1\right)/\gamma\right) W(\mathbf{X}_i/\gamma)$$
$$- \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1^*\right)/\gamma\right) W\left(\mathbf{X}_i/\gamma\right)$$
$$+ n\lambda_n\left(\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta})\|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1\right).$$

We have shown in Theorem 2.1 that $Q_n(\boldsymbol{\beta}, \mathbf{0})$ is minimized at $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_1$ and, therefore, $G_n(\boldsymbol{\theta})$ is minimized at $\hat{\boldsymbol{\theta}} = \mathbf{V}_n^{-1}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right)$. Define a convex open set

$$\mathbb{A}_n = \{\boldsymbol{\theta} \in \mathbb{R}^{k_n} : \|\boldsymbol{\theta}\|_2 < c_4\sqrt{k_n}\},$$

where $c_4 > 0$ is some constant independent of k_n. We divide $G_n(\boldsymbol{\theta})$ into two components as follows:

$$G_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}) + T_n(\boldsymbol{\theta}), \tag{A.6}$$

where $L_n(\boldsymbol{\theta}) = E[G_n(\boldsymbol{\theta})]$ and $T_n(\boldsymbol{\theta})$ is the centralized stochastic process defined by

$$T_n(\boldsymbol{\theta}) = S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*) - \mathbf{E}\left(S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*)\right),$$

where $S_n(\boldsymbol{\beta}_1) = \sum_{i=1}^n \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1\right)/\gamma\right) W\left(\mathbf{X}_i/\gamma\right)$.

We first examine the mean component $L_n(\boldsymbol{\theta})$. Using the method as in (A.2), we obtain

$$\mathbf{E}\left(S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*)\right)$$
$$= \sum_{i=1}^{n} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)^T \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$$
$$+ O(1)\left(\sum_{i=1}^{n} \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}W(\mathbf{X}_i/\gamma)\right)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2+\varepsilon}\right), \tag{A.7}$$

for some constant $0 < \varepsilon < 1$. The first term on the RHS of (A.7) is equal to $\|\boldsymbol{\theta}\|_2^2$. By condition (C4), for any $\boldsymbol{\theta} \in \mathbb{A}_n$ we have $\|\mathbf{V}_n\boldsymbol{\theta}\|_2 \le c_1 n^{-1/2}\|\boldsymbol{\theta}\|_2 \le c_2 n^{-1/2} k_n^{1/2}$ for some

constants $c_1 > 0$ and $c_2 > 0$. Then by the Cauchy-Schwarz inequality, for any $\boldsymbol{\theta} \in \mathbb{A}_n$,

$$\begin{aligned}\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right| &\leq \|\mathbf{X}_{1i}\|_2 \|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\|_2 \\ &= \|\mathbf{X}_{1i}\|_2 \|\mathbf{V}_n \boldsymbol{\theta}\|_2 \\ &\leq \|\mathbf{X}_{1i}\|_2 C n^{-1/2} k_n^{1/2}.\end{aligned}$$

By condition (C3), it follows that $W(\mathbf{X}_i/\gamma)\|\mathbf{X}_{1i}\|_2^{2+\varepsilon} \leq c_3 k_n^{(2+\varepsilon)/2}$ for some constant $c_3 > 0$. By combining above facts, for any $\boldsymbol{\theta} \in \mathbb{A}_n$ we obtain

$$\begin{aligned}\sum_{i=1}^n \mathbf{E}\left(W(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2+\varepsilon}\right) &= O(n k_n^{(2+\varepsilon)/2} n^{-(2+\varepsilon)/2} k_n^{(2+\varepsilon)/2}) \\ &= O(n^{-\varepsilon/2} k_n^{2+\varepsilon}) \\ &= o(1), \qquad \text{(A.8)}\end{aligned}$$

provided $k_n = o(n^{\varepsilon/(4+2\varepsilon)})$. Therefore, from (A.7) and (A.8) it follows that

$$\mathbf{E}\left(S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*)\right) = \|\boldsymbol{\theta}\|_2^2 + o(1). \qquad \text{(A.9)}$$

By condition (C4), $\mathbf{V}_n$ has bounded eigenvalues. Thus, for any $\boldsymbol{\theta} \in \mathbb{A}_n$ we have

$$\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta})\|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 = \tilde{\mathbf{d}}_0^T \mathbf{V}_n \boldsymbol{\theta}, \qquad \text{(A.10)}$$

where $\tilde{\mathbf{d}}_0$ is a k_n-dimensional vector with i-th component as $d_i \operatorname{sgn}(\beta_j^*)$ and $\operatorname{sgn}(\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}) = \operatorname{sgn}(\boldsymbol{\beta}_1^*)$. Now combining (A.9) and (A.10) we obtain

$$L_n(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2 + n\lambda_n \tilde{\mathbf{d}}_0^T \mathbf{V}_n \boldsymbol{\theta} + o(1), \qquad \text{(A.11)}$$

uniformly over all $\boldsymbol{\theta} \in \mathbb{A}_n$.

Now let us consider the stochastic component $T_n(\boldsymbol{\theta})$ defined in (A.6). In addition, we define $\mathbf{D} = -\left(\phi^{(1)}(e_1/\gamma), \ldots, \phi^{(1)}(e_n/\gamma)\right)^T$ and $\mathbf{U}_n = (\mathbf{Z}_n^*)^T \mathbf{D}$, where $\mathbf{Z}_n^*$ is defined in the beginning of Section 2.2. Let

$$\Omega_n(\boldsymbol{\theta}) = \sum_{i=1}^n \phi\left(\left(e_i - \mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right)/\gamma\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^n \phi(e_i/\gamma) W(\mathbf{X}_i/\gamma) - \mathbf{U}_n^T\boldsymbol{\theta}.$$

It is easy to check that $\mathbf{E}\left(\mathbf{U}_n^T\boldsymbol{\theta}\right) = 0$. Therefore, we can write $T_n(\boldsymbol{\theta})$ as

$$T_n(\boldsymbol{\theta}) = \mathbf{U}_n^T\boldsymbol{\theta} + h_n(\boldsymbol{\theta}), \tag{A.12}$$

where $h_n(\boldsymbol{\theta}) = \Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)$. By Lemma 2.1, we know that for any $\varepsilon > 0$,

$$\mathbf{P}\left(|h_n(\boldsymbol{\theta})| > \varepsilon\right) \leq \exp\left(-C\varepsilon a_n k_n\right),$$

for some positive constant C and a sequence $a_n \to \infty$. Define $H_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}) - n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} - \mathbf{U}_n^T\boldsymbol{\theta}$. By the definition of $G_n(\boldsymbol{\theta})$, $H_n(\boldsymbol{\theta})$ and $\|\boldsymbol{\theta}\|_2^2$ are convex functions in $\boldsymbol{\theta}$. In addition, $h_n(\boldsymbol{\theta})$ can be written as

$$h_n(\boldsymbol{\theta}) = H_n(\boldsymbol{\theta}) - \|\boldsymbol{\theta}\|_2^2 - o(1).$$

It is easy to show that for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathbb{A}_n$,

$$\left|\|\boldsymbol{\theta}_1\|_2^2 - \|\boldsymbol{\theta}_2\|_2^2\right| = \left|\left(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\right)^T\left(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right)\right| \leq O\left(k_n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty\right).$$

Therefore, all the conditions needed for Lemma 4 in Fan et al. (2014) are satisfied. Now applying their lemma, for any compact set $A_{k_n} = \{\|\boldsymbol{\theta}\|_2 \leq c_5\sqrt{k_n}\} \subset \mathbb{A}_n$ with $c_5 < c_4$, we have

$$\sup_{\boldsymbol{\theta}\in A_{k_n}} |h_n(\boldsymbol{\theta})| = o_p(1). \tag{A.13}$$

Now combining (A.6), (A.11) and (A.12), we obtain

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \|\boldsymbol{\theta}\|_2^2 + n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} + \mathbf{U}_n^T\boldsymbol{\theta} + h_n(\boldsymbol{\theta}) + o(1) \\ &= \|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2^2 - \|\boldsymbol{\zeta}_n\|_2^2 + h_n(\boldsymbol{\theta}) + o(1), \end{aligned} \tag{A.14}$$

where

$$\boldsymbol{\zeta}_n = -\frac{1}{2}\left(n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0 + \mathbf{U}_n\right).$$

By condition (C2), $\mathbf{Var}\left(\phi^{(1)}(e_i/\gamma)\right) = \sigma_\gamma^2$. Then using a classical weak convergence argument we have

$$\mathbf{c}^T\left(\left(\mathbf{Z}_n^*\right)^T\mathbf{Z}_n^*\right)^{-1/2}\mathbf{U}_n \xrightarrow{\mathcal{D}} N\left(0, \sigma_\gamma^2\right),$$

for any vector $\mathbf{c} \in \mathbb{R}^{k_n}$ such that $\mathbf{c}^T\mathbf{c} = 1$. It then follows that

$$\mathbf{c}^T \left(\left(\mathbf{Z}_n^*\right)^T \mathbf{Z}_n^* \right)^{-1/2} \left(2\boldsymbol{\zeta}_n + n\lambda_n \mathbf{V}_n \tilde{\mathbf{d}}_0 \right) \xrightarrow{\mathcal{D}} N\left(0, \sigma_\gamma^2\right). \tag{A.15}$$

It remains to show that for any $\iota > 0$,

$$\mathbf{P}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\zeta}_n\|_2 > \iota \right) \to 0,$$

i.e., the minimizer $\hat{\boldsymbol{\theta}}$ of $G_n(\boldsymbol{\theta})$ is close to $\boldsymbol{\zeta}_n$. Theorem 2.3 will then follow from (A.15) and Slutsky's Theorem.

Let $B^*(n)$ denote a ball with center $\boldsymbol{\zeta}_n$ and radius $\iota > 0$. It can be shown that $\frac{\mathbf{U}_n^T \mathbf{U}_n}{\sqrt{k_n}}$ is uniformly tight, i.e., $\mathbf{U}_n^T\mathbf{U}_n = O_P(\sqrt{k_n})$. In addition, $n\mathbf{V}_n\mathbf{V}_n^T$ has bounded eigenvalues. Then using the assumption imposed in the theorem, we obtain

$$\begin{aligned} \|\boldsymbol{\zeta}_n\|_2 &= \left\| \frac{1}{2} \left(n\lambda_n \mathbf{V}_n \tilde{\mathbf{d}}_0 + \mathbf{U}_n \right) \right\|_2 \\ &\leq \frac{1}{2} \left(\|\mathbf{U}_n\|_2 + \|n\lambda_n \mathbf{V}_n \tilde{\mathbf{d}}_0\|_2 \right) \\ &= \frac{c_6\sqrt{k_n}}{2}(1 + O_p(1)), \end{aligned}$$

for some constant $c_6 > 0$. Now choose the constant c_4 in $\mathbb{A}_n$ large enough so that $c_4 > c_6/2$. Then the constant c_5 in A_{k_n} can be chosen large enough to contain $B^*(n)$. Then, by (A.13) we have

$$\Delta_n \doteq \sup_{\boldsymbol{\theta} \in B^*(n)} |h_n(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta} \in A_{k_n}} |h_n(\boldsymbol{\theta})| = o_p(1). \tag{A.16}$$

Let us now consider the behavior of $G_n(\boldsymbol{\theta})$ outside of $B^*(n)$. Let $\boldsymbol{\theta} = \boldsymbol{\zeta}_n + b\mathbf{u} \in \mathbb{R}^{k_n}$, where $\mathbf{u} \in \mathbb{R}^{k_n}$ is a unit vector and $b > \iota > 0$ are some constants. Let $\boldsymbol{\theta}^*$ be the boundary point of $B^*(n)$ that lies on the line segment from $\boldsymbol{\zeta}_n$ to $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}^*$ can be written as $\boldsymbol{\theta}^* = \boldsymbol{\zeta}_n + \iota\mathbf{u} = \left(1 - \frac{\iota}{b}\right)\boldsymbol{\zeta}_n + \frac{\iota}{b}\boldsymbol{\theta}$. Now using the convexity of G_n, definition of Δ_n, (A.14) and (A.16), we obtain

$$\frac{\iota}{b} G_n(\boldsymbol{\theta}) + \left(1 - \frac{\iota}{b}\right) G_n\left(\boldsymbol{\zeta}_n\right) \geq G_n\left(\boldsymbol{\theta}^*\right) \geq \iota^2 - \|\boldsymbol{\zeta}_n\|_2^2 - \Delta_n \geq \iota^2 + G_n\left(\boldsymbol{\zeta}_n\right) - 2\Delta_n.$$

Since $b > \iota$, for large n it follows that

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\zeta}_n\|_2>\iota} G_n(\boldsymbol{\theta}) \geq G_n\left(\boldsymbol{\zeta}_n\right) + \frac{b}{\iota}\left(\iota^2 - o_p(1)\right) > G_n\left(\boldsymbol{\zeta}_n\right).$$

The preceding result suggests that the minimum of $G_n(\boldsymbol{\theta})$ cannot occur at any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2 > \iota$. Hence, with probability tending to one, we have $\|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2 \leq \iota$. This completes the proof.

Proof of Theorem 2.4.

