LUO, BIN, Ph.D. Robust Penalized Regression for Complex High-dimensional Data. (2020)
Directed by Dr. Xiaoli Gao. 169 pp.

Robust high-dimensional data analysis has become an important and challenging task in complex Big Data analysis due to the high-dimensionality and data contamination. One of the most popular procedures is the robust penalized regression. In this dissertation, we address three typical robust ultra-high dimensional regression problems via penalized regression approaches. The first problem is related to the linear model with the existence of outliers, dealing with the outlier detection, variable selection and parameter estimation simultaneously. The second problem is related to robust high-dimensional mean regression with irregular settings such as the data contamination, data asymmetry and heteroscedasticity. The third problem is related to robust bi-level variable selection for the linear regression model with grouping structures in covariates.

In Chapter 1, we introduce the background and challenges by overviews of penalized least squares methods and robust regression techniques. In Chapter 2, we propose a novel approach in a penalized weighted least squares framework to perform simultaneous variable selection and outlier detection. We provide a unified link between the proposed framework and a robust M-estimation in general settings. We also establish the non-asymptotic oracle inequalities for the joint estimation of both the regression coefficients and weight vectors. In Chapter 3, we establish a framework of robust estimators in high-dimensional regression models using Penalized Robust Approximated quadratic M estimation (PRAM). This framework allows general settings such as random errors lack of symmetry and homogeneity, or covariates are not sub-Gaussian. Theoretically, we show that, in the ultra-high dimension setting,

the PRAM estimator has local estimation consistency at the minimax rate enjoyed by the LS-Lasso and owns the local oracle property, under certain mild conditions. In Chapter 4, we extend the study in Chapter 3 to robust high-dimensional data analysis with structured sparsity. In particular, we propose a framework of high-dimensional M-estimators for bi-level variable selection. This framework encourages bi-level sparsity through a computationally efficient two-stage procedure. It produces strong robust parameter estimators if some nonconvex redescending loss functions are applied. In theory, we provide sufficient conditions under which our proposed two-stage penalized M-estimator possesses simultaneous local estimation consistency and the bi-level variable selection consistency, if a certain nonconvex penalty function is used at the group level. The performances of the proposed estimators are demonstrated in both simulation studies and real examples. In Chapter 5, we provide some discussions and future work.

ROBUST PENALIZED REGRESSION FOR COMPLEX HIGH-DIMENSIONAL DATA

by

Bin Luo

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation written by Bin Luo has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Xiaoli Gao

Committee Members _____
Sat Gupta

_____
Quefeng Li

_____
Scott Richter

_____
Haimeng Zhang

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

Foremost, I wish to express my deep gratitude to my advisor Dr. Xiaoli Gao for her strong support and insightful guidance throughout my Ph.D. study and research, for her patience, enthusiasm and continuous encouragement. Without her persistent help, I would not have achieved what I have now.

I would also like to thank Dr.s Sat Gupta, Quefeng Li, Scott Richter and Haimeng Zhang for their service on my committee. I really appreciate the precious learning opportunity and environment provided by the Department of Mathematics and Statistics at UNC Greensboro.

I am very grateful to my family, who are always doing their best to support me throughout my life.

Lastly, I would like to thank my wife, Yang, for being extremely supportive and making countless sacrifices to help me get to this point.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

I.1. Background and Challenges

Due to the rapid development of advanced technologies over the last decades, high-dimensional data arise in many scientific fields, with the trend towards radically larger numbers of variables (p) but relatively small number of observations (n), i.e. $p \gg n$. For example, in biomedical studies, huge numbers of magnetic resonance images (MRI) and functional MRI data are collected for each subject with hundreds of subjects involved. Satellite imagery has been used in natural resource discovery and agriculture, collecting thousands of high resolution images. These kind of examples are plentiful among fields of science, engineering and humanities and new knowledge need to be discovered by using these massive high-throughput data [D$^+$00, FL06].

The high-dimensionality of data has posted some challenges in data analysis. One of them is the intensive computation inherent in these high-dimensional mathematical problems. Systematically searching through a high-dimensional space is usually computational infeasible. At the same time, high-dimensionality has significantly challenged traditional statistical theory. For instance, in term of asymptotic theory, the traditional approximation assumes that $n \to \infty$ while p remain smaller order than n or usually fixed. However, the high-dimensional scenario would imagine that p goes to infinity faster than n [JT09]. Other challenges incurred by high-dimensionality also include how to efficiently estimate model parameters in high-dimensional spaces and how to obtain an interpretable model with a large number of variables.

In recent decades, a great number of statistical methods, algorithms and theories have been developed to perform high-dimensional data analysis (HDDA). Among them, penalized least squares (PLS) methods have become very popular in high-dimensional linear regression analysis since the introduction of the LASSO [Tib96a]. A PLS approach is to minimize the penalized objective function combined with both the ℓ_2 loss and a penalty on the coefficients vector. When the penalty is designed to obtain exactly zeros for some coefficients, and nonzero for others, the PLS can perform a simultaneous coefficient estimation and variable selection process, which is attractive in HDDA.

However, the PLS approach may lose its efficiency in both estimation and variable selection in presence of irregular settings such as data contamination. In fact, high-dimensional data can be complex in general: (a) the data are contaminated in both response and a large number of variables [RL05]; (b)the data are highly skewed and heteroscedastic [ZFB14, FLW17]; (c) the covariates possess complex grouping structures [YL06, HBM12]. Hence more sophisticated methods are needed to deal with the high-dimensional complex data.

I.1.1. Data Contamination

In real applications, the data can be contaminated due to the existence of outliers. An outlier is defined as an observation that is very different from other observations based on certain measure. The presence of outliers can lead to biased estimation of parameters, misspecification of the model and misleading predictions. This phenomenon become even more common and challenging in high-dimensional settings. For example, in gene expression analysis, outliers are often produced due to the complicated data generation process. To perform robust variable selection and parameter estimation in HDDA, extensive work on penalized robust M-estimators

has been investigated, such as Huber-Lasso [H$^+$64, LLZ$^+$11] and LAD-Lasso [GH10, Wan13]. Besides, outliers detection also plays a fundamental role in dealing with data contamination. It has important applications in the field of fraud detection, network robustness analysis and intrusion detection. To detect outliers or influential points in high-dimensional regression model, a few diagnosis measures, such as High-dimension Influence Measure (HIM) [ZLL$^+$13], have been proposed.

I.1.2. Data Asymmetry and Heteroscedasticity

Asymmetry along with heteroscedasticity or contamination often occurs with the growth of data dimensionality. In high-dimensional settings, particularly when random errors follow irregular distributions such as asymmetry and heteroscedasticity, simultaneous mean estimation and variable selection are still of interest in many applications. For example, in economics where asymmetric data is prevalent, it is still of interest to study how mean GDP is affected by many features. Another example can be found in RNA-seq data analysis, the highly skewed nature and mean-variance dependency of RNA-Seq data may pose difficulties on building prognostic gene signatures.

[H$^+$64] implies that the location estimator generated by Huber's method is possibly biased for certain fixed asymmetric contamination. For M-estimation in linear regression model, [Car79,CW88] indicate that data asymmetry does not affect the slope estimation asymptotically when the error and covariates distributions are independent. However, the case of asymmetric and heteroscedastic errors was not well addressed. [FLW17] further points out that most of penalized robust M-estimators generate bias to the conditional mean regression function when the error distribution is asymmetric and heteroscedastic. Thus it remains a challenge to effectively reduce the bias generated by asymmetric distribution under data contamination in high-dimensional settings.

I.1.3. Grouping Structure in Covariates

Covariates often function group-wisely in many applications. For example, in gene expression analysis, genes from the same biological pathways may exhibit similar activities. In high-dimensional data analysis, bi-level sparsity is often assumed when covariates function group-wisely and sparsity can appear either at the group level or within certain groups. Penalized least squares approaches with penalties incorporating grouping structures, such as the group Lasso [YL06], have become very popular in recent decades. To avoid the all-in or all-out variable selection at the group level, extensive methods such as the sparse group Lasso [FHT10, SFHT13] have been investigated to perform bi-level variable selection. However, when the data are contaminated or heavy-tailed in high-dimension settings, it remains a challenge to perform robust bi-level variable selection and parameter estimation.

I.1.4. Real Data Example

We close this section by introducing a real data example. The NTC-60 data is a gene expression data set collected from Affymetrix HG-U133A chip, which is corresponding to a high-dimensional case ($p > n$). The dataset consists of data on 60 human cancer cell lines and can be downloaded via the web application CellMiner (http://discover.nci.nih.gov/cellminer/). The study is to predict the protein expression on the KRT18 antibody from other gene expression levels. The expression levels of the protein *keratin* 18 is known to be persistently expressed in carcinomas [OBC96]. And the response variable is chosen from variables with the largest MAD. After removing the missing data, there are $n = 59$ samples with $21,944$ genes in the dataset. One can refer [SRN$^+$07] for more details.

[LLLP11] applies only non-robust regression methods to this data and obtains models with hundreds of predictors that are thus difficult to interpret. In this thesis,

considering the possible irregularity in the dataset, the robust high-dimensional data analysis approaches are applied.

I.2. Penalized Least Squares Method

To enhance model interpretability and make statistical inference feasible in high-dimensional regression models, the sparsity condition is proposed that among a large set of variables only a few of them are relevant. In such cases, variable selection techniques are crucial for identifying important variables and improving estimation accuracy. For last decades perhaps the most popular approaches for sparse high-dimensional models are the Penalized Least Squares (PLS) methods. The other techniques include sequential approaches (e.g. LARS [EHJ$^+$04], Forward Regression [Wan09], Sequential Lasso [LC14]) and screening methods (e.g. Sure Independence Screening [FL08], Sure Independent Ranking and Screening [ZLLZ11], nonparametric independence screening [FFS11]).

Consider a high-dimensional linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad 1 \leq i \leq n, \tag{I.1}$$

where y_i and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ are the observed response variable and covariates vector, $\epsilon_1, \ldots, \epsilon$ are i.i.d. random variables with mean 0 and variance σ^2. Note that $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)^T \in \mathbb{R}^p$ is an s-sparse coefficient vector (only include s nonzero elements) and $p \gg n$. A class of PLS estimators for $\boldsymbol{\beta}^*$ takes the following form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{I.2}$$

where ρ_λ is a penalty function and λ is a tuning parameter in the penalty. The form of ρ_λ determines the flavor of penalized regression and λ controls the magnitude of the penalty. Specially, when $\lambda = 0$, the penalty term goes away and we are left

with the ordinary least squares estimator. In most scenarios, the penalty function is coordinate-separable such that

$$\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho_\lambda(\beta_j),$$

for some scalar function $\rho_\lambda : \mathbb{R} \mapsto \mathbb{R}$.

The work of AIC [Aka98] and BIC [S$^+$78] suggests to choose a parameter $\boldsymbol{\beta}$ that minimizes the penalized least squares in (I.2) with the ℓ_0-norm penalty

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbf{I}(\beta_j \neq 0),$$

when the random error ϵ is normal. With the ℓ_0-norm penalty, the PLS method can be viewed as a model selection approach that penalizes the number of variables in the model. However, this penalized ℓ_0 regression is unstable with respect to small perturbations in the data, since the ℓ_0 penalty is not continuous. It is also computational infeasible in the high-dimensional space.

[FF93] generalizes the penalized ℓ_0 regression to the bridge regression by considering

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{P} |\beta_j|^\gamma \text{ for } 0 < \gamma \leq 2.$$

It bridges the penalized ℓ_0 regression ($\gamma \to 0$) to the ridge regression [HK70] ($\gamma = 2$). When $\gamma \leq 1$, the component of $\boldsymbol{\beta}$ in (I.2) can be shrunk to zero if λ is sufficiently large, thus achieving simultaneous coefficient estimation and variable selection. While the bridge penalty with $\gamma < 1$ is continuous, its infinite derivative at the origin may cause numerical problem.

The special case when $\gamma = 1$ is related to the least absolute shrinkage and selection operator (Lasso) [Tib96a], which is a very popular shrinkage method for

variable selection. The Lasso penalty (ℓ_1 penalty) can be viewed as a convex surrogate of the ℓ_0 penalty. But it is more stable due to its continuity and computationally feasible for high-dimensional data. From the Bayesian perspective, the Lasso estimator can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double-exponential) priors [PC08].

The statistical properties of the Lasso estimator have been extensively studied (e.g. [KF00], [EHJ+04], [Zou06], [ZY06], [ZH06], [ZH+08],[MY+09] and [BRT09]). [FL01] shows that the Lasso shrinkage produces biased estimates for the large coefficient. [BRT09] presents that the Lasso is asymptotically equivalent to the Dantzig selector [CT07], with the ℓ_2 error rate of prediction or estimation being $s/n \log(p)$, where the number of variable p can be much larger than the sample size n. [ZY06] characterizes the model selection consistency of the Lasso by proposing the property of sign consistency,

$$P\left(sgn(\boldsymbol{\beta}^*) = sgn(\hat{\boldsymbol{\beta}})\right) \to 1 \text{ as } n \to \infty,$$

where $sgn(\boldsymbol{\beta})$ is a vector of signs of β_js and $sgn(0)$ is defined as 0. They show that the Lasso is sign consistent if the following irrepresentable condition is satisfied,

$$\|\mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} sgn(\boldsymbol{\beta}_1)\|_\infty < 1,$$

where $\boldsymbol{\beta}_1$ is the subvector of $\boldsymbol{\beta}^*$ on its support $supp(\boldsymbol{\beta}^*)$, and $\mathbf{X}_1$ and $\mathbf{X}_2$ are the submatrices of the $n \times p$ design matrix $\mathbf{X}$ formed by its columns in $supp(\boldsymbol{\beta}^*)$ and its complement, respectively. However, the irrepsentable condition is easily violated in present of highly correlated variables and therefore very restricted in high dimensions. This explains why the Lasso estimator tend to include many false positive in the selected model [FL10].