We begin by considering the minimizer of $\hat{Q}_n(\boldsymbol{\beta})$ in the subspace which contains $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T$, with $\boldsymbol{\beta}_2 = \mathbf{0}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1$ with $\tilde{b}_n = \sqrt{k_n \log n/n} + \lambda_n\left(\|\mathbf{d}_0^*\|_2 + c_3^* c_1^* \sqrt{k_n \log p_n/n}\right)$, and $\|\boldsymbol{\nu}_1\|_2 = C$ for some large enough constant $C > 0$. Note that each element of $\boldsymbol{\beta}_1^*$ is the true coefficient corresponding to the important covariates. First note that

$$\hat{Q}_n\left(\boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1, \mathbf{0}\right) - \hat{Q}_n\left(\boldsymbol{\beta}_1^*, \mathbf{0}\right) = L_1(\boldsymbol{\nu}_1) + L_2(\boldsymbol{\nu}_1), \tag{A.17}$$

where

$$L_1(\boldsymbol{\nu}_1) = \sum_{i=1}^{n} \phi\left(\frac{Y_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1\right)}{\gamma}\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^{n} \phi\left(\frac{Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1^*}{\gamma}\right) W(\mathbf{X}_i/\gamma),$$

and

$$L_2(\boldsymbol{\nu}_1) = n\lambda_n\left(\|\hat{\mathbf{d}}_0 \circ (\boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1)\| - \|\hat{\mathbf{d}}_0 \circ \boldsymbol{\beta}_1^*\|\right).$$

Now again consider

$$Z_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)} \frac{1}{n}\left|\upsilon_n(\boldsymbol{\beta}) - \upsilon_n(\boldsymbol{\beta}^*) - \mathbf{E}\left(\upsilon_n(\boldsymbol{\beta}) - \upsilon_n(\boldsymbol{\beta}^*)\right)\right|, \tag{A.18}$$

where $\upsilon_n(\boldsymbol{\beta}) = \sum \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma)$. By Lemma 2.2, the probability of the event $A_n = \{Z_n(\mu) \leq 2\mu\sqrt{k_n(\log n)/n}\}$ goes to one as $n \to \infty$. Therefore, with probability tending to one, it follows that $|L_1(\boldsymbol{\nu}_1) - \mathbf{E}\left(L_1(\boldsymbol{\nu}_1)\right)| \leq nZ_n(\tilde{b}_n) \leq 2\tilde{b}_n\sqrt{k_n n \log(n)}\|\boldsymbol{\nu}_1\|_2$

holds true for sufficiently large $\boldsymbol{\nu}_1$. Furthermore, it is easy to show that

$$\mathbf{E}\left(L_1(\boldsymbol{\nu}_1)\right) = \tilde{b}_n^2 \boldsymbol{\nu}_1^T \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\boldsymbol{\nu}_1 + \frac{q(\gamma)}{2\gamma} n\tilde{b}_n^2\|\boldsymbol{\nu}_1\|_2^2 o(1).$$

Then by the triangle inequality, we have

$$L_1(\boldsymbol{\nu}_1) \geq \tilde{b}_n^2 \boldsymbol{\nu}_1^T \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\boldsymbol{\nu}_1 + \frac{q(\gamma)}{2\gamma} n\tilde{b}_n^2\|\boldsymbol{\nu}_1\|_2^2 o(1) - 2\tilde{b}_n\sqrt{k_n n\log(n)}\|\boldsymbol{\nu}_1\|_2. \quad \text{(A.19)}$$

Using the Cauchy-Schwarz inequality, the second term in (A.17) can be bounded as

$$|L_2(\boldsymbol{\nu}_1)| \leq n\lambda_n\|\hat{\mathbf{d}}_0 \circ (\tilde{b}_n\boldsymbol{\nu}_1)\|_1 \leq n\lambda_n\|\hat{\mathbf{d}}_0\|_2\|\tilde{b}_n\boldsymbol{\nu}_1\|_2. \quad \text{(A.20)}$$

Also, by the triangle inequality, it follows that

$$\|\hat{\mathbf{d}}_0\|_2 \leq \|\hat{\mathbf{d}}_0 - \mathbf{d}_0^*\|_2 + \|\mathbf{d}_0^*\|_2 \leq O\left(\|\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1^*\|_2 + \|\mathbf{d}_0^*\|_2\right) \leq O\left(\sqrt{k_n(\log p_n)/n} + \|\mathbf{d}_0^*\|_2\right).$$

Now combining (A.17), (A.19), (A.20) and the preceding bound, we obtain

$$\hat{Q}_n\left(\boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1, \mathbf{0}\right) - \hat{Q}_n\left(\boldsymbol{\beta}_1^*, \mathbf{0}\right) \geq \tilde{b}_n^2\boldsymbol{\nu}_1^T\mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\boldsymbol{\nu}_1 + \frac{q(\gamma)}{2\gamma}o(1)n\tilde{b}_n^2\|\boldsymbol{\nu}_1\|_2^2 - 2\tilde{b}_n\sqrt{k_n n\log(n)}\|\boldsymbol{\nu}_1\|_2 - n\tilde{b}_n\lambda_n\left(\|\mathbf{d}_0^*\|_2 + O\left(\sqrt{k_n(\log p_n)/n}\right)\right)\|\boldsymbol{\nu}_1\|_2.$$

By condition (C4), the right-hand side of the preceding inequality is larger than

$$C\tilde{b}_n^2 n\|\boldsymbol{\nu}_1\|_2^2 + \frac{q(\gamma)}{2\gamma}o(1)n\tilde{b}_n^2\|\boldsymbol{\nu}_1\|_2^2 - 2\tilde{b}_n\sqrt{k_n(\log n)n}\|\boldsymbol{\nu}_1\|_2 - n\tilde{b}_n\lambda_n\left(\|\mathbf{d}_0^*\|_2 + O\left(\sqrt{k_n(\log p_n)/n}\right)\right)\|\boldsymbol{\nu}_1\|_2 = \tilde{b}_n n\|\boldsymbol{\nu}_1\|_2\left(C\tilde{b}_n\|\boldsymbol{\nu}_1\|_2 + \frac{q(\gamma)}{2\gamma}o(1)\tilde{b}_n\|\boldsymbol{\nu}_1\|_2 - 2\sqrt{k_n(\log n)/n} - \lambda_n\left(\|\mathbf{d}_0^*\|_2 + O\left(\sqrt{k_n(\log p_n)/n}\right)\right)\right),$$

where $C > 0$ is some constant. By making the radius $\|\boldsymbol{\nu}_1\|_2$ sufficiently large, it then follows that with probability tending to one, $\hat{Q}_n\left(\boldsymbol{\beta}_1^* + \tilde{b}_n\boldsymbol{\nu}_1, \mathbf{0}\right) - \hat{Q}_n\left(\boldsymbol{\beta}_1^*, \mathbf{0}\right) > 0$. Then,

with asymptotic probability one, there exists a minimizer $\hat{\boldsymbol{\beta}}_1$ of $\hat{Q}_n(\boldsymbol{\beta}_1, \mathbf{0})$ such that $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq O(\tilde{b}_n)$. In other words, there exists a positive constant c_2^* such that $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq c_2^* \tilde{b}_n = b_n$.

The next step is to show that the estimator $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T\right)^T$ is a global minimizer of $\hat{Q}_n(\boldsymbol{\beta})$. For this purpose, we show that

$$\|(\hat{\mathbf{d}}_1)^{-1} \circ \mathbf{Q}^T \Psi^{(1)}(\mathbf{Y} - \mathfrak{X}\hat{\boldsymbol{\beta}})\|_\infty \leq cn\lambda_n, \tag{A.21}$$

with probability tending to 1, where $\Psi(\cdot)$ is defined in (A.5), $(\hat{\mathbf{d}}_1)^{-1} = \left(\hat{d}_{k_n+1}^{-1}, \ldots, \hat{d}_{p_n}^{-1}\right)^T$, and $c > 0$ is some constant. Then by KKT conditions and (A.21), it can be easily verified that $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T\right)^T$ is a global minimizer of $\hat{Q}_n(\boldsymbol{\beta})$.

For $j = k_n+1, \ldots, p_n$, each entry in $\boldsymbol{\beta}^*$ has $\beta_j^* = 0$, and hence $d_j^* = p_{\lambda_n}^{(1)}(0^+)$. Moreover, by assumptions in Theorem 2.4, the initial estimate $\hat{\beta}_j^*$ satisfies $\left|\hat{\beta}_j^*\right| \leq c_1^* \sqrt{k_n(\log p_n)/n}$. Thus $\min_{j>k_n} p_{\lambda_n}^{(1)}\left(\left|\hat{\beta}_j^*\right|\right) \geq p_{\lambda_n}^{(1)}\left(c_1^* \sqrt{k_n(\log p_n)/n}\right)$. Then by condition (C5), we have

$$\|(\hat{\mathbf{d}}_1)^{-1}\|_\infty = \left(\min_{j>k_n} p_{\lambda_n}^{(1)}\left(\left|\hat{\beta}_j^*\right|\right)\right)^{-1} < 2/p_{\lambda_n}^{(1)}(0^+) = 2\|(\mathbf{d}_1^*)^{-1}\|_\infty < \infty, \tag{A.22}$$

since $\min_{j \geq k_n+1} d_j^*$ is strictly positive. In the proof of Lemma 2.3 below, we show that for any $\boldsymbol{\beta} \in \mathbb{N} = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0},\ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n\}$ with some sequence $v_n \to 0$,

$$\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n} \|\mathbf{Q}^T \Psi^{(1)}(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta})\|_\infty < n\lambda_n(O(1) + o(1))$$

holds with asymptotic probability one. Now let $v_n = b_n$, then we have

$$\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq b_n} \|\mathbf{Q}^T \Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right)\|_\infty < n\lambda_n(O(1) + o(1)) \tag{A.23}$$

holds with asymptotic probability one. Then from (A.22) and (A.23), we conclude that

$$\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq b_n} \|(\hat{\mathbf{d}}_1)^{-1} \circ \mathbf{Q}^T \Psi^{(1)}(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta})\|_\infty < cn\lambda_n, \tag{A.24}$$

with asymptotic probability one, where $c > 0$ is some constant. From (A.24), it now follows that (A.21) holds with probability tending to one because $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq b_n$ holds

with probability tending to one. This completes the proof.

A.2: Proofs of lemmas in Chapter 2

Lemma 2.1. *Assume that conditions of Theorem* 2.3 *hold. Let* $\Omega_n(\boldsymbol{\theta}) = \sum_{i=1}^n \Omega_{ni}(\boldsymbol{\theta})$, *where* $\Omega_{ni}(\boldsymbol{\theta}) = \phi\left((e_i - \mathbf{Z}_{ni}^T\boldsymbol{\theta})/\gamma\right) W\left(\mathbf{X}_i/\gamma\right) - \phi(e_i/\gamma)W(\mathbf{X}_i/\gamma) + \phi^{(1)}(e_i/\gamma)(\mathbf{Z}_{ni}^*)^T\boldsymbol{\theta}$. *Then for any* $\varepsilon > 0$,

$$\mathbf{P}\left(|\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)| > \varepsilon\right) \leq \exp\left(-Ca_nk_n\varepsilon\right),$$

where $a_n > 0$ *is some sequence tending to infinity such that* $a_nk_n^{2\delta+3}n^{-\delta} \to 0$, *and* $C > 0$ *is some constant.*

Proof of Lemma 2.1.

Define $\xi_i = \Omega_{ni}(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right)$. Then we have $\sum_{i=1}^n \xi_i = \Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)$. Applying Markov's inequality yields that for any $\varepsilon > 0$ and $t > 0$,

$$\begin{aligned}\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right) \geq \varepsilon\right) &\leq \exp(-t\varepsilon)\mathbf{E}\left(\exp\left(t\sum_{i=1}^n \xi_i\right)\right) \\ &= \exp\left(-t\varepsilon - t\sum_{i=1}^n \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right)\right)\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right).\end{aligned} \tag{A.25}$$

Using a similar argument as in the proof of (A.9), we obtain

$$t\sum_{i=1}^n \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right) = t\|\boldsymbol{\theta}\|_2^2 + o(1).$$

Then, from (A.25) it follows that

$$\begin{aligned}\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta}) \geq \varepsilon\right)\right) &\leq \exp(-t\varepsilon)\mathbf{E}\left(\exp\left(t\sum_{i=1}^n \xi_i\right)\right) \\ &\propto \exp\left(-t\varepsilon - t\|\boldsymbol{\theta}\|_2^2\right)\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right).\end{aligned} \tag{A.26}$$

By Taylor expansion, we have

$$\begin{aligned}\mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right) &= \mathbf{E}\left(1 + t\Omega_{ni}(\boldsymbol{\theta}) + O\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right) \\ &= 1 + t\mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right) + O\left(\mathbf{E}\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right) \\ &= 1 + \frac{t\|\boldsymbol{\theta}\|_2^2}{n} + O\left(\mathbf{E}\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right).\end{aligned}$$

For $x_i > 0$, we have $\prod_{i=1}^n (1 + x_i) \leq \exp\left(\sum_{i=1}^n x_i\right)$. Using this type of bound, we obtain

$$\begin{aligned}\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right) &\leq \exp\left(\sum_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta}) - 1\right)\right)\right) \\ &= \exp\left(t\|\boldsymbol{\theta}\|_2^2 + O\left(\mathbf{E}\left(\sum_{i=1}^n t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right)\right).\end{aligned}$$

Therefore, from (A.26) we obtain

$$\mathbf{P}\left(\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)\right) > \varepsilon\right) \leq \exp\left(-t\varepsilon + O\left(\mathbf{E}\left(\sum_{i=1}^n t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right)\right).$$

Now let $t = a_n k_n$ with a_n being some diverging sequence such that $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$. Then, it follows that

$$\mathbf{P}\left(\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)\right) > \varepsilon\right) \leq \exp\left(-C\varepsilon k_n a_n\right).$$

It now remains to show that $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$. Using the definition of $\Omega_{ni}(\boldsymbol{\theta})$,

$$\begin{aligned}&\Omega_{ni}^2(\boldsymbol{\theta}) = \\ &\left(\phi\left(\left(Y_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)\right)/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) - \phi\left(e_i/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) + \phi^{(1)}(e_i/\gamma)\left(\mathbf{Z}_{ni}^*\right)^T\boldsymbol{\theta}\right)^2,\end{aligned}$$

and then using the mean value theorem, we obtain

$$\mathbf{E}\left(\Omega_{ni}^2(\boldsymbol{\theta})\right) = \frac{1}{\gamma^2}\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)W(\mathbf{X}_i/\gamma) - \phi^{(1)}(e_i^*/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)W(\mathbf{X}_i/\gamma)\right)^2, \tag{A.27}$$

where e_i^* is a value between e_i and $e_i - \mathbf{Z}_{ni}^T\boldsymbol{\theta} = Y_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)$. Then using the δ-Lipschitz condition of $\phi^{(1)}$ (see condition (C1)), we have

$$\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)-\phi^{(1)}(e_i^*/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right)^2$$
$$\leq c\mathbf{E}\left(\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right|^\delta \mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right)^2 \leq c\mathbf{E}\left(\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right),$$

where $c>0$ is some constant. Therefore, we now have from (A.27) for $t>0$,

$$t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) \leq tc\sum_{i=1}^n \mathbf{E}\left(W^2(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right).$$

Using an argument similar to derive (A.8), we obtain

$$\sum_{i=1}^n \mathbf{E}\left(W^2(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right) = O(n^{-\delta}k_n^{2(\delta+1)}).$$

Then we have $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$ since $t = a_n k_n$ and by the assumptions on k_n in Theorem 2.3. Using a similar argument for $\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right) \leq -\varepsilon\right)$ completes the proof.

Lemma 2.2. *Assume that conditions* (C1)–(C3) *hold. Let* $Z_n(\mu)$ *be defined by* (A.18). *Then, for any* $t>0$ *we have*

$$\mathbf{P}\left(Z_n(\mu) > C_3\mu\sqrt{k_n/n}+t\right) \leq \exp\left(-\frac{nt^2}{8C_1^2\mu^2 k_n}\right),$$

where C_1 *and* C_3 *are some positive constants.*

Proof of Lemma 2.2.

First we write $Z_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)} \frac{1}{n}\left|\sum_{i=1}^n \left(Z_i(\boldsymbol{\beta}) - \mathbf{E}(Z_i(\boldsymbol{\beta}))\right)\right|$, where

$$Z_i(\boldsymbol{\beta}) = \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)W(\mathbf{X}_i/\gamma) - \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta}^*)/\gamma\right)W(\mathbf{X}_i/\gamma).$$

By condition (C1), ϕ satisfies the Lipschitz condition. By (C3), $x_i W(\mathbf{x})$ is bounded for

$x_i \in \mathbf{x} \in \mathbb{R}^{p_n}$. Therefore, for any $\boldsymbol{\beta} \in \mathbb{B}(\mu)$ we have

$$\begin{aligned} |Z_i(\boldsymbol{\beta})| &\leq C_0 \left| W(\mathbf{X}_i/\gamma)\mathbf{X}_i^T(\boldsymbol{\beta}^* - \boldsymbol{\beta})/\gamma \right| \\ &\leq C_1 \mu \sqrt{k_n}, \end{aligned} \tag{A.28}$$

where C_0 and C_1 are some positive constants. Then for any $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \mathbf{0}^T\right)^T \in \mathbb{B}(\mu)$, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i(\boldsymbol{\beta}))^2 \leq \left(C_1 \mu \sqrt{k_n}\right)^2. \tag{A.29}$$

Let $D_1, D_2, \ldots, D_n$ be a Rademacher sequence, independent of $Z_1, Z_2, \ldots, Z_n$. Then using (A.28) together with the symmetrization and the contraction inequality (see, e.g., Theorems 14.3 and 14.4 in Bühlmann and van de Geer (2011)), we obtain

$$\begin{aligned} \mathbf{E}\left(\frac{1}{n}\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\sum Z_i(\boldsymbol{\beta}) - \mathbf{E}\left(Z_i(\boldsymbol{\beta})\right)\right|\right) &\leq 2\mathbf{E}\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\frac{1}{n}\sum D_i\left(Z_i(\boldsymbol{\beta})\right)\right|\right) \\ &\leq 4C_0\mathbf{E}\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\frac{1}{n}\sum D_i\left(W(\mathbf{X}_i/\gamma)\mathbf{X}_i^T\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)/\gamma\right)\right|\right). \end{aligned} \tag{A.30}$$

Again using (C3) and the moment inequality, it follows that

$$\begin{aligned} RHS\ (A.30) &\leq 4C_2\sqrt{k_n}\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2\mathbf{E}\left|\frac{1}{n}\sum D_i\right| \\ &\leq 4C_2\mu\sqrt{k_n}\left(\mathbf{E}\left(\frac{1}{n}\sum D_i\right)^2\right)^{1/2} \\ &\leq 4C_2\mu\sqrt{k_n/n}, \end{aligned} \tag{A.31}$$

where $C_2 > 0$ is some constant. Then from (A.30) and (A.31), we obtain

$$\mathbf{E}[Z_n(\mu)] \leq C_3\mu\sqrt{k_n/n}, \tag{A.32}$$

where $C_3 = 4C_2$. Now combining (A.29)–(A.32) and applying Massart's concentration theorem (see, e.g., Theorem 14.2 in Bühlmann and van de Geer (2011)), for any $t > 0$ we obtain

$$\mathbf{P}\left(Z_n(\mu) > C_3\mu\sqrt{k_n/n} + t\right) \leq \exp\left(-\frac{nt^2}{8C_1^2\mu^2 k_n}\right).$$

This completes the proof.

Lemma 2.3. *Assume that conditions* (C1)–(C4) *hold. Suppose* $\mathbb{N} = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0},\ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n\}$ *denotes a neighborhood around* $\boldsymbol{\beta}^*$ *for some sequence* $v_n \to 0$. *Assume that* $\min_{j>k_n} d_j$ *is strictly positive,* $\sqrt{(1+v_n k_n^{(2+\delta)/2})}\log_2 n = o(\sqrt{n}\lambda_n)$, $\lambda_n > 2\sqrt{(1+c)(\log p_n)/n}$, *and* $\lambda_n > c_0 v_n\sqrt{k_n}$ *for some constant* $c_0 > 0$. *Then*

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{N}} \|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T\Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right)\|_\infty \geq n\lambda_n\right) \to 0,$$

where $\Psi(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}) = (\phi((Y_1 - \mathbf{X}_1^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_1/\gamma), ..., \phi((Y_n - \mathbf{X}_n^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_n/\gamma))^T$.

Proof of Lemma 2.3.

Consider the following decomposition:

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\frac{1}{n}\mathbf{Q}^T\Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right)\right\|_\infty \leq & \\ \sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\frac{1}{n}\mathbf{E}\left(\mathbf{Q}^T\left(\Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right) - \Psi^{(1)}(\mathbf{e})\right)\right)\right\|_\infty & \\ + \left\|\frac{1}{n}\mathbf{Q}^T\Psi^{(1)}(\mathbf{e})\right\|_\infty + \max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\frac{1}{n}\sum|r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)| = I_1 + I_2 + I_3,\end{aligned}$$

where $\mathbf{e} = (e_1, \ldots, e_n)^T$ and

$$\begin{aligned}r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i) = X_{ij}\left(\phi^{(1)}\left((Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right)W(\mathbf{X}_i/\gamma) & \\ - \mathbf{E}\left(X_{ij}\left(\phi^{(1)}\left((Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma\right) - \phi^{(1)}(e_i/\gamma)\right)W(\mathbf{X}_i/\gamma)\right) & \quad \text{(A.33)}\end{aligned}$$

for fixed j satisfying $k_n + 1 \leq j \leq p_n$. We first study I_1. It is easy to see that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\mathbf{E}\left(\mathbf{Q}^T\left(\Psi^{(1)}\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta}\right) - \Psi^{(1)}(\mathbf{e})\right)\right)\right\|_\infty & \\ = \max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\mathbf{E}\left(\sum_{i=1}^n X_{ij}\left(\phi^{(1)}\left((Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma\right) - \phi^{(1)}(e_i/\gamma)\right)W(\mathbf{X}_i/\gamma)\right)\right|. & \quad \text{(A.34)}\end{aligned}$$

Since the e_i's are independent of the $\mathbf{X}_i$'s, from (A.34) we obtain

$$\sup_{\boldsymbol{\beta}\in\mathbb{N}} \left\|\mathbf{E}\left(\mathbf{Q}^T\left(\Psi^{(1)}\left(\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta}\right)-\Psi^{(1)}(\mathbf{e})\right)\right)\right\|_\infty$$
$$= \max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\mathbf{E}\left(\sum_{i=1}^n X_{ij}W(\mathbf{X}_i/\gamma)\phi^{(1)}\left((Y_i-\mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma\right)\right)\right|. \quad \text{(A.35)}$$

By condition (C2) together with the independence of the e_i's and the $\mathbf{X}_i$'s, we obtain

$$\mathbf{E}\left(\sum_{i=1}^n X_{ij}W(\mathbf{X}_i/\gamma)\phi^{(1)}\left((Y_i-\mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma\right)\right)$$
$$= \sum_{i=1}^n \mathbf{E}\left(X_{ij}W(\mathbf{X}_i/\gamma)\left(g(\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)+o(|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)|)\right)\right). \quad \text{(A.36)}$$

Now by combining (A.34)–(A.36), from the assumptions in Lemma 2.3, we obtain that

$$I_1 \leq \lambda_n(O(1)+o(1)). \quad \text{(A.37)}$$

We now study I_2. It is easy show that $\mathbf{E}\left(X_{ij}\phi^{(1)}\left(e_i/\gamma\right)W(\mathbf{X}_i/\gamma)\right)=0$ for all i and $k_n+1\leq j\leq p_n$. By condition (C3), $|(X_{ij}/\gamma)W(\mathbf{X}_i/\gamma)|$ is bounded for all (i,j). Therefore, there exists a constant $M^*>0$ such that $\left|X_{ij}\phi^{(1)}\left(e_i/\gamma\right)W(\mathbf{X}_i/\gamma)\right|\leq M^*$ for all i and $k_n+1\leq j\leq p_n$. By applying Hoeffding's Inequality, if $\lambda_n>2\sqrt{(1+c)(\log p_n)/n}$ with some positive constant c, then we have

$$\mathbf{P}\left(\|\mathbf{Q}^T\Psi^{(1)}(\mathbf{e})\|_\infty\geq n\lambda_n\right)\leq\sum_{k_n+1}^{p_n}2\exp\left(\frac{-2n^2\lambda_n^2}{n\gamma^2(M^*)^2}\right)$$
$$= O(p_n^{-c_9}),$$

where $c_9>0$ is some constant. Therefore, we conclude that $\|\frac{1}{n}\mathbf{Q}^T\Psi^{(1)}(\mathbf{e})\|_\infty\leq\lambda_n$ holds with probability tending to one, since $p_n\to\infty$ as $n\to\infty$. Thus, with probability tending to one we have

$$I_2=\left\|\frac{1}{n}\mathbf{Q}^T\Psi^{(1)}(\mathbf{e})\right\|_\infty=o_p(\lambda_n). \quad \text{(A.38)}$$

It now remains to study $I_3=\max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\frac{1}{n}\sum_{i=1}^n\left|r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i)\right|$. We will show that $I_3=o_p(\lambda_n)$. First, for each i, define the functional space $\Gamma_j=\{r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i):\boldsymbol{\beta}\in\mathbb{N}\}$.

Endow Γ_j with the (random) norm

$$\|r_{\boldsymbol{\beta},j}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} r^2_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)}.$$

For $\varepsilon > 0$, let $N(\varepsilon, \Gamma_j, \|\cdot\|_n)$ denote the covering number of space $(\Gamma_j, \|\cdot\|_n)$ for each j. Using condition (C3), it is easy to show that $\|r_{\boldsymbol{\beta},j}\|_n \leq C^*$ for each j, where C^* is some positive constant. Using conditions (C1) and (C3), for any $\boldsymbol{\beta} \in \mathbb{N}$ and $\boldsymbol{\beta}' \in \mathbb{N}$ we have

$$\begin{aligned}|r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i) - r_{\boldsymbol{\beta}',j}(\mathbf{X}_i, Y_i)| &\leq W(\mathbf{X}_i/\gamma)\left|\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}')\right|^\delta + \mathbf{E}\left(W(\mathbf{X}_i/\gamma)\left|\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}')\right|^\delta\right) \\ &\leq c k_n^{\delta/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^\delta,\end{aligned}$$

where $c > 0$ is a constant. Note that $\|\cdot\|_2^\delta$ is a metric for $0 < \delta < 1$. Then by Theorem 2.7.11 of van der Vaart and Wellner (1996), the covering numbers of the spaces Γ_j and $\mathbb{N}$ satisfy

$$\begin{aligned}N\left(2^{2-s}, \Gamma_j, \|\cdot\|_n\right) &\leq N\left(\frac{2^{2-s}}{c_8 k_n^{\delta/2}}, \mathbb{N}, \|\cdot\|_2^\delta\right) \\ &\leq N\left(\frac{2^{2-s}}{c_8 k_n^{\delta/2}}, \mathbb{N}, \|\cdot\|_2\right),\end{aligned}$$

where the last inequality follows from the fact that the neighborhood $\{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^\delta \leq v_n\}$ is a subset of the neighborhood $\{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n\}$. By Lemma 14.27 in Bühlmann and van de Geer (2011), the ball $\mathbb{N}$ can be covered by $\left(1 + \frac{2v_n}{\varepsilon}\right)^{k_n}$ balls with radius $\varepsilon > 0$. Then it follows that

$$N\left(\frac{2^{2-s}}{c_8 k_n^{\delta/2}}, \Gamma_j, \|\cdot\|_n\right) \leq \left(1 + \frac{2c_8 v_n k_n^{\delta/2}}{2^{2-s}}\right)^{k_n},$$

and hence we have

$$\begin{aligned}\log\left(1 + N\left(\frac{2^{2-s}}{c_8 k_n^{\delta/2}}, \Gamma_j, \|\cdot\|_n\right)\right) &\leq \log 6 + k_n \log\left(1 + 2^{s-1} c_8 k_n^{\delta/2} v_n\right) \\ &\leq 2c_7\left(1 + v_n k_n^{(2+\delta)/2}\right) 2^{2s}\end{aligned}$$

for some constant $c_7 > 0$. Thus, we have shown that all the conditions in Corollary 14.4 in Bühlmann and van de Geer (2011) are satisfied for $\frac{1}{n}\sum_{i=1}^n r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)$. Now applying

this corollary we obtain for any $t > 0$,

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\frac{1}{n}\sum_{i=1}^{n} r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)\right| \geq \frac{C^*}{\sqrt{n}}\left(3\sqrt{(1+v_n k_n^{(2+\delta)/2})}\log_2 n + 4 + 4t\right)\right) \leq \exp\left(-\frac{nt^2}{8}\right).$$

Taking $t = C\sqrt{\log p_n/n}$ with $C > 0$ a large enough constant, it then follows from the union bound that

$$\mathbf{P}\left(\max_{p_n\geq j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\frac{1}{n}\sum_{i=1}^{n} r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)\right| \geq \frac{4C^*}{\sqrt{n}}\left(3\sqrt{(1+v_n k_n^{(2+\delta)/2})}\log_2 n\right)\right) \leq (p_n - k_n)\exp\left(-\frac{C\log p_n}{8}\right),$$

which goes to zero as $n \to \infty$. Therefore, if $\sqrt{(1+v_n k_n^{(2+\delta)/2})}\log_2 n = o(\sqrt{n}\lambda_n)$, then we have

$$I_3 = o_p(\lambda_n). \tag{A.39}$$

Now combining (A.37), (A.38) and (A.39) finishes the proof.

Appendix B: Supplementary material for Chapter 3

B.1: Proofs of theorems in Chapter 3

We begin by defining the following: for some constant $\mu > 0$, we define the set

$$\mathbb{B}(\mu) = \{\boldsymbol{\beta} =(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \mu,\ \text{supp}(\boldsymbol{\beta}) \subseteq \text{supp}(\boldsymbol{\beta}^*)\}, \tag{B.1}$$

where $\text{supp}(\boldsymbol{\beta}^*) = \{1, ..., k_n\}$. Also define the function

$$Z_n(\mu) = \sup_{\boldsymbol{\beta} \in \mathbb{B}(\mu)} \frac{1}{n} \left| \sum_{i=1}^{n} \left(Z_i(\boldsymbol{\beta}) - \mathbf{E}(Z_i(\boldsymbol{\beta}))\right) \right|, \tag{B.2}$$

where $Z_i(\boldsymbol{\beta}) = \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) - \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta}^*)/\gamma\right) W(\mathbf{X}_i/\gamma)$.

Proof of Theorem 3.1.

Note that $\mathbb{B}(\mu)$ is defined in (B.1). Let

$$I = \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta}^*)/\gamma\right) W(\mathbf{X}_i/\gamma).$$

We first show that for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T \in \mathbb{B}(\mu)$,

$$\mathbf{E}(I) \geq cn \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2, \tag{B.3}$$

for sufficiently large n and when μ is properly chosen, where $c > 0$ is some constant. Let

$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + \boldsymbol{\kappa}$, where $\|\boldsymbol{\kappa}\|_2 = \mu$. We rewrite I as

$$\begin{aligned} I &= \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T(\boldsymbol{\beta}_1^* + \boldsymbol{\kappa})\right)/\gamma\right) W(\mathbf{X}_i/\gamma) - \sum_{i=1}^{n} \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1^*\right)/\gamma\right) W(\mathbf{X}_i/\gamma) \\ &= \sum_{i=1}^{n} \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right)\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right) dt \\ &\qquad - \frac{1}{\gamma}\sum_{i=1}^{n} W\left(\mathbf{X}_i/\gamma\right)\phi^{(1)}\left(e_i/\gamma\right)\mathbf{X}_{1i}^T\boldsymbol{\kappa}. \quad \text{(B.4)} \end{aligned}$$

Since e_i is independent of $\mathbf{X}_i$, using condition (D1) and Taylor expansion we obtain from (B.4) that

$$\begin{aligned} \mathbf{E}(I) &= \mathbf{E}\left(\sum_{i=1}^{n} \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right)\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right) dt\right) \\ &= \sum_{i=1}^{n} \mathbf{E}\left(\frac{1}{\gamma} W\left(\mathbf{X}_i/\gamma\right) \int_0^{-\mathbf{X}_{1i}^T\boldsymbol{\kappa}} \left(g(\gamma)t + o(|t|\, q(\gamma)\right) dt\right) \\ &= \frac{g(\gamma)}{2\gamma}\boldsymbol{\kappa}^T\mathbf{E}\left(\sum_{i=1}^{n} \mathbf{X}_{1i}\mathbf{X}_{1i}^T W\left(\mathbf{X}_i/\gamma\right)\right)\boldsymbol{\kappa} + o(1)\frac{q(\gamma)}{2\gamma} n\,\|\boldsymbol{\kappa}\|_2^2\,. \quad \text{(B.5)} \end{aligned}$$

Then by condition (D3), we have $\boldsymbol{\kappa}^T\mathbf{E}\left(\sum_{i=1}^{n} \mathbf{X}_{1i}\mathbf{X}_{1i}^T W\left(\mathbf{X}_i/\gamma\right)\right)\boldsymbol{\kappa} \geq cn\,\|\boldsymbol{\kappa}\|_2^2$ for some constant $c > 0$. Further, $\|\boldsymbol{\kappa}\|_2 = \mu$. Therefore, (B.3) now follows from (B.5).