[FL01] introduces the oracle property for model selection. Let $S = \{j : \beta_j^* \neq 0\}$ be the index set of important variable. We call the PLS method in (I.2) an oracle procedure if $\hat{\boldsymbol{\beta}}$ satisfies (asymptotically) the following oracle properties:

(1) Consistency of variable selection, $\{j : \hat{\beta}_j \neq 0\} = S$ and

(2) Asymptotic normality, $\sqrt{n}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*) \to_d \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$,

where $\boldsymbol{\Sigma}^*$ is the covariance matrix knowing the true subset model. [FL01] studies the oracle properties of nonconcave penalized likelihood estimators in the finite-dimensional setting. They propose the Smoothly Clipped Absolute Deviation (SCAD) penalty given as follows:

$$\rho_\lambda(t) = \begin{cases} \lambda|t| & for\ |t| \leq \lambda, \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & for\ \lambda < |t| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & for\ |t| > a\lambda, \end{cases} \tag{I.3}$$

where $a > 2$ is a fixed parameter. They show that the local minimum in (I.2) with the SCAD penalty satisfies the oracle properties under some regular conditions. [FP+04] further extends this result to a high-dimensional setting with $p = o(n^{1/5})$ or $p = o(n^{1/3})$. Due to the concavity of the SCAD penalty, it suffers from the multiple minima issue. [KCO08] later shows that with high probability the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ is actually a local minimum of the PLS with SCAD penalty, allowing p to grow with n exponentially. They also provide sufficient conditions to check when a local minimum becomes a global minimum.

[Zou06] shows that the Lasso estimator does not have the oracle properties in general and proposes the adaptive lasso that uses the weighted ℓ_1 penalty,

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^{p} w_j |\beta_j|,$$

where $w_j = 1/|\tilde{\beta}_j|^\gamma$ and $\tilde{\boldsymbol{\beta}}$ is an root-n consistent estimator of $\boldsymbol{\beta}^*$ which serves as an initial estimator for the adaptive Lasso procedure. Note that for any fixed λ, the penalty for zero-initial estimation goes to infinity, while weights for nonzero initials converge to a finite constant. Consequently, by allowing a relatively higher penalty for zero-coefficients and lower penalty for nonzero coefficients, the adaptive lasso is able to reduce the estimation bias and improve variable selection accuracy. Similar to the Lasso, solving for the adaptive Lasso is also a convex optimization problem and thus it does not have the issue of multiple local minima.

For fixed p, [Zou06] proves that the adaptive LASSO has the oracle property. In high dimension setting, for $p \gg n$, [HMZ08] shows that under the partial orthogonality and certain other conditions, the adaptive LASSO obtains variable selection consistency and estimation efficiency, when the marginal regression estimators are used as the initial estimators.

[Z$^+$10] proposes the Minimax Concave Penalty (MCP) that shares a similar spirit as the SCAD penalty. The MCP takes the form

$$\rho_\lambda(t) = sign(t)\lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

with a fixed parameter $b > 0$. It minimizes the maximum concavity

$$\kappa(\rho) := \sup_{0<t_1<t_2} \{\rho'_\lambda(t_1) - \rho'_\lambda(t_2)\}/(t_2 - t_1)$$

subjects to the following unbiasedness and selection features

$$\rho'_\lambda(t) = 0 \quad \text{for } t \geq b\lambda \quad \text{and} \quad \rho'_\lambda(0+) = \lambda.$$

It has been proved that the local minima of the PLS in (I.2) with MCP have the oracle properties under some regular conditions. Specially, [Z$^+$10] proposes the Penalized Linear Unbiased Selection (PLUS) algorithm with MCP to obtain local minimizers that equal the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$, with the probability converging to 1.

The above motioned folded-concave penalty, i.e. the SCAD penalty and the MCP, can be viewed as interpolations between the ℓ_0 penalty and the ℓ_1 (Lasso) penalty. One one hand, the folded-concave penalties possess smoothness over the ℓ_0 penalty to gain flexibility and stability in computations. On the other hand, they can reduce the bias of the Lasso and thus improve model selection accuracy and obtain oracle properties. [FL11] investigates the penalized likelihood approaches using a general class of folded-concave penalty functions in the context of generalized liner model. They demonstrates that such methods have oracle properties with the dimensionality of non-polynomial order of the sample size.

Although these methods enjoy many attractive statistical properties, they do not work well when the covariates are highly correlated or have certain grouping structures. For example, in gene expression analysis, genes from the same biological pathways may have strong correlations. [Tib96a] points out that when there are highly correlated predictor in high-dimensional settings, the prediction performance of the Lasso is dominated by the ridge regression. [ZH05] demonstrates that the Lasso tends to select one variable among a group of highly correlated covariates.

To address these issues, [ZH05] proposes to use the elastic net (Enet) penalty, which is the linear combination of the ℓ_1 and ℓ_2 penalties

$$\rho_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2,$$

where $\lambda_1, \lambda_2 > 0$ are the tuning parameters. The Enet penalty can encourage the sparsity and grouping effects simultaneously. [YL07] and [JY10] investigate its selection consistency in the settings when p is fixed and $p \gg n$, respectively. They show that the Enet estimator is selection consistent under an irrepresentable condition and certain other conditions.

[ZZ09] proposed the adaptive Enet estimator to reduce the asymptotically biasedness caused by the ℓ_1 component, following the same rationale behind the adaptive Lasso estimator. Their oracle results require that the singular values of the design matrix is bounded away from zero and infinity, which excludes the case of highly correlated covariates and only applicable when $p < n$. To overcome these limitations, [HBMZ10] replaces the ℓ_1 component by the MCP and proposes the Mnet approach. They show that the Mnet estimator is selection consistent and equalt to the oracle estimator under some regular conditions, applicable to the situation when $p \gg n$. Similarly, the SCAD-ridge penalty is also studied in [ZX14, DSA18]. The main drawback of these methods is that they essentially treat each variable individually and are not able to incorporate grouping structures among covariates to improve the selection accuracy.

When the p covariates form J non-overlapping groups, the linear regression model in (I.1) can be written as

$$y_i = \sum_{j=1}^{J} \mathbf{x}_{ij}^T \boldsymbol{\beta}_j^* + \epsilon_i, \quad i = 1, \cdots, n. \tag{I.4}$$

Here $\mathbf{x}_{ij}$s are independent and identically distributed (i.i.d) d_j-dimensional covariate vectors corresponding to the jth group, $\boldsymbol{\beta}_j^*$ is the d_j-dimensional true regression coefficient vector of the jth group. Then $p = \sum_{j=1}^J d_j$. Let $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{iJ}^T)^T$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \cdots, \boldsymbol{\beta}_J^{*T})^T$. Since the highly-correlated predictors in the same group tend to be in or out of the model together, the group sparsity condition is often assumed: there exists $S \subseteq \{1, \cdots, J\}$ such that $\boldsymbol{\beta}_j^* = \mathbf{0}$ for all $j \notin S$.

[B$^+$99] first proposes to use the group Lasso (GLasso) and is later developed by [YL06]. The GLasso estimator is defined as a minimizer of (I.2) with the penalty

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^J \sqrt{d_j} \|\boldsymbol{\beta}_j\|_2.$$

As a nature extension of the Lasso, the GLasso selects variables at group level by applying the Lasso penalty on the ℓ_2 norm of coefficients associated with each group of variables. [HZ$^+$10] demonstrates that the GLasso is superior to the Lasso under the strong group sparsity and certain other conditions. While the selection consistency is established under a variant of the irrepresentable condition [Bac08, NR$^+$08], [WH10] shows that the GLasso is not group selection consistent in general and proposes the adaptive GLasso following the same spirit of the standard adaptive Lasso. They show that the adaptive GLasso enjoys the consistency in group selection under some regular conditions, when the group Lasso is used as the initial estimator.

[WCL07] proposes to select groups of time-varying coefficients by the group SCAD

$$\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^J \rho_\lambda(\|\boldsymbol{\beta}_j\|_2),$$

where the scalar version of ρ_λ is the SCAD penalty in (I.3). They also establish the oracle result in fixed dimensional settings. [GZWW15] studies the oracle property of

the Group SCAD in the high-dimensional setting when the number of groups can grow at a certain polynomial rate. Similarly, the computational and theoretical properties of the group MCP estimator are also investigated in [MHW+11, YHZ14].

The above mentioned group penalties essentially penalize the ℓ_2 norm of coefficients associated with each group of variables and thus can only perform variable selection at the group level, not at the individual level. However, this is not appropriate for some situations. For example, in genetic association study, while the variants belong to the same gene form a group, it is not necessary that all variants in the same group are associated with the decease. In such cases, the bi-level sparsity is often assumed: the sparsity can appear either at the group level or within certain groups. [HMXZ09] proposes the group bridge penalty to encourage the bi-level sparsity,

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_j\|_1^\gamma,$$

where $\gamma \in (0, 1)$ is the bridge index and c_j are constants adjustable for the dimension of the group, e.g. $c_j = d_j^\gamma$. The group bridge penalty applies the bridge penalty on the ℓ_1 norm of the coefficients for each group and thus perform bi-level variable selection. [HMXZ09] shows that the global solution of the group bridge enjoys consistency in group selection in low dimensional settings. [HBM12] further proposes the concave ℓ_1 norm penalty

$$\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_1, \sqrt{d_j}\lambda).$$

Here the ρ function is a folded concave penalty, such as the SCAD penalty and the MCP. While the concave ℓ_1 norm penalty does indeed provide the bi-level selection, [SMS20] shows that in general the concave ℓ_1-norm penalty can only perform consistent group selection, not the individual variable-level selection.

[BH09] proposes a framework of the composite penalty that applies an outer penalty ρ_O to the sum of an inner penalty ρ_I, which can be written as

$$\sum_{j=1}^{J} \rho_O \left(\sum_{i=1}^{d_j} \rho_I(\beta_{ij}) \right),$$

where β_{ij} is the ith component of the coefficients vector in jth group. It is easy to verify that the GLasso penalty, the group bridge penalty, the concave ℓ_1-norm penalty and the concave ℓ_2 norm penalty all fit into this framework. To perform bi-level selection, the paper proposes the composite MCP where the penalty ρ_O and ρ_I are the MCP penalty. They also point out that the corresponding composite SCAD penalty displays less grouping effect than the composite MCP. However, no oracle results are available for the composite penalty even under the fixed-dimensional setting.

For other approaches that achieve bi-level selection, see for examples the composite absolute penalty (CAP) [ZRY$^+$09], the hierarchical Lasso [ZZ10], the sparse group Lasso (SGL) [FHT10, SFHT13] and the sparse adaptive group Lasso (adSGL) [FWZ$^+$15].

In this section we have introduced the PLS approaches in three differnt categories: the individual variable selection approaches, the group selection approaches and the bi-level selection approaches. While some of the methods enjoy nice statistical properties, such as estimation consistency and oracle properties, almost all of them require the random error at least to be sub-Gaussian, since the quadratic loss in (I.2) is very sensitive to outliers or heavy-tailed random errors. In addition, most of the statistical results require certain forms of the restricted eigenvalue condition, which may not hold when the predictors are not sub-Gaussian. In this thesis, we propose three different high-dimensional M-estimation frameworks to deal with these issues.

From both the theoretical and computational aspects, we will show that our methods are robust to the irregular settings motioned in I.1.

I.3. Robust Penalized Regression Method

The need for robust methods in statistical inference is widely recognized. Especially in high-dimensional settings, the data unusually suffers from irregularities, such as data contamination or heavy-tailed errors. However, the Penalized Least Squares (PLS) methods are very sensitive to outliers and thus not able to provide robust variable selection and parameter estimation.

[Box53] and [BA55] first bring robustness into the statistical scene. Later [H$^+$64], [Ham68] and [Bic75] lay the comprehensive foundation of the theory of robust statistics. In particular, Huber's seminal work [H$^+$64] establishes the asymptotic property of the M-estimators and proposes a minimax approach for constructing regression functions that are insensitive to deviations from normality. In addition to the classical concept of efficiency, [Ham68] proposes the influential function to describe the local stability of an estimator in the presence of a small proportion of outliers. [DH83] introduces the breakdown point, which represents the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values, to measure the global robustness of an estimator. Since then, many significant steps have been taken toward designing and analyzing robust statistical methods – notably in the work of the Least median of squares (LMS) [Rou84], the Least-trimmed squares (LTS) [Rou84], the S-Estimators [RY84], the MM estimator [Yoh87], among many others.

While the classical robust regression techniques ignore variable selection out of necessity, the advance of technologies on collecting and analyzing high-dimensional data has driven statisticians to work on penalized robust regression approaches. Consider

a high-dimensional regression model in (I.1), due to the sensitivity of the quadratic loss to heavy-tailed errors or outliers, a robust penalized selection and estimation procedure replaces the sum of squares loss in (I.2) by a certain robust loss function . Hence, the corresponding robust estimator $\hat{\boldsymbol{\beta}}$ takes the following form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\boldsymbol{\beta}; Z_1^n) + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{I.5}$$

where $\mathcal{L}_n(\boldsymbol{\beta}; Z_1^n)$ is the empirical loss function, $Z_1^n = \{Z_1, Z_2, \cdots, Z_n\}$ denote a collection of n samples and $Z_i = (\mathbf{x}_i, y_i)$ for $i = 1, \cdots, n$. Note that a penalized robust procedure is characterized by its loss function $\mathcal{L}_n(\boldsymbol{\beta}; Z_1^n)$ and the penalty function encourages a certain sparsity on the parameter vector $\boldsymbol{\beta}$. Compared to the sum of squares loss, a robust loss function is able to accommodate the data's irregularity and the model misspecification. For the rest of this section, we will review some widely used penalized robust approaches.

I.3.1. Penalized Quantile Regression and Its Variants

Since its inception in [KBJ78], the quantile regression (QR) has become a significant and broadly used technique to study the conditional quantiles of a response variable. A penalized quantile regression estimator consider the loss function as follows

$$\mathcal{L}_n(\boldsymbol{\beta}; Z_1^n) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}),$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the check function of [KBJ78] at a given quantile level $0 < \tau < 1$. Suppose the random error ϵ_i in (I.1) satisfies $P(\epsilon_i \leq 0|\mathbf{x}_i) = \tau$ and we ignore the intercept for brevity here. Hence, $\mathbf{x}^T\boldsymbol{\beta}^*$ becomes the $100\tau\%$ quantile of the response y given $\mathbf{x}$. In fact, $\boldsymbol{\beta}^*$ is the population minimizer of the check function

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} E_{y|\mathbf{x}}[\rho_\tau(y - \mathbf{x}^T\boldsymbol{\beta})].$$

Compared to the least squares procedures, robust procedures based on the QR is more resistant to the outliers and the influential points in the response measurement. Theirs unique advantages also lie in the capability to capture data heteroscedasticity through estimates on different quantiles.

The penalized QR approaches have been extensively studied for the last decades. [Koe04] applies the ℓ_1-norm quantile regression (ℓ_1-QR) for longitudinal data to encourage sparsity in estimating the random effect. [LZ08] proposes an efficient algorithm to compute the solution path of the ℓ_1-QR. [WL09] establishes oracle properties of the SCAD and adaptive-Lasso penalized QR for fixed dimension p. [BC$^+$11] investigates the ℓ_1-QR in a high-dimensional setting. They show the estimator is consistent at a near-oracle rate and provide sufficient conditions under which the selected model includes the true model, uniformly over a compact set of quantile indices. [WWL12] considers non-convex penalized QR in an ultra-high dimensional sparse model and demonstrates that the oracle estimator is a local minimum of the non-convex penalized QR, under certain mild assumptions on the error distribution. [FFB14] proposes a weighted ℓ_1-QR estimator and constructs its oracle results and asymptotic normality in an ultra-high dimensional setting.

To obtain a more comprehensive understanding of the response-predictors relationship, [ZY08a] proposes the simultaneous multiple QR (SMQR) method to estimate multiple conditional qunatiles jointly, of which the loss function is

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)}).$$

Here $\boldsymbol{\beta}^{(k)} = (\boldsymbol{\beta}_1^{(k)}, \boldsymbol{\beta}_2^{(k)}, \cdots, \boldsymbol{\beta}_p^{(k)})^T$ be the coefficients vector from the τ_k conditional quantile function of y given $\mathbf{x}$ for $k = 1, 2, \cdots, K$. Note that the above loss function

reduces to the check function when $K = 1$. [ZY08a] penalizes the above loss function by a norm of the coefficient matrix that encourages the column-wise sparsity, of which the penalty is defined as

$$\rho_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \max_k \{|\boldsymbol{\beta}_j^k|\}.$$

Note that the SMQR method is preferable only when it is reasonable to assume the same subset of the predictors are associated with multiple conditional quantile of the response.

[ZY+08b] proposes an adaptive-lasso-penalized composite quantile regression (ACQR) procedure. In that paper the conditional $100\tau\%$ quantile of Y given $\mathbf{x} = \mathbf{x}_i$ is assumed to be

$$\sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j^* + b_\tau^*,$$

where b_τ^* is the $100\tau\%$ quantile of ϵ and uniquely defined for any $0 < \tau < 1$. The loss function for the ACQR method takes the form

$$\sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - b_{\tau_k} - \mathbf{x}_i^T\boldsymbol{\beta}). \tag{I.6}$$

They show that the ACQR method works well for the data contaminated with outliers or generated from infinite-variance errors for fixed-dimensional settings. A weighted version of (I.6) is proposed by [BFW11] termed as the composite quasi-likelihood approaches. Considering the high-dimensional linear model, the loss function of [BFW11] is defined as

$$\sum_{k=1}^{K}\sum_{i=1}^{n} w_k\rho_k(y_i - \mathbf{x}_i^T\boldsymbol{\beta}),$$

where $\rho_1, \cdots, \rho_K$ are the convex functions and $w_1, \cdots, w_K$ are constant weights chosen to minimize the asymptotic variance of the resulting estimator. From the perspective of non-parametric statistics, the convex functions $\rho_1, \cdots, \rho_K$ can be viewed as the

basis functions used to approximate the unknown log-likelihood function of the error distribution. With the weighted ℓ_1 penalty to alleviate the bias generated by the ℓ_1 penalty, they show that the proposed estimator enjoys selection consistency and estimation efficiency for the true non-zero parameters, under certain mild conditions.

It is worth noting that the QR regression becomes the least absolute deviation (LAD) regression when we choose the quantile level $\tau = 0.5$ in the check function. The LAD loss function is defined as follows

$$\sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|.$$

The LAD regression estimates the conditional median function and is well known for its robustness to outliers in the response or heavy-tail errors.

Penalized LAD regression methods haven been studied to perform simultaneous robust estimation and variable selection. [WLJ07] shows that in low-dimensional setting, the LAD-Lasso estimator has the same asymptotic efficiency as the unpenalized LAD estimator obtained under the true model. [GH10] provides sufficient conditions under which the LAD-Lasso enjoys the estimation and selection consistency in a sparse high-dimensional regression model. [Ars12] proposes the weighted LAD-Lasso to address the problem that the LAD-Lasso is not resistant to outliers in covariates. They apply the LAD-Lasso to the transformed data set $(w_i y_i, w_i \mathbf{x}_i)$ for $i = 1, \cdots, n$ where the weights w_i are computed using a certain robust distant in covariates. [Wan13] shows that the LAD-Lasso achieves the near-oracle risk performance with a nearly universal penalty parameter and also establishes its sure screening property for high-dimensional settings.

The penalized QR methods are attractive in that they are resistant to heavy-tail errors or outliers while enjoying oracle results if an appropriate penalty function is

used. They can also capture the data heterocedasticity by jointly estimating multiple conditional quantiles. The main drawback is that they essentially provide the median (quantile) regression instead of mean regression. Using quantile approaches may generate bias respective to the mean estimation when the underlying error distribution is not symmetric. Hence, the penalized QR methods are not applicable in robust high-dimensional regression when the mean estimation is still of interest.

I.3.2. Penalized Robust M-estimator

Define $t_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ as the residual for the ith observation. Recall the PLS method considers the loss function $\sum_{i=1}^n t_i^2$, which produces an unstable result if outliers occur in the data. To reduce the effect of outliers or heavy-tail errors, [H$^+$64] proposes to replace the squared loss by another function of residuals, yielding

$$\sum_{i=1}^{n} l(t_i), \tag{I.7}$$

where $l : \mathbb{R} \mapsto \mathbb{R}$ is the residual function or the loss function. [God60] shows that choosing a loss function l proportional to $\log f_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is the best choice, where $f_{\boldsymbol{\beta}}(\mathbf{x}, y)$is the density function of observations. [H$^+$64] further derives the optimal minimax function l when the model $f_{\boldsymbol{\beta}}(\mathbf{x}, y)$ is only approximately true and calls the solution in minimizing (I.7) an M-estimator. The least squares method takes $l(t) = t^2$ and the LAD method takes $l(t) = |t|$, which are special cases of M-estimators. Note that for some M-estimators, the residual function is applied to a scaled residual instead, such as $l(t_i/\hat{s})$, where the scale estimator $\hat{s}$ can be obtained from a certain robust procedure. We omit it in this introduction for the sake of brevity.

The penalized robust M-estimation approaches have become very popular in robust variable selection and estimation since the last decade. [LLZ$^+$11] points out that LAD approaches suffer a loss of efficiency for normally distributed data and proposes the following loss function with concomitant scale parameter s

$$\mathcal{L}_H(\boldsymbol{\beta}, s) = \begin{cases} ns + \sum_{i=1}^n l_\gamma\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{s}\right) s & for\ s > 0, \\ 2M\sum_{i=1}^n |y_i - \mathbf{x}_i^T\boldsymbol{\beta}| & for\ s = 0, \\ +\infty & for\ s < 0, \end{cases} \tag{I.8}$$

where the residual function l_γ with $\gamma > 0$ is the Huber loss function in [H$^+$64]

$$l_\gamma(t) = \begin{cases} t^2 & for\ |t| \le \gamma, \\ 2\gamma|t| - \gamma^2 & for\ |t| > \gamma. \end{cases} \tag{I.9}$$

Note that γ controls the robustness of the Huber loss in that l_γ applies the quadratic function to smaller errors and the absolute function to larger errors. By combing the adaptive Lasso penalty with the loss function in (I.8), [LLZ$^+$11] shows that the proposed estimator is resistant to the heavy-tailed errors or outliers in response and enjoys oracle properties for fixed dimension p.

[WJHZ13] proposes a class of penalized regression estimators based on the exponential squared loss, of which the residual function is defined as follows

$$l_\gamma(t) = 1 - \exp\left\{-t^2/\gamma\right\},$$

Similarly, $\gamma > 0$ is a tuning parameter that controls the degree of robustness for the estimators. In particular, when γ is large, the summand can be approximated as the quadratic loss and thus the proposed estimator behaves similarly to the PLS

estimator. For a small γ, the observation with large residual yields a bounded loss and therefore has a limited effect on the estimator of $\boldsymbol{\beta}^*$. [WJHZ13] establishes the root-n consistency and oracle properties under defined regularity conditions for fixed dimension p. They also demonstrate that the proposed estimators achieve the highest breakdown point of 1/2 and bounded influence functions with respect to the outliers in either the response or the covariates.

[CRW18] proposes a robust Lasso regression method using Tukey's biweight criterion, of which the residual function takes the form

$$l_\gamma(t) = \begin{cases} \frac{\gamma^2}{6}\left\{1 - \left[1 - \left(\frac{t}{\gamma}\right)^2\right]^3\right\} & for\ |t| \leq \gamma, \\ \frac{\gamma^2}{6} & for\ |t| > \gamma. \end{cases}$$

Here $\gamma > 0$ controls the robustness of the estimator by truncating the residuals that are larger than γ to the constant $\frac{\gamma^2}{6}$, and therefore the impact of the corresponding observation is alleviated. [CRW18] proposes estimator is applied to high-dimensional data where $p > n$ but the corresponding statistical properties are not available.

The above mentioned robust residual function l_γ all share the same characteristics such that their derivative, denoted by ψ_γ, are bounded. It has been shown that the influential function [Ham68] of M-estimators, which measures the influence of an observation on the value of estimated parameter, is proportional to its derivative function ψ_γ. Hence, the bounded ψ_γ alleviates the impact of observations with large residuals and achieves robustness with respect to outlier in the response or heavy-tailed errors. Compared to other loss functions, the Huber loss function is more advantageous in that its convexity yields unique minimization and more stable computations. However, the non-convex loss function, e.g. the exponential loss

function and Tukey's biweight loss function, may achieve stronger robustness through producing redescending M-estimators. In the robust regression literature, we call an M-estimator redescending if the derivative function ψ_γ becomes 0 or decreases to 0 smoothly for all residual greater at some points. In that case, large residuals can be downweighted or ignored completely. See [Mul04] and [SMS08] for more discussions.

[NRW$^+$12] proposes a unified framework of penalized M-estimator for high-dimensional data analysis. They provide sufficient conditions under which the penalized M-estimator is consistent at a certain optimal rate. But they do not provide the oracle properties and require the loss function to be convex. [Loh17] establishes the local estimation consistency and oracle properties for a framework of high-dimensional M-estimators, which allows both the loss function and the penalty function to be non-convex. Although their results are applicable for the heavy-tailed distribution and/or outliers in additive errors and covariates, they do not address the issue of asymmetry and heteroscedasticity.

I.3.3. Outlier Detection for High-dimensional Data Analysis

The presence of outliers may result in biased estimation, model misspecification and misleading predictions. While all the above mentioned approaches perform direct robust estimation against outliers, it is also nature to detect and remove outliers before fitting regression models. Typical approaches for outlier diagnostics are based on refitting the regression model after deleting one case at a time [AS03], These diagnostic methods are helpful in the discovery of outliers, including Cook's distance [Coo77], studentized residuals [Pop76] and jackknifed residuals [VR13], among many others. For high-dimensional models, [ZLL$^+$13] proposes a diagnosis measure called High-dimension Influence Measure (HIM), that uses a marginal correlation to measure observation's influence. [WL17] uses outlier detection measures based on distance

correlation. The work of [RRSY19] studies a few measures for gauging the influence of an observation on Lasso model selection. However, these methods only focus on single-case diagnostics. To deal with multiple influential observations that give rise to the "masking" and "swamping" effects, [ZLNL19] studies two extreme statistics based on a marginal-correlation-based influence measure. [WLCL18] proposed to obtain a clean set using the sure independence screening method and the least trimmed squares regression estimates, followed by the multiple outliers detection through testing procedures.

Another line of research focuses on simultaneous outlier detection and robust estimation via the penalized regression in high-dimensional regression models. Consider the following mean-shift linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \theta_i + \epsilon_i, \quad 1 \leq i \leq n,$$

where the mean-shift parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$ is assumed to be sparse that θ_i is non-zero only when the observation i is an outlier. [LMJ07] proposes the robust Lasso estimator which takes the form

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\theta} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \theta_i) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{i=1}^{n} |\theta_i| \right\}. \tag{I.10}$$

The above Lasso penalties encourage the sparsity on both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Hence, the proposed estimator performs simultaneous outlier detection and variable selection. [SO12] consider a general penalty function on $\boldsymbol{\theta}$ and propose the so-called Θ-IPOD estimator

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \theta_i) + \lambda_2 \sum_{i=1}^{n} \rho(\theta_i) \right\},$$

where $\rho : \mathbb{R} \mapsto \mathbb{R}$ is a penalty function that encourages sparsity on $\boldsymbol{\theta}$ and is allowed to be non-convex. The authors established the connection between the Θ-IPOD estimators and M-estimators. They also applied their estimator to high-dimensional data by considering the sparsity on both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. [XJ13] proposes the sparse robust outlier shrinkage (SROS) estimator which applies the adaptive Lasso penalty and the weighted ridge penalty on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. They show that the SROS estimator enjoys the selection consistency and preserves full asymptotic efficiency for normal data in low-dimensional settings. [NT12] demonstrates that the estimator in (I.10) can faithfully recover both the parameter vector $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ under certain conditions. [KBW18] modifies the estimator in (I.10) by applying the adaptive Lasso penalty on the mean-shift parameter and developed nice theoretical properties for their approach.