Now consider a new vector $\tilde{\boldsymbol{\beta}}^*$ defined by $\tilde{\boldsymbol{\beta}}^* = \left((\tilde{\boldsymbol{\beta}}_1^*)^T, \mathbf{0}^T\right)^T$, where

$$\tilde{\boldsymbol{\beta}}_1^* = M\hat{\boldsymbol{\beta}}_1 + (1 - M)\boldsymbol{\beta}_1^*,$$

with $M = \mu/(\mu + \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2)$. Then $\tilde{\boldsymbol{\beta}}^* \in \mathbb{B}(\mu)$. Using the convexity of the objective function $Q_n(\boldsymbol{\beta})$ defined by (3.3) and the definition of $\hat{\boldsymbol{\beta}}_1$, we have

$$Q_n(\tilde{\boldsymbol{\beta}}^*) \leq MQ_n(\hat{\boldsymbol{\beta}}_1, \mathbf{0}) + (1 - M)Q_n(\boldsymbol{\beta}_1^*, \mathbf{0}) \leq Q_n(\boldsymbol{\beta}^*).$$

For $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T \in \mathbb{B}(\mu)$, define function $\upsilon_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi\left((Y_i - \mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right) W(\mathbf{X}_i/\gamma)$.

By the triangle inequality, we have

$$\begin{aligned}\mathbf{E}\left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*)-\upsilon_n(\boldsymbol{\beta}^*)\right) &= (\upsilon_n(\boldsymbol{\beta}^*)-\mathbf{E}\upsilon_n(\boldsymbol{\beta}^*))-\left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*)-\mathbf{E}\upsilon_n(\tilde{\boldsymbol{\beta}}^*)\right)\\ &\quad + Q_n(\tilde{\boldsymbol{\beta}}^*)-Q_n(\boldsymbol{\beta}^*)+n\lambda_n\|\mathbf{d}_0\circ\boldsymbol{\beta}_1^*\|_1-n\lambda_n\|\mathbf{d}_0\circ\tilde{\boldsymbol{\beta}}_1^*\|_1 \qquad \text{(B.6)}\\ &\le nZ_n(\mu)+n\lambda_n\|\mathbf{d}_0\circ(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1^*)\|_1,\end{aligned}$$

where $Z_n(\mu)$ is defined by (B.2). Using the Cauchy-Schwarz inequality again, it follows that $n\lambda_n\|\mathbf{d}_0\circ(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1^*)\|_1\le n\lambda_n\|\mathbf{d}_0\|_2\mu$. Hence from (B.6),

$$\mathbf{E}\left(\upsilon_n(\tilde{\boldsymbol{\beta}}^*)-\upsilon_n(\boldsymbol{\beta}^*)\right)\le nZ_n(\mu)+n\lambda_n\|\mathbf{d}_0\|_2\mu. \qquad \text{(B.7)}$$

Define the event $A_n=\{Z_n(\mu)\le 2\mu\sqrt{k_n(\log n)/n}\}$. By Lemmas 3.2 and 3.3, $\mathbf{P}(A_n)\to 1$ as $n\to\infty$. Then from (B.3) and (B.7), on the event A_n we have

$$cn\|\tilde{\boldsymbol{\beta}}_1^*-\boldsymbol{\beta}_1^*\|_2^2\le 2\mu\sqrt{k_n n\log n}+n\lambda_n\|\mathbf{d}_0\|_2\mu. \qquad \text{(B.8)}$$

Now take $\mu=O(\sqrt{k_n/n}+\lambda_n\|\mathbf{d}_0\|_2)$. Then by assumptions in Theorem 3.1, $\mu\to 0$. Hence from (B.8), we obtain

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_1^*-\boldsymbol{\beta}_1^*\|_2^2 &\le O\left(\left(\sqrt{k_n(\log n)/n}+\lambda_n\|\mathbf{d}_0\|_2\right)\left(\sqrt{k_n/n}+\lambda_n\|\mathbf{d}_0\|_2\right)\right),\\ &\le O\left(\sqrt{k_n(\log n)/n}+\lambda_n\|\mathbf{d}_0\|_2\right)^2. \qquad \text{(B.9)}\end{aligned}$$

Note that $\|\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1^*\|_2\le O(\mu)$ implies $\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2\le O(\mu)$. So, from (B.9) it follows that on the event A_n,

$$\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2\le O\left(\sqrt{k_n(\log n)/n}+\lambda_n\|\mathbf{d}_0\|_2\right).$$

A bound for the l_1-loss follows from the inequality $\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_1\le\sqrt{k_n}\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^*\|_2$. This completes the proof of Theorem 3.1.

Proof of Theorem 3.2.

Theorem 3.1 shows that $\hat{\boldsymbol{\beta}}_1$ is a minimizer of $Q_n(\boldsymbol{\beta}_1,\mathbf{0})$. We now that $\hat{\boldsymbol{\beta}}=\left(\hat{\boldsymbol{\beta}}_1^T,\mathbf{0}^T\right)^T$ is a global minimizer of $Q_n(\boldsymbol{\beta})$ in the space $\mathbb{R}^{p_n}$. For this purpose, it is enough to show

that the following condition holds with probability tending to 1:

$$\|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T \psi^{(1)} \left(\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta} \right) / \gamma \right) \|_\infty < Cn\lambda_n, \tag{B.10}$$

where

$$\psi((\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta})/\gamma) = (\phi((Y_1 - \mathbf{X}_1^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_1/\gamma), ..., \phi((Y_n - \mathbf{X}_n^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_n/\gamma))^T,$$

$\mathbf{d}_1^{-1} = \left(d_{k_n+1}^{-1}, \ldots, d_{p_n}^{-1} \right)^T$, $C > 0$ is some constant and $\boldsymbol{\beta} \in \mathbb{N}$ with neighborhood $\mathbb{N}$ defined in Lemma 3.4. Then the result follows from Karush-Kuhn-Tuker (KKT) conditions in optimization theory.

Define events

$$N_1 = \{\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \alpha_n\},$$

and

$$N_2 = \{\sup_{\boldsymbol{\beta} \in \mathbb{N}} \|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T \psi^{(1)} \left(\left(\mathbf{Y} - \mathfrak{X}\boldsymbol{\beta} \right) / \gamma \right) \|_\infty < C_1 n\lambda_n\},$$

where α_n is defined in Theorem 3.1. Then from Theorem 3.1 and Lemma 3.4, it follows that $\mathbf{P}\left(N_1 \cap N_2\right) \to 1$. Therefore, (B.10) holds on the event $N_1 \cap N_2$, since $\hat{\boldsymbol{\beta}} \in \mathbb{N}$ on the event N_1. Again, a bound for the l_1-loss follows from the inequality $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_1 \leq \sqrt{k_n}\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2$. This completes the proof.

Proof of Theorem 3.3.

Let $\boldsymbol{\theta} = \mathbf{V}_n^{-1}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)$ with $\mathbf{V}_n = \left(\mathbf{E}\left(\frac{g(\gamma)}{2\gamma} \sum_{i=1}^n \mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma) \right) \right)^{-1/2}$. Next, denote $G_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\beta}_1, \mathbf{0}) - Q_n(\boldsymbol{\beta}_1^*, \mathbf{0})$, where $Q_n(\boldsymbol{\beta})$ is defined by (3.3). Then it follows that

$$\begin{aligned} G_n(\boldsymbol{\theta}) = & \sum_{i=1}^n \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1 \right) / \gamma \right) W(\mathbf{X}_i/\gamma) \\ & - \sum_{i=1}^n \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1^* \right) / \gamma \right) W\left(\mathbf{X}_i/\gamma \right) \\ & + n\lambda_n \left(\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}) \|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 \right). \end{aligned} \tag{B.11}$$

Note that Theorem 3.1 shows that $Q_n(\boldsymbol{\beta}, \mathbf{0})$ is minimized at $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_1$ and hence $G_n(\boldsymbol{\theta})$

is minimized at $\hat{\boldsymbol{\theta}} = \mathbf{V}_n^{-1}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right)$. Now we study terms in $G_n(\boldsymbol{\theta})$ in details. First write $G_n(\boldsymbol{\theta})$ as follows:

$$G_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}) + T_n(\boldsymbol{\theta}), \tag{B.12}$$

where $L_n(\boldsymbol{\theta}) = \mathbf{E}[G_n(\boldsymbol{\theta})]$ and

$$T_n\left(\boldsymbol{\theta}\right) = S_n\left(\boldsymbol{\beta}_1\right) - S_n\left(\boldsymbol{\beta}_1^*\right) - \mathbf{E}\left(S_n\left(\boldsymbol{\beta}_1\right) - S_n\left(\boldsymbol{\beta}_1^*\right)\right),$$

with $S_n(\boldsymbol{\beta}_1) = \sum_{i=1}^n \phi\left(\left(Y_i - \mathbf{X}_{1i}^T\boldsymbol{\beta}_1\right)/\gamma\right) W\left(\mathbf{X}_i/\gamma\right)$. A technique similar to (B.5) gives

$$\begin{aligned}
&\mathbf{E}\left(S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*)\right) \\
&\qquad = \sum_{i=1}^n \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)^T \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\mathbf{X}_{1i}\mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma)\right)\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right) \\
&\qquad\qquad + O(1)\left(\sum_{i=1}^n \mathbf{E}\left(\frac{g(\gamma)}{2\gamma}W(\mathbf{X}_i/\gamma)\right)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2+\varepsilon}\right),
\end{aligned} \tag{B.13}$$

for some constant $0 < \varepsilon < 1$. Now define a set

$$\mathbb{A}_n = \{\boldsymbol{\theta} \in \mathbb{R}^{k_n} : \|\boldsymbol{\theta}\|_2 < c_0\sqrt{k_n}\},$$

where $c_0 > 0$ is some constant independent of k_n. The first term on the RHS of (B.13) is equal to $\|\boldsymbol{\theta}\|_2^2$. By condition (D3), we have $\|\mathbf{V}_n\boldsymbol{\theta}\|_2 \le c_1 n^{-1/2}\|\boldsymbol{\theta}\|_2 \le c_2 n^{-1/2}k_n^{1/2}$ for any $\boldsymbol{\theta} \in \mathbb{A}_n$, where some constants $c_1 > 0$ and $c_2 > 0$. Then by the Cauchy-Schwarz inequality, for any $\boldsymbol{\theta} \in \mathbb{A}_n$ we obtain

$$\begin{aligned}
\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right| &\le \|\mathbf{X}_{1i}\|_2\|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\|_2 \\
&= \|\mathbf{X}_{1i}\|_2\|\mathbf{V}_n\boldsymbol{\theta}\|_2 \\
&\le \|\mathbf{X}_{1i}\|_2 C n^{-1/2}k_n^{1/2}.
\end{aligned}$$

By condition (D2), we have $W(\mathbf{X}_i/\gamma)\|\mathbf{X}_{1i}\|_2^{2+\varepsilon} \le c_3 k_n^{(2+\varepsilon)/2}$ for some constant $c_3 > 0$.

Then for any $\boldsymbol{\theta} \in \mathbb{A}_n$ and $k_n = o(n^{\varepsilon/(4+2\varepsilon)})$, it follows that

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{E}\left(W(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2+\varepsilon}\right) &= O(nk_n^{(2+\varepsilon)/2}n^{-(2+\varepsilon)/2}k_n^{(2+\varepsilon)/2}) \\ &= O(n^{-\varepsilon/2}k_n^{2+\varepsilon}) \\ &= o(1). \end{aligned} \tag{B.14}$$

From (B.13) and (B.14) we now have

$$\mathbf{E}\left(S_n(\boldsymbol{\beta}_1) - S_n(\boldsymbol{\beta}_1^*)\right) = \|\boldsymbol{\theta}\|_2^2 + o(1). \tag{B.15}$$

The matrix $\mathbf{V}_n$ has bounded eigenvalues from condition (D3). Hence, for any $\boldsymbol{\theta} \in \mathbb{A}_n$ we obtain

$$\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta})\|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 = \tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta}, \tag{B.16}$$

where $\tilde{\mathbf{d}}_0$ is a k_n-dimensional vector with i-th component as $d_i \operatorname{sgn}(\beta_j^*)$ and $\operatorname{sgn}(\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}) = \operatorname{sgn}(\boldsymbol{\beta}_1^*)$. Then from (B.15) and (B.16), one obtains

$$L_n(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2 + n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} + o(1), \tag{B.17}$$

uniformly over all $\boldsymbol{\theta} \in \mathbb{A}_n$.