I.3.4. Robust High-dimensional Asymmetric Data Analysis

Ever since [H$^+$64] implies that the location estimation based on Huber's method is possibly biased for fixed asymmetric contamination, lots of effort have been made in robust statistics that deal with asymmetric data. Consider a distribution function that is governed by the standard normal density on the set $[-d, d]$ and is otherwise arbitrary, [Col76] studies a class of M-estimator with continuous skew-symmetric ψ functions that vanish outside a certain set $[-c, c]$ and establishes the estimation consistency. For M-estimation in linear regression model, [Car79, CW88] address that the data asymmetry does not affect the slope estimation asymptotically when the error and covariates distributions are independent. However, the case of asymmetric and heteroscedastic errors was not well addressed. While transformation methods (e.g. [BC64]) are extensively used to obtain symmetric and homogeneous errors, such transformations may not exist when both asymmetry and heteroscedasticity are present. Moreover, transformations essentially modify the relationship between

the response and covariates and thus alter the original problem. [Wil97] proposes a regression method based on modeling the error distribution using the S_U distribution in [Joh49]. But the method is not appropriate for inferences on the slop parameters in the presence of both data asymmetry and heteroscedasticity.

Recently, [XC18] proposes a modify Huber function (MHF) to deal with asymmetric data as follows

$$l_{m_1,m_2}(t) = \begin{cases} m_1 t - \frac{1}{2}m_1^2 & for\ t \leq m_1, \\ \frac{1}{2}t^2 & for\ m_1 < u < m_2, \\ m_2 t - \frac{1}{2}m_2^2 & for\ t \geq m_2, \end{cases}$$

where $m_1 = -\frac{2k\gamma}{1+k}$ and $m_2 = \frac{2\gamma}{1+k}$. Here $\gamma > 0$ controls the robustness of the estimator and $k > 0$ is a data-adaptive parameter that accommodates the data asymmetry. When $k = 1$, the proposed MHF is reduced to the Huber loss function. When $k > 1$, the proposed MHF puts more weights to the longer tail one the left side and vice versa. However, the method is only investigated in low-dimensional space.

In high-dimensional regression models, [FLW17] points out that most of penalized robust M-estimators generate bias to the conditional mean regression function for asymmetric data. They proposes the regularized approximate quadratic (RA-Lasso) estimator which uses the Huber loss function in (I.9) but refer $\gamma > 0$ as a diverging parameter that balances the bias and robustness. They establish nice asymptotic properties of the RA-Lasso estimator, and prove its estimation consistency at the minimax rate enjoyed by LS-Lasso. [SZF19] regards this method as a adaptive Huber regression and investigates the theoretical framework that deals with heavy-tailed error with bounded $(1 + \delta)$-moment for any $\delta > 0$.

I.3.5. Robust High-dimensional Group Variable Selection

When there exists certain grouping structures in covariates, it is desirable to select variables at both the group level and the within the group level. However, the PLS methods for group variable selection are not robust to non-normal data and/or data including outliers. To handle outliers in the response, [Lil15] proposes the LAD-GLasso estimator that minimizes the combination between the LAD loss and the group Lasso penalty. That paper also introduces a weighted version of LAD-GLasso estimator to allow outliers in predictors. [WT16] investigates a general penalized M-estimators framework using convex loss functions and concave ℓ_2-norm penalties for the partially linear model with grouped covariates. Under regular conditions, they show that the robust estimator enjoys the oracle property in a high-dimensional setting. But those robust estimators only select variables at group levels. Considering the linear model with grouping structures in (I.4), [WT16] studies the penalized quantile regression estimator to perform robust bi-level selection, which takes the form as follows

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \sum_{j=1}^{J} \mathbf{x}_{ij}^T \boldsymbol{\beta}_j) + \lambda \sum_{j=1}^{J} \left(\|\boldsymbol{\beta}_j\|_1 \right)^{\frac{1}{2}} \right\},$$

where the check function $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ at a given quantile level $0 < \tau < 1$. That paper also establishes the oracle property in low-dimensional settings. However, as we discussed before, estimators based on quantile regression essentially perform median (quantile) regression and thus may generate bias for mean regression.

I.4. Main Contributions

I.4.1. Penalized Weighted Least Squares Method

In Chapter 2, we propose to run sparse robust HDDA and outlier detection in a weighted least squares framework. To be more specific, we relate each observation's

irregularity to a weight value: weights of regular observations being 1 and weights of irregular observation being smaller than 1. In a penalized weighted least squares framework, we introduce a shrinkage rule for the weight vector to perform simultaneous outlier detection, variable selection and robust estimation. Here the term "irregularity" represents a sample's departure from the majority of the observation due to either the heterogeneity or outlying phenomena. We call our model as the PAWLS method in general since the weighted least squares model is considered and a penalization approach is linked to the proposed weight shrinkage rule.

The contribution can be summarized as follows. First, we provide an efficient robust approach for simultaneous outlier detection and variable selection in ultra high-dimensional settings; Second, to our knowledge, this is the first work of obtaining a data-adaptive weight vector estimation using penalization or shrinkage rule in high-dimensional settings; Third, some non-asymptotic oracle properties for weight vector estimation are studied under $p \gg n$ settings; Fourth, we build a unified link between the weight shrinkage rule and the robust M-estimation. This can facilitate the further investigation of M-estimation in $p \gg n$ settings.

I.4.2. Penalized Robust Approximated Quadratic M-estimators

In Chapter 3, We consider high-dimensional linear regression in more general irregular settings: the data can be contaminated or include possible large outliers in both random errors and covariates, the random errors may lack of symmetry and homogeneity. In particular, we investigate both statistical and computational properties of high-dimensional mean regression in the penalized M-estimator framework with diverging robustness parameters. This framework allows both the loss function and the penalty to be non-convex. Our perspective is different from [Loh17] since all loss functions considered in our study converge to a quadratic loss when the corresponding

robustness parameter diverges. To be more specific, we proposed a class of Penalized Robust Approximated quadratic M-estimators (PRAM) to address all irregular settings in (a-c) mentioned above. Inspired by [FLW17], PRAM uses a family of loss functions with a diverging parameter α to control the robustness as well as the discrepancy to the quadratic loss. By controlling the divergent rate of α, PRAM estimators are able to reduce the bias induced by asymmetric error distribution and meanwhile preserve the robustness to approximate the mean estimators. Additionally, we extend the PRAM to a more general setting by relaxing the sub-Gaussian assumption on covariates.

Our theoretical contributions in this chapter include the investigation of statistical properties for a class of PRAM estimators with only weak assumptions on both random errors and covariates. In particular, We first introduce sufficient conditions under which a PRAM estimator has local estimation consistency at the same rate as the minimax rate enjoyed by the LS-Lasso. We then show that the PRAM estimator actually equals the local oracle solution with the correct support if an appropriate non-convex penalty is used. Based on this oracle result we further establish the asymptotic normality of the PRAM estimators. As we will see, with the devise of diverging parameters in the loss functions, our theoretical result is applicable for a wide class of PRAM estimators which are robust to general irregular settings, when the dimensionality of data grows with the sample size at an almost exponential rate.

Computationally, we also implement the PRAM estimator through a two-step optimization procedure and investigate the performance of six PRAM estimators generated from three types of loss function approximation (the Huber loss, Tukey's biweight loss and Cauchy loss) combined with two types of penalty functions (the Lasso and MCP penalties). While our numerical results demonstrate satisfactory finite sample performance of the PRAM estimators under general irregular settings,

it suggests that in practice, when the data are heavy-tailed or contaminated, a well-behaved PRAM estimator can be chosen by considering a redescending loss function approximation and a concave penalty, using the RA-Lasso as an initial.

I.4.3. High-dimensional M-estimation for Bi-level Variable Selection

In Chapter 4, we consider high-dimensional linear regression with grouped covariates, in irregular settings that the data (random errors and/or covariates) may be contaminated or heavy-tailed. In particular, we propose a novel high-dimensional bi-level variable selection method through a two-stage penalized M-estimator framework: penalized M-estimation with a concave ℓ_2-norm penalty achieving the consistent group selection at the first stage, and a post-hard-thresholding operator to achieve the within-group sparsity at the second stage. Our perspective at the first stage is different from [WT16] since we allow the loss function to be non-convex and thus it is more general. In addition, our proposed two-stage framework is able to separate the groups selection and the individual variables selection efficiently, since the post-hard-thresholding operator at the second stage nearly poses no additional computational burden to the first stage. More importantly, our framework includes a wide range of M-estimators with strong robustness if a redescending loss function is adopted. Furthermore, we extend our framework to a more general setting by relaxing the sub-Gaussian assumption enforced on covariates.

Theoretically, we investigate statistical properties of our proposed two-stage framework with weak assumptions on both random errors and covariates. We first show that with certain mild conditions on the loss function, a penalized M-estimator at the first stage has the local estimation consistency at the minimax rate enjoyed by the LS-GLasso. We further establish that with an appropriate group concave ℓ_2-norm penalty, the estimator from our first stage has a group-level oracle property. We

then show that these nice statistical properties can be carried over directly to the post-hard-thresholding estimators at the second stage and thus we establish its bi-level variable selection consistency. As we will reveal later, those theoretical results are applicable when the data are heavy-tailed or contaminated, allowing the dimensionality of data grows with the sample size at an almost exponential rate.

Computationally, we propose to implement an efficient algorithm through a two-step optimization procedure. We compare the performance of estimators generated from different types of loss functions (e.g. the Huber loss and Cauchy loss) combined with a concave penalty (e.g. MCP penalty). Our numerical results demonstrate satisfactory finite sample performances of the proposed estimators under different settings. Additionally, it also suggests that a well-behaved two-stage M-estimator can be usually obtained by considering a redescending loss (e.g. Cauchy loss) with a concave penalty, when the data are heavy-tailed or strongly contaminated.

CHAPTER II

PENALIZED WEIGHTED LEAST SQUARES METHOD

II.1. Introduction

High-dimensional data arise in many scientific areas due to the rapid development of advanced technologies. In recent decades, a great number of statistical methods, algorithms and theories have been developed to perform high-dimensional data analysis (HDDA). Among them, penalized least squares (PLS) methods have become very popular in high-dimensional linear regression analysis since the introduction of the Lasso [Tib96a]. However, a penalized least squares approach may lose its efficiency and produce unstable result in both estimation and variable selection due to the existence of either outliers or heteroscedasticity. Although many robust analysis tools were proposed in low-dimensional data analysis and also extended in high-dimensional data settings, most of them do not identify outliers in particular, which themselves can provide important scientific findings. Most of existing outliers detection methods, such as visualizing tools or diagnosis statistics, can fail due to the masking and swamping phenomena in presence of multiple outliers. For a HDDA method with separate outliers detection and variable selection process, the problem became more complicated since the damage of high-dimensionality and data contamination can be intertwined.

In this chapter, we aim to introduce a shrinkage rule for the weight vector to perform simultaneous outliers detection, variable selection and robust estimation in a penalized weighted least squares framework. The rest of this chapter is organized as follows. In Section II.2, we introduce the basic setup and define the PAWLS model,

along with a brief discussion of its Bayesian understanding. We also establish a unified link between the PAWLS and a regularized robust M-estimation in this section. We discuss the PAWLS implementation, including both the Algorithm and tuning parameter selection in Section II.3. Some non-asymptotic oracle inequalities of the PAWLS estimation error for both the weights and coefficients vectors are discussed in detail in Section II.4. In Section II.5, we conduct some numerical studies including some simulation studies and real data analysis under both $p < n$ and $p \gg n$ settings.

II.2. Weight Shrinkage

Consider a weighted linear regression

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \eta_i, \quad 1 \le i \le n, \tag{II.1}$$

where y_i and $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ip})'$ are the observed response variable and covariates vector, $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \cdots, \beta_p^*)'$ is the coefficients vector, η_i is the random error with mean 0 and variance σ_i^2. In particular, we let $\sigma_i = \sigma/w_i^*$ for $0 \le \sigma < \infty$. We make an important assumption that the majority number of w_i^*s are 1, except a few others. Thus, the heteroscedasticity or irregularity only exists among a few observations. Such a model assumption is defined as the *irregularity sparsity* in this Chapter.

If the weight vector $\mathbf{w} = (w_1, \cdots, w_n)'$ in (II.1) is given or represented as *a priori*, then we can obtain a sparse estimation of $\boldsymbol{\beta}$ by minimizing a penalized weighted least squares loss with a penalty on $\boldsymbol{\beta}$ (no penalty on intercept),

$$\tilde{\boldsymbol{\beta}}(\lambda_{1n}, \mathbf{w}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n} w_i^2(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + P_{\lambda_{1n}}(\boldsymbol{\beta}). \tag{II.2}$$

For example, an LAD-Lasso takes $w_i = |y_i - \mathbf{x}_i'\boldsymbol{\beta}|^{-1/2}$ and $P_{\lambda_{1n}}(\boldsymbol{\beta}) = \lambda_{1n}\sum_{j=1}^{p}|\beta_j|$ [GH10], [WLJ07], [Wan13]. A sparse LTS [ACG$^+$13] takes $w_i = 0$ for some selected outliers and $w_i = 1$ for others. In some heterscedacity settings, w_i is chosen to be

smaller for clusters with larger variation and larger for clusters with smaller variation.