Next we deal with the stochastic component $T_n(\boldsymbol{\theta})$ in (B.12). Let us define $\mathbf{M} = -\left(\phi^{(1)}(e_1/\gamma), \ldots, \phi^{(1)}(e_n/\gamma)\right)^T$ and $\mathbf{U}_n = (\mathbf{Z}_n^*)^T\mathbf{M}$. Then $\mathbf{E}\left(\mathbf{U}_n^T\boldsymbol{\theta}\right) = 0$. Define

$$\Omega_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\phi\left(\left(e_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)\right)/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) - \sum_{i=1}^{n}\phi\left(e_i/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) - \mathbf{U}_n^T\boldsymbol{\theta}.$$

We write $T_n(\boldsymbol{\theta})$ as follows:

$$T_n(\boldsymbol{\theta}) = \mathbf{U}_n^T\boldsymbol{\theta} + h_n(\boldsymbol{\theta}), \tag{B.18}$$

where $h_n(\boldsymbol{\theta}) = \Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)$. From Lemma 3.1, for any $\varepsilon > 0$ it follows that

$$\mathbf{P}\left(|h_n(\boldsymbol{\theta})| > \varepsilon\right) \leq \exp\left(-c_4\varepsilon a_n k_n\right),$$

for some constant $c_4 > 0$ and a sequence $a_n \to \infty$. Define $J_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}) - n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} -$

$\mathbf{U}_n^T\boldsymbol{\theta}$. Note that $G_n(\boldsymbol{\theta})$, $J_n(\boldsymbol{\theta})$ and $\|\boldsymbol{\theta}\|_2^2$ are convex functions in $\boldsymbol{\theta}$. Furthermore, $h_n(\boldsymbol{\theta})$ can be written as

$$h_n(\boldsymbol{\theta}) = J_n(\boldsymbol{\theta}) - \|\boldsymbol{\theta}\|_2^2 - o(1).$$

In addition, for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathbb{A}_n$ we have

$$\left|\|\boldsymbol{\theta}_1\|_2^2 - \|\boldsymbol{\theta}_2\|_2^2\right| = \left|(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\right| \leq O\left(k_n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty\right).$$

Therefore, we have now verified that all the conditions in Lemma 4 in Fan et al. (2014) are satisfied for $h_n(\boldsymbol{\theta})$. Now applying this lemma, for any compact set $A_{k_n} = \{\|\boldsymbol{\theta}\|_2 \leq c_5\sqrt{k_n}\} \subset \mathbb{A}_n$ with $c_5 < c_0$, we obtain

$$\sup_{\boldsymbol{\theta}\in A_{k_n}} |h_n(\boldsymbol{\theta})| = o_p(1). \tag{B.19}$$

Then from (B.12), (B.17) and (B.18), we now have

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \|\boldsymbol{\theta}\|_2^2 + n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} + \mathbf{U}_n^T\boldsymbol{\theta} + h_n(\boldsymbol{\theta}) + o(1) \\ &= \|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2^2 - \|\boldsymbol{\zeta}_n\|_2^2 + o(1), \end{aligned} \tag{B.20}$$

where

$$\boldsymbol{\zeta}_n = -\frac{1}{2}\left(n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0 + \mathbf{U}_n\right). \tag{B.21}$$

Then using a weak convergence argument it follows that

$$\mathbf{c}^T\left((\mathbf{Z}_n^*)^T\mathbf{Z}_n^*\right)^{-1/2}\mathbf{U}_n \xrightarrow{\mathcal{D}} N\left(0, \sigma_\gamma^2\right), \tag{B.22}$$

for any vector $\mathbf{c} \in \mathbb{R}^{k_n}$ such that $\mathbf{c}^T\mathbf{c} = 1$, where $\sigma_\gamma^2 = \mathbf{Var}\left(\phi^{(1)}(e_i/\gamma)\right)$. From (B.21) and (B.22), we have

$$\mathbf{c}^T\left((\mathbf{Z}_n^*)^T\mathbf{Z}_n^*\right)^{-1/2}\left(2\boldsymbol{\zeta}_n + n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0\right) \xrightarrow{\mathcal{D}} N\left(0, \sigma_\gamma^2\right). \tag{B.23}$$

It is now enough to show that for any $\iota > 0$,

$$\mathbf{P}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\zeta}_n\|_2 > \iota\right) \to 0, \tag{B.24}$$

and hence the minimizer $\hat{\boldsymbol{\theta}}$ of $G_n(\boldsymbol{\theta})$ is close to $\boldsymbol{\zeta}_n$. Then the proof of Theorem 3.3 is completed by using Slutsky's Theorem along with (B.23) and (B.24).

Let $B^*(n)$ denote a ball with center $\boldsymbol{\zeta}_n$ and radius $\iota > 0$. Using Lemma 3.5 and the assumption that $\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2 = O(\sqrt{k_n})$, we have

$$\begin{aligned}\|\boldsymbol{\zeta}_n\|_2 &= \left\|\frac{1}{2}\left(n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0 + \mathbf{U}_n\right)\right\|_2 \\ &\leq \frac{1}{2}\left(\|n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0\|_2 + \|\mathbf{U}_n\|_2\right) \\ &= \frac{c_6\sqrt{k_n}}{2}\left(1 + O_p(1)\right),\end{aligned}$$

for some constant $c_6 > 0$. Now choose the constant c_0 in $\mathbb{A}_n$ large enough so that $c_0 > c_6/2$. Then the constant c_5 in A_{k_n} can be chosen large enough to contain $B^*(n)$. Then, by (B.19) we have

$$\Delta_n \doteq \sup_{\boldsymbol{\theta}\in B^*(n)} |h_n(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta}\in A_{k_n}} |h_n(\boldsymbol{\theta})| = o_p(1). \tag{B.25}$$

Let us now consider the behavior of $G_n(\boldsymbol{\theta})$ outside of $B^*(n)$. Let $\boldsymbol{\theta} = \boldsymbol{\zeta}_n + b\mathbf{u} \in \mathbb{R}^{k_n}$, where $\mathbf{u} \in \mathbb{R}^{k_n}$ is a unit vector and $b > \iota > 0$ are some constants. Let $\boldsymbol{\theta}^*$ be the boundary point of $B^*(n)$ that lies on the line segment from $\boldsymbol{\zeta}_n$ to $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}^*$ can be written as $\boldsymbol{\theta}^* = \boldsymbol{\zeta}_n + \iota\mathbf{u} = \left(1 - \frac{\iota}{b}\right)\boldsymbol{\zeta}_n + \frac{\iota}{b}\boldsymbol{\theta}$. Using the convexity of $G_n(\boldsymbol{\theta})$, definition of Δ_n, (B.20) and (B.25), we then obtain

$$\frac{\iota}{b}G_n(\boldsymbol{\theta}) + \left(1 - \frac{\iota}{b}\right)G_n\left(\boldsymbol{\zeta}_n\right) \geq G_n\left(\boldsymbol{\theta}^*\right) \geq \iota^2 - \|\boldsymbol{\zeta}_n\|_2^2 - \Delta_n \geq \iota^2 + G_n\left(\boldsymbol{\zeta}_n\right) - 2\Delta_n.$$

Since $b > \iota$, for large n we have

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\zeta}_n\|_2>\iota} G_n(\boldsymbol{\theta}) \geq G_n\left(\boldsymbol{\zeta}_n\right) + \frac{b}{\iota}\left(\iota^2 - o_p(1)\right) > G_n\left(\boldsymbol{\zeta}_n\right). \tag{B.26}$$

Following from (B.26), we show that the minimum of $G_n(\boldsymbol{\theta})$ cannot occur at any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2 > \iota$. Hence, with probability tending to 1, we have $\|\boldsymbol{\theta} - \boldsymbol{\zeta}_n\|_2 \leq \iota$. This completes the proof.

B.2: Proofs of lemmas in Chapter 3

Lemma 3.1. *Let* $\Omega_n(\boldsymbol{\theta}) = \sum_{i=1}^n \Omega_{ni}(\boldsymbol{\theta})$, *where* $\Omega_{ni}(\boldsymbol{\theta}) = \phi\left((e_i - \mathbf{Z}_{ni}^T\boldsymbol{\theta})/\gamma\right) W\left(\mathbf{X}_i/\gamma\right) - \phi(e_i/\gamma)W(\mathbf{X}_i/\gamma) + \phi^{(1)}(e_i/\gamma)(\mathbf{Z}_{ni}^*)^T\boldsymbol{\theta}$. *Suppose that the conditions of Theorem* 3.3 *are satisfied, then for any* $\varepsilon > 0$,

$$\mathbf{P}\left(|\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)| > \varepsilon\right) \leq \exp\left(-Ca_nk_n\varepsilon\right),$$

where $a_n > 0$ *is some sequence tending to infinity such that* $a_n k_n^{2\delta+3} n^{-\delta} \to 0$, *and* $C > 0$ *is some constant.*

Proof of Lemma 3.1.

Define $\xi_i = \Omega_{ni}(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right)$. Then we have $\sum_{i=1}^n \xi_i = \Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)$. By Markov's inequality, we have that for any $\varepsilon > 0$ and $t > 0$,

$$\begin{aligned}\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right) \geq \varepsilon\right) &\leq \exp(-t\varepsilon)\mathbf{E}\left(\exp\left(t\sum_{i=1}^n \xi_i\right)\right) \\ &= \exp\left(-t\varepsilon - t\sum_{i=1}^n \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right)\right)\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right).\end{aligned} \tag{B.27}$$

Then we show that

$$t\sum_{i=1}^n \mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right) = t\|\boldsymbol{\theta}\|_2^2 + o(1).$$

Then, from (B.27) we have

$$\begin{aligned}\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right) \geq \varepsilon\right) &\leq \exp(-t\varepsilon)\mathbf{E}\left(\exp\left(t\sum_{i=1}^n \xi_i\right)\right) \\ &\propto \exp\left(-t\varepsilon - t\|\boldsymbol{\theta}\|_2^2\right)\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right).\end{aligned} \tag{B.28}$$

From (B.28), applying Taylor expansion gives

$$\begin{aligned}\mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right) &= \mathbf{E}\left(1 + t\Omega_{ni}(\boldsymbol{\theta}) + O\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right) \\ &= 1 + t\mathbf{E}\left(\Omega_{ni}(\boldsymbol{\theta})\right) + O\left(\mathbf{E}\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right) \\ &= 1 + \frac{t\|\boldsymbol{\theta}\|_2^2}{n} + O\left(\mathbf{E}\left(t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right).\end{aligned}$$

For $x_i > 0$, we have $\prod_{i=1}^n (1 + x_i) \leq \exp\left(\sum_{i=1}^n x_i\right)$. Using this type of bound, we obtain

$$\begin{aligned}\prod_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta})\right)\right) &\leq \exp\left(\sum_{i=1}^n \mathbf{E}\left(\exp\left(t\Omega_{ni}(\boldsymbol{\theta}) - 1\right)\right)\right) \\ &= \exp\left(t\|\boldsymbol{\theta}\|_2^2 + O\left(\mathbf{E}\left(\sum_{i=1}^n t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right)\right).\end{aligned}$$

Therefore, (B.28) is followed by

$$\mathbf{P}\left(\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)\right) > \varepsilon\right) \leq \exp\left(-t\varepsilon + O\left(\mathbf{E}\left(\sum_{i=1}^n t^2\Omega_{ni}^2(\boldsymbol{\theta})\right)\right)\right).$$

Now let $t = a_n k_n$ with a_n being some diverging sequence such that $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$. Then, it follows that

$$\mathbf{P}\left(\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right)\right) > \varepsilon\right) \leq \exp\left(-C\varepsilon k_n a_n\right).$$

It now remains to show that $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$. Using the definition of $\Omega_{ni}(\boldsymbol{\theta})$,

$$\begin{aligned}&\Omega_{ni}^2(\boldsymbol{\theta}) = \\ &\left(\phi\left(\left(Y_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)\right)/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) - \phi\left(e_i/\gamma\right)W\left(\mathbf{X}_i/\gamma\right) + \phi^{(1)}(e_i/\gamma)\left(\mathbf{Z}_{ni}^*\right)^T\boldsymbol{\theta}\right)^2,\end{aligned}$$

and then by the mean value theorem, we obtain

$$\mathbf{E}\left(\Omega_{ni}^2(\boldsymbol{\theta})\right) = \frac{1}{\gamma^2}\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)W(\mathbf{X}_i/\gamma) - \phi^{(1)}(e_i^*/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)W(\mathbf{X}_i/\gamma)\right)^2, \tag{B.29}$$

where $e_i^* \in \left(e_i, e_i - \mathbf{Z}_{ni}^T\boldsymbol{\theta} = Y_i - \mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\right)\right)$. Then by condition (D1), we have

$$\begin{aligned}\mathbf{E}\left(\phi^{(1)}(e_i/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) - \phi^{(1)}(e_i^*/\gamma)\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right)^2 \\ \leq C_4\mathbf{E}\left(\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{\delta}\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right)^2 \\ \leq C_4\mathbf{E}\left(\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right),\end{aligned}$$

where $C_4 > 0$ is some finite constant. Therefore, we now have from (B.29) for $t > 0$,

$$t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) \leq tC_4\sum_{i=1}^n \mathbf{E}\left(W^2(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right).$$

In addition, we obtain $\sum_{i=1}^n \mathbf{E}\left(W^2(\mathbf{X}_i/\gamma)\left|\mathbf{X}_{1i}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right|^{2(\delta+1)}\right) = O(n^{-\delta}k_n^{2(\delta+1)})$. Then we have $t\mathbf{E}\left(\sum_{i=1}^n \Omega_{ni}^2(\boldsymbol{\theta})\right) = o(1)$ since $t = a_n k_n$ and by the assumptions on k_n in Theorem 3.3. Using a similar argument for $\mathbf{P}\left(\Omega_n(\boldsymbol{\theta}) - \mathbf{E}\left(\Omega_n(\boldsymbol{\theta})\right) \leq -\varepsilon\right)$ completes the proof.

Lemma 3.2. *Let $Z_n(\mu)$ be defined by* (B.2). *Assume that conditions* (D1) *to* (D2) *hold. Then for any $t > 0$ we have*

$$\mathbf{P}\left(Z_n(\mu) > \frac{2R_n}{\sqrt{n}}(3\sqrt{(1 + C_0\mu^2 k_n)}\log_2 n + 4) + t\right) \leq \exp\left(-\frac{nt^2}{8R_n^2}\right) + \mathbf{P}\left(T_n(\mu) > \frac{t}{2}\right),$$

where $R_n = Ck_n^{(\delta+1)/2}\mu^{\delta+1}$, $T_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\frac{1}{n}\left|\sum_{i=1}^n \frac{1}{\gamma}W\left(\mathbf{X}_i/\gamma\right)\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\phi^{(1)}\left(e_i/\gamma\right)\right|$, and $C_0 > 0,\ C > 0$ are some finite constants.

Proof of Lemma 3.2.

By an application of Taylor expansion,

$$\begin{aligned}Z_i(\boldsymbol{\beta}) &= \int_0^{-\mathbf{X}_i^T\boldsymbol{\theta}} \frac{1}{\gamma}W\left(\mathbf{X}_i/\gamma\right)\left(\phi^{(1)}\left(\left(e_i + t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right)dt - \frac{1}{\gamma}W\left(\mathbf{X}_i/\gamma\right)\mathbf{X}_i^T\boldsymbol{\theta}\phi^{(1)}\left(e_i/\gamma\right) \\ &= \tilde{Z}_i(\boldsymbol{\theta}) - \frac{1}{\gamma}W\left(\mathbf{X}_i/\gamma\right)\mathbf{X}_i^T\boldsymbol{\theta}\phi^{(1)}\left(e_i/\gamma\right).\end{aligned}$$

Note that $\mathbf{E}(\tilde{Z}_i(\boldsymbol{\theta})) = \mathbf{E}(Z_i(\boldsymbol{\beta}))$ for all i from condition (D1). Using the δ-Lipschitz prop-

erty of $\phi^{(1)}$ and condition (D2), for any $\boldsymbol{\beta} \in \mathbb{B}(\mu)$ we have

$$\begin{aligned}
\left|\tilde{Z}_i(\boldsymbol{\theta})\right| &\leq \frac{W\left(\mathbf{X}_i/\gamma\right)}{\gamma} \int_0^{-\mathbf{X}_i^T\boldsymbol{\theta}} \left|\left(\phi^{(1)}\left(\left(e_i+t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right)\right| dt \\
&\leq \frac{W\left(\mathbf{X}_i/\gamma\right)}{\gamma} \int_0^{-\mathbf{X}_i^T\boldsymbol{\theta}} \frac{1}{\gamma^{\delta}} \left|t\right|^{\delta} dt \\
&\leq \frac{1}{\gamma^{\delta+1}} W\left(\mathbf{X}_i/\gamma\right) \left|\mathbf{X}_i^T\boldsymbol{\theta}\right|^{\delta+1} \\
&\leq c k_n^{(\delta+1)/2} \mu^{\delta+1},
\end{aligned}$$

where $c > 0$ is some constant. Then for any $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \mathbf{0}^T\right)^T \in \mathbb{B}(\mu)$, it follows that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\tilde{Z}_i(\boldsymbol{\theta})\right)^2} \leq R_n, \tag{B.30}$$

where $R_n = C k_n^{(\delta+1)/2} \mu^{\delta+1}$.