However, in general, $\mathbf{w}$ is unknown and needed be estimated data-adaptively with $\boldsymbol{\beta}$. In the PAWLS approach we develop here, we allow weights to be data-driven and propose to obtain $\hat{\mathbf{w}}$ and $\hat{\boldsymbol{\beta}}$ simultaneously. In particular, a PAWLS method with the Lasso penalty is to solve

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}})(\lambda_{1n}, \lambda_{2n}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, 0<w_i\leq 1}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^{n} w_i^2(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_{1n}\sum_{j=1}^{p}|\beta_j| + \lambda_{2n}\sum_{i=1}^{n}|1-w_i| \right\}, \tag{II.3}$$

where $\lambda_{1n}\sum_{j=1}^{p}|\beta_j|$ is to encourage the model sparsity by shrinking all coefficients to 0, while $\lambda_{2n}\sum_{i=1}^{n}|1-w_i|$ is to encourage the irregularity sparsity by shrinking all weights from some small amount to 1. Here $\lambda_{1n} \geq 0$ and $\lambda_{2n} \geq 0$ are two tuning parameters controlling the size of a sparse model and the ratio of irregular observations, respectively.

Remark 1: The non-differentiability of penalty $|1-w_i|$ over $w_i = 1$ implies that some of the components of $\widehat{\mathbf{w}}$ may be exactly equal to one. Thus those observations corresponding to $\widehat{w}_i = 1$ survive the irregularity screening, while those corresponding to $\widehat{w}_i \neq 1$ are suspected to be irregular observations. Therefore, the PAWLS can perform simultaneous robust variable selection and irregular or outlying observation detection.

There is a Bayesian understanding of the PAWLS model in (II.3). Suppose we have independent prior distributions: $\beta_0 \propto 1$, $\pi(\beta_j) \propto e^{-\lambda_{10}|\beta_j|}$ for $1 \leq j \leq p$, and $\pi(w_i) \propto (w_i)^{-1}e^{-\lambda_{20}|1-w_i|}I(0 < w_i \leq 1)$ for $1 \leq i \leq n$, where $I(\cdot)$ is the indicator function. The joint posterior distribution of the parameters,

$$\pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y}) \propto \prod_{i=1}^{n} \exp\left\{-w_i^2(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 - \lambda_{20}|1-w_i|\right\} \prod_{j=1}^{p} \exp\left\{-\lambda_{10}|\beta|_j\right\}.$$

Thus the PAWLS estimation $(\hat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})$ in (II.3) with $\lambda_{1n} = \lambda_{10}/(2n)$ and $\lambda_{2n} = \lambda_{20}/(2n)$ is equivalent to a corresponding posterior mode of $\boldsymbol{\beta}$ and $\mathbf{w}$. In the left panel of Figure II.1, we plot three sample curves of $\pi(w_i)$ for $\lambda_{20} = 4, 8, 15$. It is observed that, $w_i = 1$ with a large probability for a large λ_{20}, and $w_i = 0$ with a large probability for a small λ_{20}. The convexity of $\pi(w_i)$ between 0 and 1 justifies the outlier detection ability of the PAWLS in (II.3) from a Bayesian perspective.

II.2.1. A General Threshold Rule and Its Link to Sparse M-estimation

In fact, the PAWLS with Lasso in (II.3) can be generalized to a series of weight shrinkage estimation which enjoys strong robustness. To understand this property, we first define a class of *scale* shrinkage rule as follows.

Definition II.1. (Scale Threshold Function) For any threshold parameter $\lambda > 0$, a positive function $\Theta_\lambda(t),\ t \in \mathbf{R}$ is defined to be a scale threshold function if it satisfies

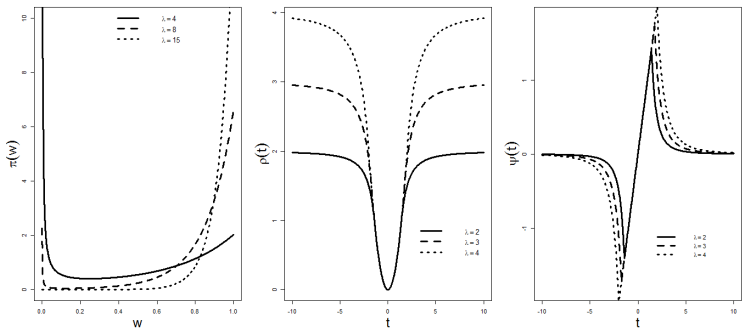
(1) (Symmetric) $\Theta_\lambda(t) = \Theta_\lambda(-t)$,

(2) (Non-increasing) $\Theta_\lambda(t) \geq \Theta_\lambda(t')$ for $0 \leq t \leq t'$ and

(3) (Two extremes) $\lim_{t\to 0} \Theta_\lambda(t) = 1$ and $\lim_{t\to\infty} \Theta_\lambda(t) = 0$.

The scale threshold function in Definition II.1 shares the similar spirit as one in [SO12], but these two types threshold functions have different features. Specifically, $\Theta_\lambda(\cdot)$ here is designed to shrink any small positive values (close to 0) to 1, while the one in [SO12] is to shrink any large values to 0. Based upon the above scale shrinkage rule, we can establish an interesting connection between the PAWLS estimation and the sparse M-estimation. Such a connection explains strong robustness properties of the proposed PAWLS in (II.3).

Figure II.1. Display of Some Functions. Left: The Shape of $\pi_\lambda(w_i)$ Function with $\lambda = 4, 8, 15$; Middle: The ρ_λ Function with Tuning Parameter $\lambda = 2, 3, 4$; Right: The ψ_λ Function with Tuning Parameter $\lambda = 2, 3, 4$



Theorem II.2. *Suppose* $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(0, \tilde{\mathbf{w}})$ *is a solution in* (II.2) *for* $\lambda_{1n} = 0$ *and* $\widetilde{w}_i^2 = \Theta_\lambda(y_i - \mathbf{x}_i\tilde{\boldsymbol{\beta}})$, $1 \le i \le n$. *Here* $\Theta_\lambda(\cdot)$ *for some* $\lambda > 0$ *is a threshold function defined in Definition II.1. Then* $\tilde{\boldsymbol{\beta}}$ *is also an M-estimator such that* $\tilde{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho_\lambda(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. *In particular,* $\psi_\lambda(t) = \frac{d\rho_\lambda(t)}{dt}$ *satisfies,*

$$\psi_\lambda(t) = t\Theta_\lambda(t). \tag{II.4}$$

The proof of Theorem II.2 is given in Appendix. Theorem II.2 tells us that a weight generated from any given scale threshold rule can be linked to a corresponding M-estimator. For example, the PAWLS with the Lasso in (II.3) indicates that $\hat{w}_i = \{n\lambda_{2n}/(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\} \wedge 1$. Thus, if we let $\lambda = n\lambda_{2n}$, then the scale shrinkage rule for (II.3) becomes

$$\Theta_\lambda(t) = \begin{cases} \lambda^2/t^4 & if\ t^2 > \lambda, \\ 1 & if\ t^2 \le \lambda. \end{cases} \tag{II.5}$$

From Theorem II.2, the PAWLS estimation in (II.3) is linked to a corresponding sparse M-estimator with ψ function with

$$\psi_\lambda(t) = \begin{cases} \lambda^2/t^3 & if\ t^2 > \lambda, \\ t & if\ t^2 \leq \lambda, \end{cases} \tag{II.6}$$

and the corresponding ρ function,

$$\rho_\lambda(t) = \begin{cases} -\lambda^2/(2t^2) + \lambda, & if\ t^2 > \lambda, \\ t^2/2, & if\ t^2 \leq \lambda. \end{cases} \tag{II.7}$$

See the middle and right panels in Figure II.1 for three curves of $\rho_\lambda(t)$ and $\psi_\lambda(t)$ under $\lambda = 2, 3, 4$. Notice that $\lim_{t\to\infty} \psi_\lambda(t) = 0$ and $\lim_{t\to\infty} \rho_\lambda(t) = \lambda$. Thus the ρ function in (II.7) gives a weakly redescending M estimation with strong robustness. Naturally, the PAWLS solution in (II.3) can be understand as a regularized robust M-estimator with the Lasso penalty. From now on, our investigation is focused on this particular PAWLS estimator. Without being addressed in particular, the Lasso penalty is used in the PAWLS approach.

II.3. Implementation

II.3.1. Coordinate Decent Algorithm for PAWLS

We first notice that (II.3) is not a convex optimization problem. This is not surprising due to the link to a regularized redescending M estimator and strong robustness discussed in Section II.2.1. However, for a given $\mathbf{w}$, the function of $\boldsymbol{\beta}$ is a convex optimization problem, and the vice versa. Therefore, the objective function (II.3) is a bi-convex function. This biconvexity guarantees that the algorithm has promising convergence properties [GPK07]. We can compute a PAWLS estimate efficiently in Algorithm 1 using coordinate decent algorithm [GPK07].

For each pair of $(\lambda_{1n}, \lambda_{2n})$, those initialization values $\boldsymbol{\beta}^{(1)}$, $\mathbf{w}^{(1)}$ play important roles during alternative iterative process. We suggest to use a multiple iterative strategy

as follows: (1) when updating $\boldsymbol{\beta}$, we start from $\boldsymbol{\beta}^{(1)} = \mathbf{0}$ and $\mathbf{w}^{(1)} = \hat{\mathbf{w}}(\lambda_{1n}, \widetilde{\lambda}_{2n})$, where $\widetilde{\lambda}_{2n}$ is an ideal tuning parameter searched from the last tuning parameter selection process to be represented in the next section; (2) when updating $\mathbf{w}$, we start from $\mathbf{w}^{(1)} = \mathbf{1}$ and $\boldsymbol{\beta}^{(1)} = \hat{\boldsymbol{\beta}}(\widetilde{\lambda}_{1n}, \lambda_{2n})$, where $\widetilde{\lambda}_{1n}$ is an ideal tuning parameter from the last tuning parameter selection process. Thus, initial values are improved for multiple times, and $\boldsymbol{\beta}^{(k)}$ and $\mathbf{w}^{(k)}$ are alternatively updated until converge.

Algorithm 1 The PAWLS under fixed λ_{1n} and λ_{2n}

Given $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$ and $\lambda_{1n}, \lambda_{2n}$ in a fine grid,
  let $\lambda_{1j} = \lambda_{1n}$ for $1 \le j \le p$, let $\lambda_{2i} = \lambda_{2n}$ for $1 \le i \le n$
  let $k = 1$ and obtain an initial $\boldsymbol{\beta}^{(\boldsymbol{k})}$, $\mathbf{w}^{(k)}$, and $\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
While not converged **do**
  [*update* $\boldsymbol{\beta}$]
  $c_j = n^{-1}\mathbf{X}_j'\mathbf{w}^{(k)'}\mathbf{w}\mathbf{X}_j$, $z_j = n^{-1}\mathbf{X}_j'\mathbf{w}^{(k)'}\mathbf{w}\mathbf{r} + c_j\boldsymbol{\beta}_{\boldsymbol{j}}^{(\boldsymbol{k})}$
  $\boldsymbol{\beta}_{\boldsymbol{j}}^{(\boldsymbol{k+1})} = S(z_j, \lambda_{1j})^1/c_j$
  $\mathbf{r} = \mathbf{r} - \mathbf{X}_j'(\boldsymbol{\beta}^{(\boldsymbol{k+1})} - \boldsymbol{\beta}^{(\boldsymbol{k})})$
  [*update* $\mathbf{w}$]
  if $r_i^2 > n\lambda_{2i}$, $\mathbf{w}_i^{(k+1)} \leftarrow n\lambda_{2i}/\mathbf{r}_i^2$; otherwise $\mathbf{w}_i^{(k+1)} \leftarrow 1$
  converged if $\|\boldsymbol{\beta}^{(\boldsymbol{k+1})} - \boldsymbol{\beta}^{(\boldsymbol{k})}\|_\infty < \epsilon$ and $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_\infty < \epsilon$
  $k \leftarrow k + 1$
end while
deliver $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(k)}$ and $\widehat{\mathbf{w}} = \mathbf{w}^{(k)}$

II.3.2. Tuning Parameter Selection

Like many other penalized regression, the selection of tuning parameters plays an important role in producing a well-behaved PAWLS estimate. Due to the high computation efficiency of Bayesian Information Criterion (BIC) [S+78], we choose two optimal tuning parameters λ_{1n}^{opt} and λ_{2n}^{opt} by modifying BIC as follows,

$$\mathbf{BIC}(\lambda_{1n}, \lambda_{2n}) = n\log\left\{\sum_{i=1}^{n}\hat{w}_i^2(\lambda_{1n}, \lambda_{2n})(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}(\lambda_{1n}, \lambda_{2n}))^2 + \frac{p}{n+p}\right\} + \widehat{s}(\lambda_{1n}, \lambda_{2n})\log(n), \tag{II.8}$$

[1] $S(z, a) = z - a, 0$ or $z + a$ if $z > a, |z| \le a$ or $z < -a$.

where $\widehat{s}(\lambda_{1n}, \lambda_{2n}) = \widehat{s}_1 + \widehat{s}_2$ with $\widehat{s}_1 = 1 + \#\{1 \leq j \leq p : \hat{\beta}_j(\lambda_{1n}, \lambda_{2n}) \neq 0\}$ and $\widehat{s}_2 = \#\{1 \leq i \leq n : \hat{w}_i(\lambda_{1n}, \lambda_{2n}) < 1\}$. Here $\widehat{s}_1$ and $\widehat{s}_2$ are the estimated number of nonzero regression coefficients and and outliers, respectively. Different from the classical BIC, we include a term $\frac{p}{n+p}$ in the first part in (II.8) dealing with the possible blowup. This may happen if a very small λ_{1n} is used such that all $\widehat{w}_i$s are close to 0.