Now define a functional space $\Gamma^* = \{\tilde{Z}_i(\boldsymbol{\theta}) : \boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}^* \text{ with } \boldsymbol{\beta} \in \mathbb{B}(\mu)\}$. Endow Γ^* with the (random) norm

$$\left\|\tilde{Z}(\boldsymbol{\theta})\right\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\tilde{Z}_i(\boldsymbol{\theta})\right)^2}.$$

For $\varepsilon > 0$, let $N\left(\varepsilon, \Gamma^*, \|\cdot\|_n\right)$ denote the covering number of space $\left(\Gamma^*, \|\cdot\|_n\right)$, that is the minimum number of balls with radius ε necessary to cover the class Γ^*. Then from (B.30), $\left\|\tilde{Z}(\boldsymbol{\theta})\right\|_n \leq R_n$. Using conditions (D1) and (D2), for any $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\boldsymbol{\theta}' = \boldsymbol{\beta}' - \boldsymbol{\beta}^*$ with $\boldsymbol{\beta} \in \mathbb{B}(\mu)$ and $\boldsymbol{\beta}' \in \mathbb{B}(\mu)$ we have

$$\begin{aligned}
\left|\tilde{Z}_i(\boldsymbol{\theta}) - \tilde{Z}_i(\boldsymbol{\theta}')\right| &\leq \left| \frac{W\left(\mathbf{X}_i/\gamma\right)}{\gamma} \int_{-\mathbf{X}_i^T\boldsymbol{\theta}'}^{-\mathbf{X}_i^T\boldsymbol{\theta}} \left|\left(\phi^{(1)}\left(\left(e_i+t\right)/\gamma\right) - \phi^{(1)}\left(e_i/\gamma\right)\right)\right| dt \right| \\
&\leq \frac{W\left(\mathbf{X}_i/\gamma\right)}{\gamma} \int_{-\mathbf{X}_i^T\boldsymbol{\theta}'}^{-\mathbf{X}_i^T\boldsymbol{\theta}} \frac{1}{\gamma^{\delta}} \left|t\right|^{\delta} dt \\
&\leq c k_n^{(\delta+1)/2} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^{\delta+1} \\
&\leq C \mu^{\delta} k_n^{(\delta+1)/2} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2
\end{aligned}$$

where $C > 0$ is some constant. Then by Theorem 2.7.11 of van der Vaart and Wellner

(1996), the covering numbers of the spaces Γ^* and $\mathbb{B}(\mu)$ satisfy

$$N\left(2^{-s}R_n, \Gamma^*, \|\cdot\|_n\right) \leq N\left(\frac{2^{-s}R_n}{c^{'}\mu^{\delta}k_n^{(\delta+1)/2}}, \mathbb{B}(\mu), \|\cdot\|_2\right).$$
$$= N\left(\frac{2^{-s}C}{c^{'}\mu}, \mathbb{B}(\mu), \|\cdot\|_2\right)$$

for any $s \geq 1$. By Lemma 14.27 in Bühlmann and van de Geer (2011), the ball $\mathbb{B}(\mu)$ can be covered by $\left(1+\frac{2\mu}{\varepsilon}\right)^{k_n}$ balls with radius $\varepsilon > 0$. Then it follows that

$$N\left(\frac{2^{-s}C}{c^{'}\mu}, \Gamma^*, \|\cdot\|_n\right) \leq \left(1+\frac{2c^{'}\mu^2}{2^{-s}C}\right)^{k_n},$$

and hence we have

$$\log\left(1+N\left(2^{-s}R_n, \Gamma^*, \|\cdot\|_n\right)\right) \leq \log 2 + k_n \log\left(1+2^{s+1}c^{'}C^{-1}\mu^2\right)$$
$$\leq \log 2 + k_n 2^{s+1}c^{'}C^{-1}\mu^2 \leq (1+C_0\mu^2 k_n)2^{2s},$$

where $C_0 = c^{'}C^{-1}$. Thus, we have shown that all the conditions in Corollary 14.4 in Bühlmann and van de Geer (2011) are satisfied for $\frac{1}{n}\sum_{i=1}^n \tilde{Z}_i(\boldsymbol{\theta})$. Now applying this corollary we obtain for any $t > 0$,

$$\mathbf{E}\left(\frac{1}{n}\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\sum_{i=1}^n \xi_i\left(\tilde{Z}_i(\boldsymbol{\theta})\right)\right|\right) \leq \frac{R_n}{\sqrt{n}}\left(3\sqrt{(1+C_0\mu^2k_n)}\log_2 n + 4\right),$$

where $\xi_1, ..., \xi_n$ is a Rademacher sequence, independent of $\tilde{Z}_1(\boldsymbol{\theta}), ..., \tilde{Z}_n(\boldsymbol{\theta})$. Then using the preceding inequality and the symmetrization theorem (see, e.g., Theorem 14.3 in Bühlmann and van de Geer (2011)), we obtain

$$\mathbf{E}\left(\frac{1}{n}\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\sum_{i=1}^n \tilde{Z}_i(\boldsymbol{\theta}) - \mathbf{E}\left(\tilde{Z}_i(\boldsymbol{\theta})\right)\right|\right) \leq 2\mathbf{E}\left(\frac{1}{n}\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\sum_{i=1}^n D_i\left(\tilde{Z}_i(\boldsymbol{\theta})\right)\right|\right)$$
$$\leq \frac{2R_n}{\sqrt{n}}\left(3\sqrt{(1+C_0\mu^2k_n)}\log_2 n + 4\right). \quad \text{(B.31)}$$

Denote $\tilde{Z}_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\frac{1}{n}\left|\sum_{i=1}^n\left(\tilde{Z}_i(\boldsymbol{\theta}) - \mathbf{E}(\tilde{Z}_i(\boldsymbol{\theta}))\right)\right|$. Then from (B.31) we have

$$\mathbf{E}\left(\tilde{Z}_n(\mu)\right) \leq \frac{2R_n}{\sqrt{n}}\left(3\sqrt{(1+C_0\mu^2k_n)}\log_2 n + 4\right). \quad \text{(A.32)}$$

Now combining (B.31), (B.32) and applying Massart's concentration theorem (see, e.g.,

Theorem 14.2 in Bühlmann and van de Geer (2011)), for any $t > 0$ we obtain

$$\mathbf{P}\left(\tilde{Z}_n(\mu) > \frac{2R_n}{\sqrt{n}}(3\sqrt{(1+C_0\mu k_n)}\log_2 n + 4) + t\right) \leq \exp\left(-\frac{nt^2}{8R_n^2}\right). \tag{B.33}$$

Write $a_n = \frac{2R_n}{\sqrt{n}}(3\sqrt{(1+C_0\mu^2 k_n)}\log_2 n + 4)$. Then, from (B.33) for any $t > 0$ we have

$$\begin{aligned}
\mathbf{P}\left(Z_n(\mu) > a_n + t\right) &\leq \mathbf{P}\left(\tilde{Z}_n(\mu) + T_n(\mu) > a_n + t\right) \\
&\leq \mathbf{P}\left(\tilde{Z}_n(\mu) > a_n + \frac{t}{2}\right) + \mathbf{P}\left(T_n(\mu) > \frac{t}{2}\right) \\
&\leq \exp\left(-\frac{nt^2}{32R_n^2}\right) + \mathbf{P}\left(T_n(\mu) > \frac{t}{2}\right).
\end{aligned}$$

This completes the proof.

Lemma 3.3. *Assume that conditions* (D1)–(D4) *hold. Let* $T_n(\mu)$ *be as defined in Lemma* 3.2. *Then* $\lim_{n\to\infty} \mathbf{P}(T_n(\mu) > \frac{t}{2}) = 0$ *for any* $t > 0$.

Proof of Lemma 3.3.

Let

$$\tilde{T}_n(\mu) = \sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_{\boldsymbol{\theta}}(\mathbf{X}_i) \right|$$

with $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, $f_{\boldsymbol{\theta}}(\mathbf{X}_i) = \frac{1}{\mu c^* \sqrt{k_n}} W\left(\mathbf{X}_i/\gamma\right) \mathbf{X}_i^T \boldsymbol{\theta}$, and $\varepsilon_i = \phi^{(1)}\left(e_i/\gamma\right)$ for $i = 1, ..., n$, where $c^* > 0$ is a constant such that $|X_{ij} W\left(\mathbf{X}_i/\gamma\right)| \leq c^*$ for all X_{ij}. From condition (D2), $|f_{\boldsymbol{\theta}}(\mathbf{X}_i)| \leq 1$ for all $\boldsymbol{\beta} \in \mathbb{B}(\mu)$ and all $\mathbf{X}_i$. From condition (D4), $\varepsilon_1, ..., \varepsilon_n$ are sub-Gaussian, and hence, $\varepsilon_1 f_{\boldsymbol{\theta}}(\mathbf{X}_1), ..., \varepsilon_n f_{\boldsymbol{\theta}}(\mathbf{X}_n)$ are also sub-Gaussian for all $\boldsymbol{\beta} \in \mathbb{B}(\mu)$.

Define a class of functions $\mathcal{F} = \{f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}^* \text{ with } \boldsymbol{\beta} \in \mathbb{B}(\mu)\}$. Endow $\mathcal{F}$ with the norm $\|f_{\boldsymbol{\theta}}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{\boldsymbol{\theta}}(\mathbf{x}_i))^2}$. Then $\|f_{\boldsymbol{\theta}}\|_n \leq 1$ for all $f_{\boldsymbol{\theta}} \in \mathcal{F}$. For $0 < \delta < 1$, let $N(\delta, \mathcal{F}, \|\cdot\|_n)$ denote the covering number of space $(\mathcal{F}, \|\cdot\|_n)$. Using condition (D2) and the Cauchy-Schwarz inequality, for any $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\boldsymbol{\theta}' = \boldsymbol{\beta}' - \boldsymbol{\beta}^*$ with $\boldsymbol{\beta} \in \mathbb{B}(\mu)$ and

$\boldsymbol{\beta}' \in \mathbb{B}(\mu)$ we have

$$\begin{aligned}
|f_{\boldsymbol{\theta}}(\mathbf{X}_i) - f_{\boldsymbol{\theta}'}(\mathbf{X}_i)| &= \frac{1}{\mu c^* \sqrt{k_n}} \left| W\left(\mathbf{X}_i/\gamma\right) \mathbf{X}_i^T \boldsymbol{\theta} - W\left(\mathbf{X}_i/\gamma\right) \mathbf{X}_i^T \boldsymbol{\theta}' \right| \\
&\leq \frac{1}{\mu c^* \sqrt{k_n}} W\left(\mathbf{X}_i/\gamma\right) \left| \mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}') \right| \\
&\leq \frac{1}{\mu c^* \sqrt{k_n}} W\left(\mathbf{X}_i/\gamma\right) \|\mathbf{X}_i\|_2 \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2 \\
&\leq \frac{1}{\mu} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2.
\end{aligned}$$

Then by Theorem 2.7.11 of van der Vaart and Wellner (1996), the covering numbers of the spaces $\mathcal{F}$ and $\mathbb{B}(\mu)$ satisfy

$$N\left(\delta, \mathcal{F}, \|\cdot\|_n\right) \leq N\left(2^{-1}\mu\delta, \mathbb{B}(\mu), \|\cdot\|_2\right).$$

By Lemma 14.27 in Bühlmann and van de Geer (2011), the ball $\mathbb{B}(\mu)$ can be covered by $\left(1 + \frac{2\mu}{\varepsilon}\right)^{k_n}$ balls with radius $\varepsilon > 0$. Then from the preceding display it follows that

$$N\left(\delta, \mathcal{F}, \|\cdot\|_n\right) \leq \left(1 + \frac{4}{\delta}\right)^{k_n},$$

and hence, we have

$$\begin{aligned}
\log\left(1 + 2N\left(\delta, \mathcal{F}, \|\cdot\|_n\right)\right) &\leq \log 4 + k_n \log\left(1 + \frac{4}{\delta}\right) \\
&\leq \log 4 + k_n \left(\frac{4}{\delta}\right) \\
&\leq \frac{8k_n}{\delta} = \left(\frac{A_\nu}{\delta}\right)^{2(\nu)},
\end{aligned}$$

where $A_\nu = 8k_n$ and $\nu = 1/2$. Thus, conditions of Corollary 14.6 in Bühlmann and van de Geer (2011) are satisfied for $\mathcal{F}$ and $N\left(\delta, \mathcal{F}, \|\cdot\|_n\right)$ with $A_\nu = 8k_n$ and $\nu = 1/2$. Then applying this corollary, for any $t > 0$ we have

$$\begin{aligned}
\mathbf{P}\left(\exists\, f_{\boldsymbol{\theta}} \in \mathcal{F}\colon \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_{\boldsymbol{\theta}}(\mathbf{X}_i) \right| \geq K_0 A_\nu^\nu \|f_{\boldsymbol{\theta}}\|_n^{1-\nu} \left(2^{1-\nu} - 1\right)^{-1} + K_0 \|f_{\boldsymbol{\theta}}\|_n^{1-\nu} t \right) \\
\leq \exp\left[-\frac{t^2}{K_0^2}\right] (1 + 2B_\nu), \quad \text{(B.34)}
\end{aligned}$$

where K_0 and B_ν are positive constants. Since $A_\nu = 8k_n$ and $\nu = 1/2$, and $\|f_{\boldsymbol{\theta}}\|_n \leq 1$ for all $f_{\boldsymbol{\theta}} \in \mathcal{F}$, thus from (B.34) we obtain

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}(\mu)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_{\boldsymbol{\theta}}(\mathbf{X}_i)\right| \geq \frac{1}{\sqrt{n}}K_0 2\sqrt{2}\sqrt{k_n}(2^{1/2}-1)^{-1} + \frac{1}{\sqrt{n}}K_0 t\right) \leq \exp\left[-\frac{t^2}{K_0^2}\right](1+2B_{1/2}), \quad (5.1)$$

and hence

$$\mathbf{P}\left(\tilde{T}_n(\mu) \geq \frac{1}{\sqrt{n}}K_0 2\sqrt{2}\sqrt{k_n}(2^{1/2}-1)^{-1} + \frac{1}{\sqrt{n}}K_0 t\right) \leq \exp\left[-\frac{t^2}{K_0^2}\right](1+2B_{1/2}).$$

Note that $\tilde{T}_n(\mu) = \frac{1}{\mu c^*\sqrt{k_n}}T_n(\mu)$, where $T_n(\mu)$ is defined in Lemma 3.2. Then we have for any $t > 0$

$$\mathbf{P}\left(T_n(\mu) \geq \frac{\mu c^*\sqrt{k_n}}{\sqrt{n}}K_0 2\sqrt{2}\sqrt{k_n}(2^{1/2}-1)^{-1} + \frac{\mu c^*\sqrt{k_n}}{\sqrt{n}}K_0 t\right) \leq \exp\left[-\frac{t^2}{K_0^2}\right](1+2B_{1/2}),$$

that is, for any $t > 0$

$$\mathbf{P}\left(T_n(\mu) \geq \frac{k_n}{\sqrt{n}}K_0 2\sqrt{2}(2^{1/2}-1)^{-1} + \frac{t}{2}\right) \leq \exp\left[-\frac{nt^2}{4k_n(\mu c^*)^2 {K_0}^4}\right](1+2B_{1/2}). \quad (B.35)$$

From (B.35), it follows that $\lim_{n\to\infty}\mathbf{P}(T_n(\mu) > \frac{t}{2}) = 0$ for any $t > 0$ when $k_n = o(\sqrt{n})$. This completes the proof.

Lemma 3.4. *Assume that conditions* (D1) *to* (D4) *hold. Suppose that* $\mathbb{N} = \{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T,\ \boldsymbol{\beta}_2^T\right)^T \in \mathbb{R}^{p_n} : \boldsymbol{\beta}_2 = \mathbf{0},\ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n\}$ *denotes a neighborhood around* $\boldsymbol{\beta}^*$ *for some sequence* $v_n \to 0$. *If* $\min_{j>k_n} d_j > c$ *for some constant* $c > 0$, $k_n^{1/2}v_n^\delta\sqrt{(1+v_n k_n^{3/2})}\log_2 n = o(\sqrt{n}\lambda_n)$, $\lambda_n > 2\sqrt{(1+c)(\log p_n)/n}$, *and* $\lambda_n > C^*\sqrt{k_n}v_n$ *for some constant* $C^* > 0$ *are satisfied, then*

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{N}}\|\mathbf{d}_1^{-1}\circ\mathbf{Q}^T\psi^{(1)}\left((\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta})/\gamma\right)\|_\infty \geq n\lambda_n\right) \to 0,$$

where $\psi((\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta})/\gamma) = (\phi((Y_1 - \mathbf{X}_1^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_1/\gamma), ..., \phi((Y_n - \mathbf{X}_n^T\boldsymbol{\beta})/\gamma)W(\mathbf{X}_n/\gamma))^T$.

Proof of Lemma 3.4.

Consider the following decomposition:

$$\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\frac{1}{n}\mathbf{Q}^T\psi^{(1)}\left((\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta})/\gamma\right)\right\|_\infty \leq$$
$$\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\frac{1}{n}\mathbf{E}\left(\mathbf{Q}^T\left(\psi^{(1)}((\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta})/\gamma)-\psi^{(1)}(\mathbf{e}/\gamma)\right)\right)\right\|_\infty$$
$$+\left\|\frac{1}{n}\mathbf{Q}^T\psi^{(1)}(\mathbf{e}/\gamma)\right\|_\infty+\max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\frac{1}{n}\sum_{i=1}^n|r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i)|\doteq I_1+I_2+I_3,$$

where $\mathbf{e}=(e_1,\ldots,e_n)^T$ and

$$r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i)=X_{ij}\left(\phi^{(1)}\left((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)-\phi^{(1)}\left(e_i/\gamma\right)\right)W(\mathbf{X}_i/\gamma)$$
$$-\mathbf{E}\left(X_{ij}\left(\phi^{(1)}\left((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)-\phi^{(1)}(e_i/\gamma)\right)W(\mathbf{X}_i/\gamma)\right) \quad \text{(B.36)}$$

for fixed j satisfying $k_n+1\leq j\leq p_n$.

We first study I_1. Since the e_i's are independent of the $\mathbf{X}_i$'s, from (B.36) we obtain

$$\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left\|\mathbf{E}\left(\mathbf{Q}^T\left(\psi^{(1)}\left((\mathbf{Y}-\mathfrak{X}\boldsymbol{\beta})/\gamma\right)-\psi^{(1)}(\mathbf{e}/\gamma)\right)\right)\right\|_\infty$$
$$=\max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\mathbf{E}\left(\sum_{i=1}^n X_{ij}W(\mathbf{X}_i/\gamma)\left(\phi^{(1)}\left((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)\right)\right)\right|. \quad \text{(B.37)}$$

For $\boldsymbol{\beta}\in\mathbb{N}$, $\phi((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma)=\phi((Y_i-\mathbf{X}_{1i}^T\boldsymbol{\beta}_1)/\gamma)$. Then by condition (D1), we have

$$\mathbf{E}\left(\sum_{i=1}^n X_{ij}W(\mathbf{X}_i/\gamma)\left(\phi^{(1)}\left((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)\right)\right)$$
$$=\sum_{i=1}^n\mathbf{E}\left(X_{ij}W(\mathbf{X}_i/\gamma)\left(g(\gamma)\mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\right)+o(|\mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\right)|)\right)\right). \quad \text{(B.38)}$$

By the Cauchy-Schwarz inequality, $\left|\mathbf{X}_{1i}^T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\right)\right|\leq\|\mathbf{X}_{1i}\|_2\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2\leq v_n\|\mathbf{X}_{1i}\|_2$ for all $\boldsymbol{\beta}\in\mathbb{N}$. By condition (D2), $|X_{ij}W(\mathbf{X}_i/\gamma)|\|\mathbf{X}_{1i}\|_2\leq ck_n^{1/2}$ for all (i,j), where $c>0$ is some constant. Then by combining (B.37), (B.38) and using $\lambda_n>C^*\sqrt{k_n}v_n$, we obtain

$$I_1\leq\lambda_n(O(1)+o(1)).$$

Next we proceed to I_2. By definition, $I_2=\|\frac{1}{n}\sum_{i=1}^n X_{ij}\psi^{(1)}(e_i/\gamma)\|_\infty$. By condition (D1), $\mathbf{E}\left(X_{ij}\phi^{(1)}(e_i/\gamma)W(\mathbf{X}_i/\gamma)\right)=0$ for all (i,j). By condition (D2), $X_{ij}W(\mathbf{X}_i/\gamma))$ is

bounded for all (i,j), i.e., $|X_{ij}W(\mathbf{X}_i/\gamma))| \leq c_0^*$ for some constant $c_0^* > 0$ independent of (i,j). Then by condition (D4), it follows that $(K_1)^2(\mathbf{E}\exp[|X_{ij}\phi^{(1)}(e_i/\gamma)W(\mathbf{X}_i/\gamma)|^2/K_1^2] - 1) \leq \sigma_1^2$ for all (i,j), where $K_1 = c_0^* K_0$ and $\sigma_1 = c_0^*\sigma_0$. Define $R^2 = K_1^2 + \sigma_1^2$. Then by Lemma 14.16 in Bühlmann and van de Geer (2011), we have

$$\mathbf{P}\left(\left\|\sum_{i=1}^{n}\frac{1}{n}X_{ij}\psi^{(1)}(e_i/\gamma)\right\|_\infty \geq R\sqrt{8(t^2+\frac{\log(2p_n)}{n})}\right) \leq \exp(-nt^2).$$

Then by taking $t = O(\sqrt{\lambda_n^2 - \log(2p_n)/n})$, we obtain $I_2 = n^{-1}\|\mathbf{Q}^T\psi^{(1)}(\mathbf{e}/\gamma)\|_\infty = o(\lambda_n)$ with asymptotic probability one.

Next we consider I_3. We will show that $I_3 = \max_{j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\frac{1}{n}\sum_{i=1}^{n}|r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i)| = o(\lambda_n)$ with probability tending to zero. First, for each i, define the functional space $\Gamma_j = \{r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i) : \boldsymbol{\beta}\in\mathbb{N}\}$. Endow Γ_j with the (random) norm

$$\|r_{\boldsymbol{\beta},j}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}r_{\boldsymbol{\beta},j}^2(\mathbf{X}_i,Y_i)}. \tag{B.39}$$

For $\varepsilon > 0$, let $N(\varepsilon,\Gamma_j,\|\cdot\|_n)$ denote the covering number of space $(\Gamma_j,\|\cdot\|_n)$ for each j. Using the local δ-Lipschitz condition on $\phi^{(1)}$ and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left|\phi^{(1)}\left((Y_i-\mathbf{X}_i^T\boldsymbol{\beta})/\gamma\right)-\phi^{(1)}\left(e_i/\gamma\right)\right| &\leq \left|\mathbf{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right|^\delta \\ &\leq \|\mathbf{X}_i\|_2^\delta\,\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^\delta\,. \end{aligned}\tag{B.40}$$

Using condition (D2), for $0 < \delta < 1$ we have

$$|X_{ij}W(\mathbf{X}_i/\gamma)|\,\|\mathbf{X}_i\|_2^\delta \leq c_1^* k_n^{1/2} \tag{B.41}$$

for some constant $c_1^* > 0$. Then using (B.39), (B.40) and (B.41) leads to

$$\|r_{\boldsymbol{\beta},j}\|_n \leq c_2^* k_n^{1/2} v_n^\delta \tag{B.42}$$

for some constant $c_2^* > 0$ and $k_n + 1 \leq j \leq p_n$. Again using the Cauchy-Schwarz inequality,

conditions (D1) and (D2), for any $\boldsymbol{\beta} \in \mathbb{N}$ and $\boldsymbol{\beta}' \in \mathbb{N}$ we have

$$\begin{aligned}|r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i) - r_{\boldsymbol{\beta}',j}(\mathbf{X}_i, Y_i)| &\leq \left|X_{ij} W(\mathbf{X}_i/\gamma)\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}')\right|^\delta + \mathbf{E}\left(\left|X_{ij} W(\mathbf{X}_i/\gamma)\mathbf{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}')\right|^\delta\right) \\ &\leq c_3^* k_n^{1/2} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^\delta\end{aligned}$$

for some constant $c_3^* > 0$. Note that $\|\cdot\|_2^\delta$ is a metric for $0 < \delta < 1$. Then by Theorem 2.7.11 of van der Vaart and Wellner (1996), the covering numbers of the spaces Γ_j and $\mathbb{N}$ satisfy

$$\begin{aligned}N\left(2^{2-s}, \Gamma_j, \|\cdot\|_n\right) &\leq N\left(\frac{2^{2-s}}{c_7 k_n^{1/2}}, \mathbb{N}, \|\cdot\|_2^\delta\right) \\ &\leq N\left(\frac{2^{2-s}}{c_7 k_n^{1/2}}, \mathbb{N}, \|\cdot\|_2\right),\end{aligned}$$

where the last inequality follows from the fact that the neighborhood $\{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^\delta \leq v_n\}$ is a subset of the neighborhood $\{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq v_n\}$. By Lemma 14.27 in Bühlmann and Van de Geer (2011), the ball $\mathbb{N}$ can be covered by $\left(1 + \frac{2v_n}{\varepsilon}\right)^{k_n}$ balls with radius $\varepsilon > 0$. Therefore, we have

$$N\left(\frac{2^{2-s}}{c_7 k_n^{1/2}}, \Gamma_j, \|\cdot\|_n\right) \leq \left(1 + \frac{2c_7 v_n k_n^{1/2}}{2^{2-s}}\right)^{k_n},$$

and hence we obtain

$$\begin{aligned}\log\left(1 + N\left(\frac{2^{2-s}}{c_7 k_n^{1/2}}, \Gamma_j, \|\cdot\|_n\right)\right) &\leq \log 6 + k_n \log\left(1 + 2^{s-1} c_7 k_n^{1/2} v_n\right) \\ &\leq 2c_8\left(1 + v_n k_n^{3/2}\right) 2^{2s}\end{aligned}$$

for some constant $c_8 > 0$. Thus, we have shown that all the conditions in Corollary 14.4 in Bühlmann and van de Geer (2011) are satisfied for $\frac{1}{n}\sum_{i=1}^n r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)$. Now applying this corollary and using (B.42) we obtain for any $t > 0$,

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\frac{1}{n}\sum_{i=1}^n r_{\boldsymbol{\beta},j}(\mathbf{X}_i, Y_i)\right| \geq \frac{c_2^* k_n^{1/2} v_n^\delta}{\sqrt{n}}\left(3\sqrt{(1 + v_n k_n^{3/2})}\log_2 n + 4 + 4t\right)\right) \leq \exp\left(-\frac{nt^2}{8}\right).$$

Taking $t = C\sqrt{\log p_n / n}$ with $C > 0$ large enough, it then follows from the union bound

that

$$\mathbf{P}\left(\max_{p_n\geq j>k_n}\sup_{\boldsymbol{\beta}\in\mathbb{N}}\left|\frac{1}{n}\sum_{i=1}^{n}r_{\boldsymbol{\beta},j}(\mathbf{X}_i,Y_i)\right|\geq\frac{4c_2^*k_n^{1/2}v_n^{\delta}}{\sqrt{n}}\left(3\sqrt{(1+v_nk_n^{3/2})}\log_2 n\right)\right)$$
$$\leq(p_n-k_n)\exp\left(-\frac{C\log p_n}{8}\right),$$

which goes to zero as $n\to\infty$. Therefore, if $k_n^{1/2}v_n^{\delta}\sqrt{(1+v_nk_n^{3/2})}\log_2 n=o(\sqrt{n}\lambda_n)$, then we have $I_3=o_p(\lambda_n)$. This completes the proof of Lemma 3.4.

Lemma 3.5. *Assume that conditions* (D2) *and* (D3) *hold. In addition, define*

$$\mathbf{M}=-\left(\phi^{(1)}(e_1/\gamma),\ldots,\phi^{(1)}(e_n/\gamma)\right)^T$$

and

$$\mathbf{U}_n=(\mathbf{Z}_n^*)^T\mathbf{M},$$

where

$$\mathbf{Z}_n^*=\left(\frac{1}{\gamma}\mathbf{X}_{11}W\left(\mathbf{X}_1/\gamma\right),\ldots,\frac{1}{\gamma}\mathbf{X}_{1n}W\left(\mathbf{X}_n/\gamma\right)\right)^T\mathbf{V}_n$$

with

$$\mathbf{V}_n=\left(\mathbf{E}\left(\frac{g(\gamma)}{2\gamma}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^TW(\mathbf{X}_i/\gamma)\right)\right)^{-1/2}.$$

Then $\|\mathbf{U}_n\|_2=O_P(\sqrt{k_n})$.

Proof of Lemma 3.5.

Observe that

$$\begin{aligned}\mathbf{U}_n^T\mathbf{U}_n&=\mathbf{M}^T\mathbf{Z}_n^*\left(\mathbf{Z}_n^*\right)^T\mathbf{M}\\&=\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^TW(\mathbf{X}_i/\gamma)\phi^{(1)}(e_1/\gamma)\right)^T\left(n\mathbf{V}_n\mathbf{V}_n^T\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^TW(\mathbf{X}_i/\gamma)\phi^{(1)}(e_1/\gamma)\right)\\&\leq C_3\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_{1i}\mathbf{X}_{1i}^TW(\mathbf{X}_i/\gamma)\phi^{(1)}(e_i/\gamma)\right\|_2,\end{aligned}\tag{B.43}$$

the last equality follows from the fact that $n\mathbf{V}_n\mathbf{V}_n^T$ has bounded eigenvalues, see condition

(D3), where $C_3 > 0$ is some constant. By repeated application of the triangle inequality and using condition (D2) we also have

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_{1i} \mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma) \phi^{(1)}(e_i/\gamma) \right\|_2 \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\| \mathbf{X}_{1i} \mathbf{X}_{1i}^T W(\mathbf{X}_i/\gamma) \right\|_2 \left| \phi^{(1)}(e_i/\gamma) \right|$$
$$\leq C_2 \frac{\sqrt{k_n}}{\sqrt{n}} \sum_{i=1}^{n} \left| \phi^{(1)}(e_i/\gamma) \right|. \tag{B.44}$$

From an application of the central limit theorem, it follows that $\Lambda_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left| \phi^{(1)}(e_i/\gamma) \right| \xrightarrow{\mathcal{D}} N\left(\mu_\gamma, \sigma_\gamma^2\right)$, where $\mu_\gamma = \mathbf{E}(|\phi^{(1)}(e_i/\gamma)|)$ and $\sigma_\gamma^2 = \mathbf{Var}(|\phi^{(1)}(e_i/\gamma)|)$. Then by Prohorov's Theorem, $\{\Lambda_n\}$ is uniformly tight (or bounded in probability). Then from (B.43) and (B.44), $\frac{\mathbf{U}_n^T \mathbf{U}_n}{\sqrt{k_n}}$ is uniformly tight. That is, $\mathbf{U}_n^T \mathbf{U}_n = O_P(\sqrt{k_n})$. This completes the proof.