The optimal tuning parameters are searched alternatively by minimizing BIC in (II.8) from a fine grid of λ_{1n}, λ_{2n}. We first fix λ_{1n}^* and find an "ideal" λ_{2n}^* using BIC; then this λ_{2n}^* is fixed, and we continue to search an "ideal" λ_{1n}^* by minimizing the BIC. The same procedure is repeated iteratively until an optimal pair $(\lambda_{1n}^{opt}, \lambda_{2n}^{opt})$ is obtained. This alternative search has high computation efficiency and performs well in our numerical studies.

Remark 2: We suggest to search for λ_{2n} first since a well chosen λ_{2n}^* (for outlier screening) at the beginning can reduce the estimation damage caused by outliers during the iteration process significantly. This is also verified by our limited numerical experience.

Remark 3: We discard those $(\lambda_{1n}, \lambda_{2n})$ such that $\widehat{s}_2/n \geq r$, where r can be any value larger than 0.5. *This is reasonable since any single linear regression model will be invalid if a data has more than 50% outliers. In this case, subgroup analysis should be applied. In our numerical studies, we takes $r = 0.8$. In fact, we have also tried different values between $r = 0.5$ to* 0.8. *All worked very well and improved the efficiency of the tuning parameter selection process significantly.*

II.3.3. Improve the PAWLS Using the Adaptive Penalty

Since the adaptive Lasso in general has better variable selection properties than the Lasso [Zou06, HMZ08], we also consider the PAWLS with the adaptive Lasso penalty by minimizing

$$\frac{1}{2n}\sum_{i=1}^{n} w_i^2(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_{1n}\sum_{j=1}^{p}|\beta_j|/|\beta_j^{(0)}| + \lambda_{2n}\sum_{i=1}^{n}|1-w_i|/|1-w_i^{(0)}|, \qquad \text{(II.9)}$$

where $w_i^{(0)}$ and $\beta_j^{(0)}$ are two initial estimates of w_i and β_j, respectively. The computation of (II.9) is similar to Algorithm 1 by replacing λ_{1j} by $\lambda_{1n}/|\beta_j^{(0)}|$ for $1 \le j \le p$ and λ_{2i} by $\lambda_{2n}/|1-w_i^{(0)}|$ for $1 \le i \le n$. By convention, $w_i^{(0)} = \min\{w_i^{(0)}, 0.999\}$ and $\beta_j^{(0)} = \min\{\beta_j^{(0)}, 0.001\}$. If all $0 \le w_i^{(0)} < 1$ and $\beta_j^{(0)}$ for $1 \le j \le p$ are the same, respectively, then (II.9) becomes the PAWLS in (II.3).

As we know, a estimation consistent initials need to be applied in order to have an variable selection consistent adaptive Lasso estimator [Zou06, HHM08]. From those non-asymptotic properties investigated in Section II.4, the PAWLS-Lasso estimates are reasonable choices for $\beta_j^{(0)}$ and $w_i^{(0)}$ in (II.9). From our empirical experiences, the above procedure works very well in all our numerical studies in section II.5.

II.4. Non-asymptotic Properties

In this section, we will investigate the estimation properties of the PAWLS in ultra high-dimensional settings when $p = O\left(\exp(n^{\alpha})\right)$ for some $0 \le \alpha < 1$. To simplify the presentation, we omit the intercept in model (II.1) in this section. All proofs are given in Appendix.

For notation's convenience, we replace $\nu_i = 1 - w_i$ for $1 \le i \le n$ in some scenarios and assume all covariates to be standardized such that $\sum_{i=1}^{n} x_{ij}^2 = n,\ 1 \le j \le n$ in this section. We put all weights and covariates coefficients together and denote a $n+p$ dimensional unknown parameters vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1 = (\beta_1, \cdots, \beta_p)'$ with

true values $\boldsymbol{\theta}_1^* = \boldsymbol{\beta}^*$ and $\boldsymbol{\theta}_2 = (\lambda_{2n}/\lambda_{1n})(\nu_1, \cdots, \nu_n)'$ with true values $\boldsymbol{\theta}_2^* = (\lambda_{2n}/\lambda_{1n})\mathbf{w}^*$. Here $\mathbf{w}^* = (w_1^*, \cdots, w_n^*)'$. Let $S_{10} = \{1 \le j \le p : \beta_j^* \neq 0\}$ with the cardinal value $s_1 = |S_{10}|$, $S_{20} = \{1 \le i \le n : w_i^* < 1\}$ with the cardinal value $s_2 = |S_{20}|$, and $J_0 = \{1 \le k \le n+p,\ \theta_k^* \neq 0\}$ be the true active set for $\boldsymbol{\theta}^*$ with the cardinal value $|J_0| = s_1 + s_2 = s$. We also denote $a_n = \min_{i \in S_{20}} w_i^*$.

We consider the fixed design such that $|x_{ij}| \le b_n$ for all i and j and the following assumptions.

(A1): $\epsilon_i = w_i^* \eta_i$ are i.i.d. sub-Gaussian distribution with mean 0 and scale factor $\sigma > 0$.

(A2): (i) $\dfrac{sb_n}{n^{1/2}} = o(1)$; (ii) $\dfrac{s\log(n)}{na_n^2} = o(1)$.

(A3): there exists a constant $M > 0$ such that $\max_{j \in S_{10}} |\beta_j^*| < M$.

RE(s, c)**:** For some integer s, such that $1 \le s \le p + n$, and a positive c, the following restricted eigenvalue condition holds:

$$\kappa(s, c) = \min_{\substack{\mathbf{d} \neq 0 \\ \|\mathbf{d}_{J_0^c}\|_1 \le c\|\mathbf{d}_{J_0}\|_1 \\ |J_0| \le s}} \frac{\|\boldsymbol{\Psi}^{1/2}\mathbf{d}\|_2}{\|\mathbf{d}_{J_0}\|_2} > 0, \tag{II.10}$$

where $\|\cdot\|_q$ is the ℓ_q norm, $\mathbf{d} = (\mathbf{d}_1', \mathbf{d}_2')'$ and $\boldsymbol{\Psi} = \dfrac{1}{n}\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \sigma^2\boldsymbol{\Omega}^{*-2} \end{pmatrix}$ with $\boldsymbol{\Omega}^*$ being a diagonal matrix generated from $\mathbf{w}^*$.

From (A1), the standard deviation of y_i, $\sigma_{y_i} = \sigma/w_i^* \to \infty$ if $w_i^* \to 0$ for $i \in S_{20}$. Thus (A1) relaxes the normal assumption on random error in PLS regression dramatically. (A3) is a trivial condition on nonzero regression coefficients. A2(i-ii) indicate that the total number of non-zero β_j^*s and outliers cannot grow with n too fast. It also means a_n can not decay to 0 too fast. If both a_n and b_n are constants, then (ii) is redundant.

The RE(s, c) condition mimics the restricted eigenvalue condition (3.1) of [BRT09]. Consider the following three events regarding the random error ϵ,

- $\mathbb{A}_1 = \{\|\epsilon' \mathbf{X}\|_\infty < n\lambda_{1n}/4\}$;
- $\mathbb{A}_2 = \{\max_{1 \le i \le n} \epsilon_i^2 / w_i^* < n\lambda_{2n}/4\}$;
- $\mathbb{A}_3 = \{\|\epsilon' \mathbf{D}_{\tilde{\nu}} \mathbf{X}\|_\infty < n\lambda_{1n}/4\}$, where $\mathbf{D}_{\tilde{\nu}}$ is a diagonal matrix consists of any estimation $\tilde{\nu} = (\tilde{\nu}_1, \cdots, \tilde{\nu}_n)'$.

We have following results on those three events.

Lemma II.3. *On event* $\mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3$,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \le 4\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1 \tag{II.11}$$

Lemma II.4. *Under (A1), we have*

$$P(\mathbb{A}_1^c) \le 2p \exp\left\{-\frac{n\lambda_1^2}{32\sigma^2}\right\}, \tag{II.12a}$$

$$P(\mathbb{A}_2^c) \le 2n \exp\left\{-\frac{n\lambda_{2n} a_n^2}{8\sigma^2}\right\}, \tag{II.12b}$$

$$P(\mathbb{A}_3^c) \le 2 \exp\left\{-M_0 \min\left\{\frac{n\lambda_{1n}^4}{256K^2\sigma^4}, \frac{n\lambda_{1n}^2}{16K\sigma^2}\right\}\right\}, \tag{II.12c}$$

where $K = \sup_{q \ge 1} q^{-1} \left[E\left(\epsilon_1^2/\sigma^2\right)^q\right]^{1/q}$ *and* $M_1 > 0$ *is an absolute constant. In particular, if we choose* $\lambda_{1n} \ge \sigma(c_1)^{1/2}(\ln(p)/n)^{1/2}$ *for* $c_1 > 32$, *then*

$$P(\mathbb{A}_1^c) \le 2p^{-c_1/32} \to 0 \text{ when } p \to \infty.$$

If we choose $\lambda_{2n} \ge \sigma^2 c_2 \log(n)/(na_n^2)$ *for some* $c_2 > 8$, *then*

$$P(\mathbb{A}_2^c) \le 2n^{1-c_2/8} \to 0 \text{ when } n \to \infty.$$

For the above λ_{1n},

$$P(\mathbb{A}_3^c) \leq O\left(\exp\left\{-\frac{c_1 M_0 \log(p)}{16K}\min\left\{\frac{c_1\log(p)}{16Kn}, 1\right\}\right\}\right).$$

Thus if $p = O(exp(n^\alpha))$ *for* $\alpha > 0$, *then* $P(\mathbb{A}_3^c) \to 0$ *for* $\alpha \geq 1/2$.

Lemma II.3 provides an upper bound of the PAWLS estimator under three events. Lemma II.4 investigates the lower probability bounds for the occurrence of those events. We now develop the theoretical properties of the proposed PAWLS estimator. In particular, we expect to obtain some non-asymptotic oracle inequalities for both $\hat{\mathbf{w}}$ and $\hat{\boldsymbol{\beta}}$.

Theorem II.5. *Suppose A1 and RE(s,3) hold. Then with probability at least* $1 - \sum_{k=1}^{5} h_i$, *we have*

$$\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1 \leq \frac{8\lambda_{1n}s}{\kappa(s,3)^2}$$

and

$$\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_2 \leq \frac{8\lambda_{1n}s^{1/2}}{\kappa(s,3)^2},$$

Here

$$h_1 = 2p_n \exp\left\{-\frac{n\lambda_{1n}^2}{32\sigma^2}\right\},$$

$$h_2 = 2n \exp\left\{-\frac{n\lambda_{2n}a_n^2}{8\sigma^2}\right\},$$

$$h_3 \quad = 2\exp\left\{-M_0 \min\{\frac{n\lambda_{1n}^4}{256K^2\sigma^4}, \frac{n\lambda_{1n}^2}{16K\sigma^2}\}\right\} \; with \; K = \sup_{q\geq 1}\frac{1}{q}\left[E\left(\frac{\epsilon_1^2}{\sigma^2}\right)^q\right]^{1/q}$$

and $M_1 > 0$ *is an absolute constant*,

$$h_4 = \frac{48\sigma}{\kappa(s,3)}\frac{\lambda_{1n}(1+\log(2n))^{1/2}}{\lambda_{2n}}\frac{s^{1/2}}{a_n n^{1/2}},$$

$$h_5 = \frac{384\sigma}{k^2(s,3)} \frac{\lambda_{1n}(1+\log(2n))^{1/2}}{\lambda_{2n}} \frac{sb_n}{na_n}.$$

In particular, if (A2) and (A3) hold and $\lambda_{1n}/\lambda_{2n} \leq O(1)$*, then* $h_4 = o(1)$ *and* $h_5 = o(1)$.

Theorem II.5 gives the oracle inequalities of joint estimators of $\boldsymbol{\theta}$. Those properties are similar to ones for the PLS estimator (with the Lasso penalty) of $\boldsymbol{\beta}$ only when $\mathbf{w}^*$ is given in advance. When $\mathbf{w}$ is jointly estimated with $\boldsymbol{\beta}$, the non-asymptotic properties for both $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ can be obtained by letting two regularization parameters λ_{1n} and λ_{2n} changes with n dependently such that $\lambda_{1n}/\lambda_{2n} = O(1)$.

The following corollary provides an explicit, shared rate of λ_{1n} and λ_{2n} such that both $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ are estimation consistent even though p grows with n at an almost exponential rate. This is a direct result from Lemma II.4 and Theorem II.5.

Corollary II.6. *Suppose* $p = O\left(\exp(n^{\alpha})\right)$ *for* $1/2 < \alpha < 1$ *and all assumptions in Theorem II.5 hold except that A2(ii) is replaced by* $s = o\left(n^{(1-\alpha)/2}\right)$*. If we can choose* $\lambda_{1n} \geq \sigma(c_1)^{1/2}(\ln(p)/n)^{1/2}$ *for* $c_1 > 32$*, and* $\lambda_{2n} \geq \sigma^2 c_2 \log(n)/(na_n^2)$ *for some* $c_2 > 8$ *such that* $\lambda_{1n} = \lambda_{2n}$*, then with probability at least* $1 - 2p^{1-c_1/32} - 2n^{1-c_2/8}$*, we have*

$$\|\hat{\boldsymbol{\beta}}_{S_{10}} - \boldsymbol{\beta}^*_{S_{10}}\|_1 + \|\hat{\mathbf{w}}_{S_{20}} - \mathbf{w}^*_{S_{20}}\|_1 \leq \frac{8\lambda_{1n}s}{\kappa(s,3)^2}$$

and

$$\|\hat{\boldsymbol{\beta}}_{S_{10}} - \boldsymbol{\beta}^*_{S_{10}}\|_2 + \|\hat{\mathbf{w}}_{S_{20}} - \mathbf{w}^*_{S_{20}}\|_2 \leq \frac{8\sqrt{2}\lambda_{1n}s^{1/2}}{\kappa(s,3)^2}.$$

II.5. Numerical Result

In this section, we demonstrate the performance of the PAWLS using both simulation studies and real data analysis under two settings: $p < n$ and $p \gg n$.

II.5.1. Simulation Studies

In all our simulation studies, the data are generated from the mean shift model without an intercept:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma_i + \epsilon_i,\ i = 1, \cdots, n,$$

where $\mathbf{x}_i$s are simulated independently from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{C} = (0.5^{|j-k|})_{p\times p}$. All simulations are repeated for 100 times.

Apparently, the true mean shift model is a misspecified model for our weighted regression model setting in (II.1). However, we will demonstrate that the advantage of the PAWLS are still obvious compared with other methods from simulation studies.

Example II.1. (Low-dimensional case) We choose $n = 50$, $p = 8$, and set $\boldsymbol{\beta}^* = (3, 2, 1.5, 0, 0, 0, 0, 0)'$. The random error ϵ_i and the mean shift parameter γ_i are generated under the following four cases.

Case A: $\epsilon_i \sim N(0, 2^2)$, and $\gamma_i = 0$ for $i = 1, \cdots, n$;

Case B: ϵ_i follows a t distribution with degrees of freedom of 2, and $\gamma_i = 0$ for $i = 1, \cdots, n$;

Case C: similar to Case A, except that $\gamma_i = (-1)^{I(U_1<1/2)}(20 + 10U_2)$ for $1 \leq i \leq n/10$, where U_1 and U_2 are independent $U[0, 1]$.

Case D: similar to Case C, except that 10 is added on all x_{ij}s for $1 \leq i \leq n/10$ and $4 \leq j \leq 8$.

Case A includes only normal data; Case B includes heavy tails errors; Case C includes normal data with outliers in y direction; while Case D includes outliers in both x and y directions.

We compare the performance of the PAWLS with the adaptive Lasso in terms of both variable selection and outlier detection with the PLS with the adaptive Lasso (ALasso: [Zou06]) and several other sparse robust estimations including the SROS [XJ13], MMNNG [GV15], and sparse LTS (sLTS) [ACG$^+$13]. As a fair comparison, the adaptive Lasso penalty are used in all methods except for MMNNG where a nonnegative garrote method is used. The codes of both the MMNNG and sLTS are public available. The code of the SROS is provided by authors. The computation of the ALasso is the same as the PAWLS by fixing all $w_i = 1$.

If a model is correctly fitted, then $\{1 \le j \le p : \hat{\beta}_j \neq 0\} = \{1 \le j \le p : \beta^*_j \neq 0\}$; if a model is over-fitting, then $\{j : \hat{\beta}_j \neq 0\} \supset \{j : \beta^*_j \neq 0\}$. Both ratios of correctly fitting the model (CFR) and over-fitting the model (OFR) are computed. The average model size (AN: mean of $\#\{1 \le j \le p : \hat{\beta}_j \neq 0\}$) is also reported. All those results are summarized in Table II.1. Our simulation results also show that the PAWLS outperforms all other estimators in terms of variable selection in almost all cases. In particular, we have those findings. (1) The ALasso performs the best as expected when the data is normal in Case A; But the PAWLS is most comparable with the ALasso, compared with all other robust estimation. (2) When the data is heavy tailed in Case B, the ALasso behaves much worse than some of other sparse robust estimates. Among them, the PAWLS performs the best, while both the sLTS and SROS perform badly in this case. (3) When some normal data are contaminated in Case C, the ALasso loses its efficiency completely, while the PAWLS still performs quite well and beats all other robust methods. (4) When outliers exist in both x and y directions, the PAWLS also performs the best.

We also evaluate the coefficients estimation using the mean squared error (MSE), $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ out of all repetitions. Those results of MSE (after removing 10% of

largest ones) from Case A, C and D are plotted in Figure II.2. The boxplot under Case B shows the similar pattern as ones from C and D and is omitted here. It is observed that PAWLS has the best estimation efficiency by providing the smallest MSE results among all methods when the data are contaminated.

To evaluate the outlier detection performance, we compute the mean masking probability (M: fraction of undetected true outliers), the mean swamping probability (S: fraction of non-outliers labeled as outliers), and the joint outlier detection rate (JD: fraction of repetitions with 0 masking) out of all repetitions. The higher JD is, the better; the smaller M and S are, the better. Since the ALasso, MMNNG and SROS are not designed to specify outliers, we only report the outlier detection results from the PAWLS and sLTS in Table II.2. It is observed that the sLTS turns to produce a very large swamping probability in most cases. Compared with the sLTS, the PAWLS has a much better outlier detection performance.

In summary, the PAWLS is robust when the data is contaminated and does not lose much efficiency as other robust methods in normal case. Besides the PAWLS, the MMNNG performs the second best. However, compared with the PAWLS, the MMNNG is much more expensive in computation. In addition, MMNNG does not produce the outlier detection result.

Example II.2. (high-dimensional case) Similar to Example II.1, except that $n = 100$, $p = 500$ and $\boldsymbol{\beta} = (\mathbf{2}'_{10}, \mathbf{0}'_{p-10})'$, where $\mathbf{c}_k$ is a k-dimensional vector consists of all c.

Table II.1. Variable Selection Results for Example II.1 ($\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)'$)

Method	CFR (%)	OFR (%)	AN	CFR (%)	OFR (%)	AN
	Case A			**Case B**		
ALasso	88	12	3.14	80	6	2.95
sLTS	8	91	4.75	30	70	4.00
MMNNG	73	24	3.27	89	11	3.18
SROS	24	75	4.28	35	65	4.00
PAWLS	87	12	3.13	94	6	3.06
	Case C			**Case D**		
ALasso	2	1	1.59	0	19	2.49
sLTS	8	92	5.02	7	93	4.97
MMNNG	85	8	3.06	61	21	3.42
SROS	51	41	3.52	12	75	4.88
PAWLS	81	15	3.13	70	15	3.20

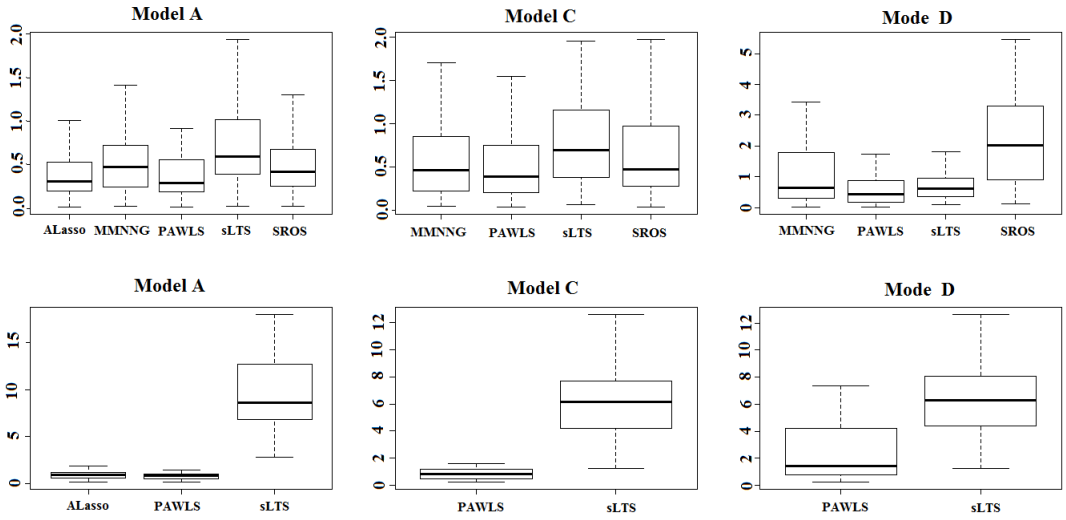
Table II.2. Outlier Detection Evaluation in Example II.1 and II.2

		sLTS			PAWLS		
	Model	M (%)	S (%)	JD(%)	M (%)	S (%)	JD(%)
Example II.1	Case A	0	5.30	100	0	1.22	100
	Case B	0	9.92	100	0	4.22	100
	Case C	0	1.87	100	0	0.67	100
	Case D	0.4	1.89	99	0	0.44	100
Example II.2	Case A	0	20.8	100	0	0.07	100
	Case B	0	18.5	100	0	1.15	100
	Case C	0	12.9	100	0.8	0.18	98
	Case D	0.1	13.0	99	27.8	0.08	100

In this example, we can only compare the PAWLS with the sLTS and ALasso since all other methods are only designed for $p < n$. We tried to implement their approaches in high-dimension where $p > n$, but failed.

All variable selection results are reported in Table II.3. Besides OFR, CFR and AN reported in Example II.1, we also report the OFR+2, the ratio of correct-fitted model and over-fitted model with at most two extra variables. Outlier detection results are reported in Table II.2. Some of MSE results are reported in those Boxplots in Figure II.2.

Figure II.2. Boxplot of MSE in Example II.1. The first row: Example II.1 (Case A, C and D from the left to right); The second row: Example II.2 (Case A, C and D from the left to the right). ALasso results are omitted in Case C and D since the MSE values are very large compared with others in those cases.



It is observed that the advantages of the PAWLS are even more obvious in high-dimensional settings, regarding variable selection, outlier detection and robust estimation. The PAWLS produces much higher CFR and CFR+2 than both the ALasso and the sLTS in contaminated cases. In this setting, sLTS turns to generate over-fitted model in most cases. When the data is normal, the PAWLS still works very well by producing high CFR value.

II.5.2. Real Data Applications

Two datasets will be studied in this section: Air pollution data ($p < n$) and NTC-60 data ($p > n$).

Table II.3. Variable Selection Results for Example II.2($\boldsymbol{\beta}' = (\mathbf{2}'_{10}, \mathbf{0}'_{p-10})$)

Method	CFR (%)	CFR+2 (%)	OFR (%)	AN (%)	CFR (%)	CFR+2 (%)	OFR (%)	AN
	Case A				**Case B**			
ALasso	55	90	45	11.0	48	74	45	13.1
sLTS	0	0	74	32.6	0	0	91	28.3
PAWLS	92	100	8	10.1	96	98	2	10.0
	Case C				**Case D**			
ALasso	0	0	5	40.2	0	0	3	39.0
sLTS	0	0	93	32.3	0	0	92	31.9
PAWLS	84	97	13	10.0	44	71	43	11.1

II.5.2.1. Air pollution

The air pollution data include information on the social and economic conditions in these areas. Their climates and some indices of air pollution potentials are available at `http://lib.stat.cmu.edu/DASL/Datafiles/SMSA.html`. The study is to investigate how the age-adjusted mortality is affected by all 14 covariates including mean January temperature (JanTemp: in degrees Fahrenheit), mean July temperature (JulyTemp: in degrees Fahrenheit), relative humidity (RelHum), annual rainfall (Rain: in inches), median education (Education), population density (PopDensity), percentage of non-whites (NonWhite), percentage of white collar workers (X.WC), population (Population), population per household (PopHouse), median income (Income), hydrocarbon pollution potential (HCPot), nitrous oxide pollution potential (NOxPot) and sulfur dioxide pollution potential (SO2Pot). Observation 21 had to be removed since it contains two missing values, resulting in $n = 59$ and $p = 14$ in our study. [GV15] analyzed the data with a QQ-plot and reveals the possible contamination of the data set. Therefore a robust regression method is needed for the air pollution data.

We consider the logarithm transformation on the pollution variables, due to their skewness. In addition, both the covariates and response variables are scaled to

have median value of zero and MAD (median absolute deviation from the median) value of one. This procedure keeps all variables within a comparable range level.

[GV15] analyzed the data with a QQ-plot and reveals the possible contamination of the data set. The PAWLS estimates of $\boldsymbol{\beta}$ are compared with output from four other methods in Table II.4. The PAWLS selects 7 variables from 14 of them. Among them, Rain, PopDensity, NonWhite, and SO2Pot are positively correlated with the log-value of the mortality rate, and JanTemp, Education, and HCPot have the negative effect. It is worthwhile to point it out that JanTemp is selected by all four robust methods, but not by ALasso. For this data, the PAWLS produces similar results as ones from MMNNG and SROS. However, the last two does not produce outlier detection results. This comparison is also consistent with the simulation studies, where MMNNG performs the second best after the PAWLS.

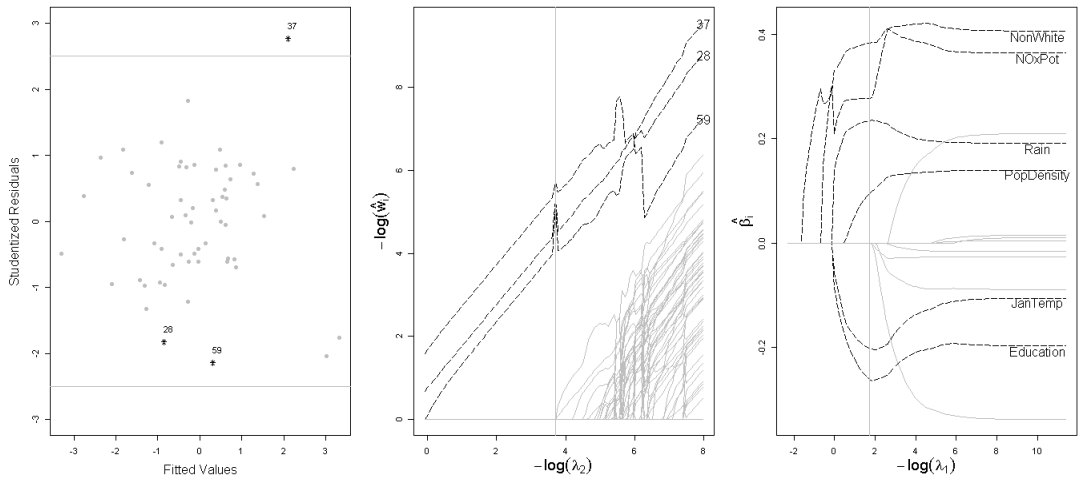
The outlier detection results from the PAWLS are reported in Figure II.3, where three suspected outliers detected by the PAWLS are highlighted by "*". See the studentized residual plot in the left panel Figure II.3. These three potential outliers are observation 28 from Lancaster, PA, observation 37 from New Orleans, LA, and observation 59 from York, PA. It is observed that the last two observations are masked using studentized residuals with cutoff value 2.5.