Appendix C: Supplementary material for Chapter 4

C.1: Proofs of theorems in Chapter 4

Proof of Theorem 4.1.

We will derive an error bound on $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ assuming that

$$\lambda_n \geq 2\|\nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s}\right)\|_\infty. \tag{C.1}$$

holds for any $\mathbf{s}$. The results in Theorems 4.2 and 4.3 show that (C.1) holds when e_i follows either a sub-Gaussian or a sub-exponential distribution.

Now let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{s}}^T\right)^T$ denote the minimizer of (4.1). Then we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{c}{2}\left(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}} - \hat{s}_i/c\right)^2 + \psi_\alpha\left(\hat{s}_i\right)\right) W(\mathbf{X}_i) + \lambda_n\|\hat{\boldsymbol{\beta}}\|_1$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{c}{2}\left(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}^* - \hat{s}_i/c\right)^2 + \psi_\alpha\left(\hat{s}_i\right)\right) W(\mathbf{X}_i) + \lambda_n\|\boldsymbol{\beta}^*\|_1.$$

Then

$$L_n^*\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}\right) - L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right) \leq \lambda_n\|\boldsymbol{\beta}^*\|_1 - \lambda_n\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda_n\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1, \tag{C.2}$$

where

$$L_n^*\left(\boldsymbol{\beta}, \boldsymbol{s}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{c}{2}\left(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - s_i/c\right)^2 + \psi_\alpha\left(s_i\right)\right) W(\mathbf{X}_i).$$

Applying a Taylor expansion on L_n^* with respect to $\boldsymbol{\beta}$, we have

$$L_n^*\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}\right) - L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right) - \left\langle \nabla_{\boldsymbol{\beta}} L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right), \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right\rangle = \frac{c}{2n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^T \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T W(\mathbf{X}_i)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \geq 0. \quad \text{(C.3)}$$

From (C.2) and (C.3), we have

$$\lambda_n \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1 + \|\nabla_{\boldsymbol{\beta}} L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right)\|_\infty \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \geq \lambda_n \|\boldsymbol{\beta}^*\|_1 - \lambda_n \|\hat{\boldsymbol{\beta}}\|_1 + \|\nabla_{\boldsymbol{\beta}} L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right)\|_\infty \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \geq 0. \quad \text{(C.4)}$$

Using assumption (C.1), from (C.4) we obtain

$$\lambda_n \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1 + \|\nabla_{\boldsymbol{\beta}} L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right)\|_\infty \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_1 \leq \frac{3}{2}\lambda_n \sqrt{k_n} \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2.$$

From Lemma 4.2, it follows that

$$\frac{\zeta}{2}\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2 \leq L_n^*\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}\right) - L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right) - \left\langle \nabla_{\boldsymbol{\beta}} L_n^*\left(\boldsymbol{\beta}^*, \hat{\boldsymbol{s}}\right), \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right\rangle \leq \frac{3}{2}\lambda_n \sqrt{k_n} \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2.$$

Hence we obtain the upper bound as claimed:

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \leq 3\zeta^{-1}\lambda_n \sqrt{k_n}.$$

Next we show that the error $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathbb{C}_1$ defined in Lemma 4.2. Using (C.1) again, from (C.4) we obtain

$$\lambda_n \|\boldsymbol{\beta}^*\|_1 - \lambda_n \|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{2}\lambda_n \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \geq 0.$$

Note that we have the following relationship:

$$\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{\beta}_T^*\|_1 - \|\hat{\boldsymbol{\beta}}_T\|_1 - \|\hat{\boldsymbol{\beta}}_{T^c}\|_1 \leq \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_T\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{T^c}\|_1. \quad \text{(C.5)}$$

Then combining (C.2) and (C.5) leads to

$$0 \leq \lambda_n \|\hat{\boldsymbol{\Delta}}_T\|_1 - \lambda_n \|\hat{\boldsymbol{\Delta}}_{T^c}\|_1 + \frac{1}{2}\lambda_n \|\hat{\boldsymbol{\Delta}}_T\|_1 + \frac{1}{2}\lambda_n \|\hat{\boldsymbol{\Delta}}_{T^c}\|_1 = \frac{3}{2}\lambda_n \|\hat{\boldsymbol{\Delta}}_T\|_1 - \frac{1}{2}\lambda_n \|\hat{\boldsymbol{\Delta}}_{T^c}\|_1,$$

which is the cone condition. This completes the proof.

Proof of Theorem 4.2.

It is easy to show that

$$\left|\nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s}\right)\right| = \left| \frac{c}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}^* - \frac{s_i}{c}\right) \mathbf{X}_i W(\mathbf{X}_i) \right|.$$

Then

$$\begin{aligned} \left\|\nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s}\right)\right\|_\infty &= \left\| \frac{c}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}^* - \frac{s_i}{c}\right) \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \\ &\leq \left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty + \left\| \frac{c}{n} \sum_{i=1}^{n} \left(\frac{s_i}{c}\right) \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \doteq I_1 + I_2, \end{aligned}$$

where $e_i = Y_i - \mathbf{X}_i^T \boldsymbol{\beta}^*$, $i \geq 1$.

We begin with the easy term $I_2 = \left\| \frac{1}{n} \sum_{i=1}^{n} s_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty$. For each $j = 1, \ldots, p_n$, by the assumption on $W(\cdot)$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} s_i X_{ij} W(\mathbf{X}_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |s_i|.$$

Then using condition 1, it follows that

$$I_2 \leq c_1 \sqrt{\frac{\log(p_n)}{n}}.$$

Now we consider I_1. Note that $\mathbf{X}_i W(\mathbf{X}_i)$ is bounded by the assumption on $W(\cdot)$. Therefeore, $e_i \mathbf{X}_i W(\mathbf{X}_i)$'s are mean zero sub-Gaussian random variables, as the e_i's are mean zero sub-Gaussian random variables. Then applying Hoeffding's inequality for sub-Gaussian random variables together with the union bound, we obtain for any $t > 0$

$$\mathbf{P}\left(\left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \geq ct \right) \leq 2p_n \exp(-nt^2/2\sigma^2),$$

where σ^2 is the variance proxy of $e_i\mathbf{X}_i W(\mathbf{X}_i)$. Then taking $t = \sqrt{2\sigma^2\varepsilon(\log p_n)/n}$ with $\varepsilon > 1$ yields that

$$\left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \leq c\sqrt{\frac{2\sigma^2 \varepsilon \log(p_n)}{n}}$$

with probability at least $1 - 2p_n \exp(-\varepsilon \log p_n)$. Then combining above results gives

$$\left\| \nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s} \right) \right\|_\infty \leq c\sqrt{\frac{2\sigma^2 \varepsilon \log p_n}{n}} + c_1 \sqrt{\frac{\log p_n}{n}}$$

with probability at least $1 - 1 - 2p_n \exp(-\varepsilon \log p_n)$. Note that the above bound holds for any $\boldsymbol{s}$ satisfying condition 1. Then for the choice

$$\lambda_n = 2\left(c\sqrt{\frac{2\sigma^2 \varepsilon \log p_n}{n}} + c_1 \sqrt{\frac{\log p_n}{n}} \right),$$

(C.1) holds with probability at least $1 - 2p_n \exp(-\varepsilon \log p_n)$ for any $\boldsymbol{s}$ satisfying condition 1.

Proof of Theorem 4.3.

Again recall the partition present in the proof of Theorem 4.2:

$$\begin{aligned} \left\| \nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s} \right) \right\|_\infty &= \left\| \frac{c}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}^* - \frac{s_i}{c} \right) \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \\ &\leq \left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty + \left\| \frac{c}{n} \sum_{i=1}^{n} \frac{s_i}{c} \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \doteq I_3 + I_4. \end{aligned}$$

Using the same technique used in the proof of Theorem 4.2 again gives $I_4 \leq c_1 \sqrt{\frac{\log(p_n)}{n}}$ holds. Now the remaining work is to derive an upper bound of $I_3 = \left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty$.

Based on the assumption on $W(\cdot)$, for each $j = 1, \ldots, p_n$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \left| e_i X_{ij} W(\mathbf{X}_i) \right|^m \leq \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} |e_i|^m.$$

Then applying Lemma 14.13 in Bühlmann and van de Geer (2011) gives for any $t > 0$,

$$\mathbf{P}\left(\left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \geq cKt + c\sqrt{2t} + c\sqrt{\frac{2\log 2p_n}{n}} + \frac{cK \log 2p_n}{n} \right) \leq \exp(-nt).$$

Note that when n is large and K is not too large, the square root terms are the lead-

ing terms. By letting $t = \log(2p_n)/n$, we obtain that with probability at least $1 - \exp(-log(2p_n))$

$$\left\| \frac{c}{n} \sum_{i=1}^{n} e_i \mathbf{X}_i W(\mathbf{X}_i) \right\|_\infty \leq c_2 \sqrt{(\log 2p_n)/n}$$

provided that $n \geq \log(2p_n)$, where $c_2 > 0$ is some constant. Then for the choice

$$\lambda_n = 2 \left(c_2 \sqrt{\frac{\log 2p_n}{n}} + c_1 \sqrt{\frac{\log p_n}{n}} \right),$$

we have

$$\lambda_n \geq 2 \| \nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s} \right) \|_\infty$$

holds with probability at least $1 - \exp(-log(2p_n))$ for any $\boldsymbol{s}$ satisfying condition 1.

C.2: Proofs of lemmas in Chapter 4

Lemma 4.1. (Lemma 12 in Loh and Wainwright 2012) *For a fixed matrix $\boldsymbol{T} \in \mathbb{R}^{p_n \times p_n}$, parameter $s > 0$ and tolerance $\delta > 0$, suppose that we have the deviation condition*

$$\left| \mathbf{v}^T \boldsymbol{T} \mathbf{v} \right| \leq \delta, \ \ \forall \ \mathbf{v} \in \mathbb{R}^{p_n}, \ \ s.t \ \ \| \boldsymbol{v} \|_0 \leq 2s \ \ and \ \ \| \boldsymbol{v} \|_2 \leq 1.$$

Then

$$\left| \mathbf{v}^T \boldsymbol{T} \mathbf{v} \right| \leq 27\delta \left(\| \boldsymbol{v} \|_2^2 + \frac{\| \boldsymbol{v} \|_1^2}{s} \right), \ \ \forall \ \mathbf{v} \in \mathbb{R}^{p_n}.$$

Lemma 4.2. *Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d random variables satisfying $\| \boldsymbol{X}_i \|_2^2 W(\boldsymbol{X}_i) \leq 1$ for some weight function $W(\cdot)$. In addition, let $\bar{\mathbf{H}}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{c}{2} \mathbf{X}_1 \mathbf{X}_1^T W(\mathbf{X}_1)$. Assume $\zeta = \lambda_{\min} \left(\mathbf{E} \left(\bar{\boldsymbol{H}}_{\boldsymbol{\beta}\boldsymbol{\beta}} \right) \right) > 0$, then for any $\boldsymbol{s}$*

$$L_n^*(\boldsymbol{\beta}, \boldsymbol{s}) - L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s} \right) - \left\langle \nabla_{\boldsymbol{\beta}} L_n^* \left(\boldsymbol{\beta}^*, \boldsymbol{s} \right), (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \right\rangle \geq \frac{\zeta}{2} \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|_2^2$$

holds with probability at least $1 - 2\exp(-nc_4)$ *over the set*

$$\mathbb{C}_1 = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\Delta}_{T^c}\|_1 \leq 3\|\boldsymbol{\Delta}_T\|_1\}$$

where $\boldsymbol{\Delta} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, ζ *and* c_4 *are some positive constants.*

Proof of Lemma 4.2.

Recall the Taylor expansion obtained at (C.3):

$$L_n^*(\boldsymbol{\beta}, \boldsymbol{s}) - L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s}) - \langle \nabla_{\boldsymbol{\beta}} L_n^*(\boldsymbol{\beta}^*, \boldsymbol{s}), (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \rangle = \frac{c}{2n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T W(\mathbf{X}_i)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$$

for any fixed $\boldsymbol{s}$. Denote a matrix $\Gamma_1 = \frac{1}{n}\sum_{i=1}^n \frac{c}{2}\mathbf{X}_i\mathbf{X}_i^T W(\mathbf{X}_i) - \mathbf{E}\left(\frac{c}{2}\mathbf{X}_i\mathbf{X}_i^T W(\mathbf{X}_i)\right)$. Then applying Lemma 4.1 together with the assumptions on the weight function $W(\cdot)$ and the design matrix $\boldsymbol{X}$, we have

$$\frac{c}{2n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T W(\mathbf{X}_i)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \geq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{E}\left(\frac{c}{2}\mathbf{X}_i\mathbf{X}_i^T W(\mathbf{X}_i)\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - 27\eta_1\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1^2}{k_n}\right),$$

where $\eta_1 > 0$ is a constant. Again, over the set $\mathbb{C}_1$ we have

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 = \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{T^c}\|_1 + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_T\|_1 \leq 3\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_T\|_1 + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_T\|_1 = 4\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_T\|_1 \leq 4\sqrt{k_n}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_T\|_2 \leq 4\sqrt{k_n}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

Then by letting $\eta_1 = \frac{1}{918}\zeta$ with $\zeta = \lambda_{\min}\left(\mathbf{E}\left(\bar{\boldsymbol{H}}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right)\right)$, we obtain

$$\frac{c}{2n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T W(\mathbf{X}_i)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \geq \frac{\zeta}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$$

for any $\mathbf{s}$.

Now it remains to check the assumption of Lemma 4.1. Note that for $\|\boldsymbol{\Delta}\|_2 \leq 1$,

$\boldsymbol{\Delta}^T \Gamma_1 \boldsymbol{\Delta}$ is the average of i.i.d bounded random variables where

$$\left| \frac{c}{2} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^* \right)^T \mathbf{X}_i \mathbf{X}_i^T W(\mathbf{X}_i) \left(\boldsymbol{\beta} - \boldsymbol{\beta}^* \right) \right| \leq c^2.$$

For any $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_0 \leq 2k_n$, applying Hoeffding's inequality, together with a discretization argument and union bound over the $\binom{p_n}{2k_n}$ choices, gives

$$\begin{aligned} \mathbf{P}\left(\left| \boldsymbol{\Delta}^T \Gamma_1 \boldsymbol{\Delta} \right| \leq \eta_1, \ \ \forall \boldsymbol{\Delta} \in \mathbb{R}^{p_n} \ \ s.t. \ \ \|\boldsymbol{\Delta}\|_0 \leq 2k_n \ \text{ and } \ \|\boldsymbol{\Delta}\|_2 \leq 1 \right) \\ \geq 1 - 2\exp\left(-c_3 n \eta_1 / c^2 + 2k_n \log(p_n) \right). \end{aligned}$$

Then for $n \gtrsim \frac{c^2 k_n \log(p_n)}{\lambda_{\min}\left(\mathbf{E}\left(\bar{\boldsymbol{H}}_{\boldsymbol{\beta}\boldsymbol{\beta}} \right) \right)}$, we have with probability at least $1 - 2\exp(-nc_4)$, $\left| \boldsymbol{\Delta}^T \Gamma_1 \boldsymbol{\Delta} \right| \leq \eta_1$ holds. This completes the proof.