We also plot the solution paths of $\hat{\beta}_j$s along a sequence of λ_{1n}. See the right panel in Figure II.3. The solution paths of $\hat{w}_i$s along a sequence of λ_{2n} is also plotted in middle panel. Instead of being removed from the regression analysis completely, those two potential outliers are still used, but with some $\hat{w}_i$ value being much smaller than 1, for the final coefficients estimation and variable selection. In this data, the estimated weights for observations 27, 36 and 58 are 0.071, 0.029, and 0.050, respectively.

Table II.4. Estimation Regression Coefficients from Air Pollution Dataset

Variable	PAWLS	ALasso	sLTS	MMNNG	SROS
JanTemp	-0.097	0	-0.015	-0.051	-0.213
JulyTemp	0	0	0	0	0
RelHum	0	0	0	0	0
Rain	0.156	0	0.277	0.149	0.253
Education	-0.213	-0.320	-0.113	0	-0.224
PopDensity	0.098	0	0.169	0	0.097
NonWhite	0.379	0.479	0.282	0.398	0.389
X.WC	0	0	-0.062	-0.137	0
Population	0	0	-0.005	0	0
PopHouse	0	0	0.025	0	0
Income	0	0	-0.017	0	0
HCPot	-0.054	0	0	-0.108	0
NOxPot	0	0	0	0	0.253
SO2Pot	0.299	0.214	0.206	0.433	0.032

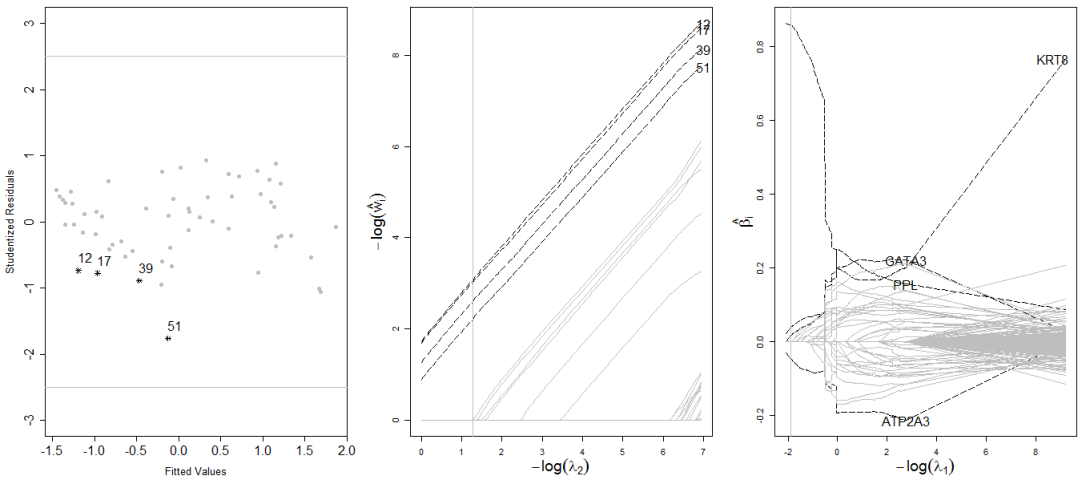
Figure II.3. Air Pollution Data Analysis. Left Panel: Studentized residuals plot (normal observations and detected outliers are highlighted by grey '·' and dark '*', separately); Middle Panel: Solution paths of $\hat{w}_i$ (curves of detected outliers (normal) observations are plotted using the dark (grey) color, the grey vertical line gives the location of the optimal λ_{2n}); Right panel: Solution paths of $\hat{\beta}_j$ (curves of selected (non-selected) variables, the grey vertical line gives the location of the optimal λ_{1n}).



II.5.2.2. NCI-60 cancer cell panel

As to the NCI-60 dataset introduced in I.1 , before the robust analysis, we perform some pre-screening and kept only p_1 genes with largest variations and then choose p_2 out of them which are most correlated with the response variable. Here the final dataset is obtained by choosing $p_1 = 2000$ and $p_2 = 500$, yielding $n = 59$ and $p = 500$. After applying the PAWLS, we select 10 genes: KRT8 (0.858), PPL(0.017), GATA3 (0.040), and ATP2A3 (-0.046), where the value in each parenthesis is the corresponding coefficient estimation. As a comparison, we also apply both the sLTS and ALasso to analyze this data, where the former selects 27 genes including KRT8 and GATA3, and the latter selects only KRT8.

Figure II.4. NCI-60 Data Analysis. Left Panel: Studentized residuals plot (normal observations and detected outliers are highlighted by grey '·' and dark '*', separately); Middle Panel: Solution paths of $\hat{w}_i$ (curves of detected outliers (normal) observations are plotted using the dark (grey) color, the grey vertical line gives the location of the optimal λ_{2n}); Right panel: Solution paths of $\hat{\beta}_j$ (curves of selected (non-selected) variables, the grey vertical line gives the location of the optimal λ_{1n}).



In addition, the PAWLS also identifies 4 outliers out of 59 samples: observations 12 (0.049), 17 (0.050), 39 (0.076), and 51 (0.112), with corresponding weight estimation given in each parenthesis. Those potential outliers are also highlighted in the studentized residuals plot in the left panel in Figure II.4. Here the studentized residuals is generated from post (Lasso) selection least squares regression. Both solution paths for all w_is and β_js are plotted in the middle and right panels, respectively. It is observed that those the weight solution paths of those potential outliers are obviously separated from ones from other observations.

The analyses are repeated for both $p_1 = 5000$, $p_2 = 1000$ and $p_1 = 3000$, $p_2 = 800$, yielding the similar results as above.

CHAPTER III

PENALIZED ROBUST APPROXIMATED QUADRATIC M-ESTIMATORS

III.1. Introduction

Asymmetry along with heteroscedasticity or contamination often occurs with the growth of data dimensionality. In high-dimensional settings, particularly when random errors follow irregular distributions such as asymmetry and heteroscedasticity, simultaneous mean estimation and variable selection are still of interest in many applications. In this chapter, we are interested in high-dimensional mean regression that is robust to the following irregular settings: (a) the data are not symmetric due to the skewness of random errors ([FLW17]); (b) the data are heteroscedastic ([DCL12], [WWL12]); and (c) the data are contaminated in both response and a large number of variables ([RL05]). However, above irregular settings are often overlooked for high-dimensional data analysis, especially for the theoretical development.

Despite the extensive work on penalized robust M-estimator in high-dimensional regression (e.g. [H$^+$64], [LLZ$^+$11], [GH10], [Wan13], [Loh17]), most of them either do not estimate the conditional mean regression function or require the error distribution to be symmetric and/or homogeneous. To tackle this problem, [FLW17] proposed a so-called RA-Lasso estimator, in which they waived the symmetry requirement by using the Huber loss with a diverging parameter in order to reduce the bias when the error distribution is asymmetric. [FLW17] obtained nice asymptotic properties of the RA-Lasso estimator, and proved its estimation consistency at the minimax rate enjoyed by LS-Lasso.

However, the Huber loss approximation used in the RA-Lasso dose not down-weight the very large residual due to its non-decreasing Ψ-function. [SMS08] showed that M-estimators given by non-decreasing Ψ-function do not possess finite variance sensitivity, meaning the asymptotic variance can be largely affected if the assume model is only approximately true. In that paper, the authors proposed to consider re-descending M-estimators with Ψ-function redescending to zero to address this problem. They further showed that redescending M-estimator can be designed by maximizing the minimum variance sensitivity under a global minimax criterion. For instance, the Smith's estimator and Tukey's biweight estimator are the optimal M-estimator with minimax variance sensitivity for a class of densities with a bounded variance and a bounded fourth moment, respectively [SMS08]. Therefore it is tempting to also include redescending M-estimator in the study of complex high-dimensional settings.

For decades both the theoretical and computational result in penalized re-descending M-estimator in high-dimensional settings have been very limited, due to the non-convexity of loss functions. Recently [Loh17] established a form of local statistical consistency for the high-dimensional M-estimators allowing both the loss and penalty functions to be non-convex. However, this study does not address the problem of asymmetry and heteroscedasticity. Also, their numerical studies neglect settings for asymmetric data and lack of comparisons among different M-estimations.

In this chapter, we consider high-dimensional linear regression in more general irregular settings: the data can be contaminated or include possible large outliers in both random errors and covariates, the random errors may lack of symmetry and homogeneity. In particular, we investigate both statistical and computational properties of high-dimensional mean regression in the penalized M-estimator framework with diverging robustness parameters.

Related Works: we end this section by highlighting a few things on how our work is different from some recent related work:

(1) As introduced earlier, the RA-Lasso proposed by [FLW17] waives the symmetry requirement by allowing the parameter of Huber loss to diverge. The idea is that by controlling the divergent rate of the parameter, while preserving certain robustness, the Huber loss becomes 'closer' to the ℓ_2 loss and thus potentially reduces the bias when the error distribution is asymmetric. Our work in this chapter relax the convexity restriction of loss functions and answer the question on how in general a loss function with strong robustness should converge to the ℓ_2 loss to achieve the estimation consistency at the minimax rate. While [FLW17] focuses exclusively on the Lasso penalty, our framework also allows concave penalties and therefore inherits certain oracle property under some conditions. Furthermore, we relax the sub-Gaussian assumption on covariates in [FLW17] by incorporating weight functions in the extension of PRAM estimators.

(2) [Loh17] also establishes a form of local statistical consistency for high-dimensional non-convex M-estimators. However, we address the problem of asymmetry and heteroscedasticity. In particular, our proposed framework is more general: we consider the empirical loss function $\mathcal{L}_{\alpha,n}$ satisfying $\lim_{\alpha\to\infty} E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$, where $\boldsymbol{\beta}^*$ is the true parameter vector and α is the diverging parameter. In contrast, [Loh17] requires the condition $E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$ for each $\alpha > 0$, which may not hold with the lack of homogeneity and symmetry in general. Additionally, [Loh17] does not suggest which estimators to be considered in real applications. We further investigate this problem by comparing different PRAM estimators in numerical studies.

The remainder of this chapter is organized as follow. In Section III.2, we introduce the basic setup regarding PRAM estimators and corresponding generalizations. In Section III.3, we establish the local estimation consistency for the PRAM estimators under sufficient conditions. For non-convex regularized PRAM estimators, we also present our statistical theory concerning the selection consistency and the asymptotic normality of PRAM estimators. We discuss the implementation of PRAM estimators including both the computational algorithm and the tuning parameter selection in Section III.4. In section III.5, we conduct some simulation studies to demonstrate the performance of the PRAM estimators under different settings. We also apply those PRAM estimators for NCI-60 data analysis and illustrates all results in Section III.6. All technical proofs are relegated to the Appendix.

Notation: We use bold symbols to denote matrices or vectors. For a matrix or a vector $\boldsymbol{\nu}$, we write $\boldsymbol{\nu}^T$ to denote its transpose. We write $\|\cdot\|_1$ and $\|\cdot\|_2$ to denote the L_1 norm and the L_2 norm of a vector, respectively. For a function $g : \mathbb{R}^p \mapsto \mathbb{R}$, we write ∇g to denote a gradient of the function. We write u_+ to denote $\max(u, 0)$ for any $u \in \mathbb{R}$.

III.2. The PRAM Method

III.2.1. Model Settings

Consider an ultra high-dimensional linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i, \tag{III.1}$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ for $i = 1, \cdots, n$ are independent and identically distributed (i.i.d) p-dimensional covariate vectors such that $E(\mathbf{x}_i) = \mathbf{0}$, $\{\epsilon_i\}_{i=1}^n$ are independent errors such that $E(\epsilon_i \mid \mathbf{x}_i) = 0$ and thus we allow the conditional heteroscedasticity.

Note $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)^T \in \mathbb{R}^p$ is an s-sparse conditional mean coefficient vector (only include s nonzero elements) and $p \gg n$.

Our model settings permit the existence of all the following irregular settings on both ϵ_is and $\mathbf{x}_i$s: (a) asymmetry of ϵ_i; (b) heteroscedasty of ϵ_i and ϵ_i may depend on $\mathbf{x}_i$; (c) data contamination of ϵ_i and $\mathbf{x}_i$.

We are interested in penalized mean regression estimators such that

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{III.2}$$

where $\mathcal{L}_{\alpha,n}$ is the empirical loss function and ρ_λ is a penalty function which encourages the sparsity in the solution. Here $\alpha > 0$ is a parameter controlling the robustness, which is allowed to diverge. As mentioned in Section III.1, we consider the loss function $\mathcal{L}_{\alpha,n}$ satisfying

$$\lim_{\alpha \to \infty} E[\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}. \tag{III.3}$$

This condition in (III.3) relaxes the condition, $E[\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$ for each $\alpha > 0$, required in [Loh17], which may be invalid with the lack of homogeneity and symmetry. The condition (III.3) permits the random error to be heterogeneous and/or asymmetric, as long as $E[\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)]$ converges to $\mathbf{0}$ with diverging α.

We also include the side condition $\|\boldsymbol{\beta}\|_1 \leq R$ in the penalized optimization problem in (III.2), in order to guarantee the existence of local/global optima, for the case where the loss function or the regularizer may be non-convex. We also require $\|\boldsymbol{\beta}^*\|_1 \leq R$ so that $\boldsymbol{\beta}^*$ is feasible in (III.2). In real applications, we can choose R to be a sufficiently large number.

III.2.2. Penalty Functions

Since the coefficients vector $\boldsymbol{\beta}^*$ is assumed to be s-sparse in the high-dimensional linear regression model in (III.1), we only consider penalties which generate sparse solutions. In particular, we require the penalty function ρ_λ in (III.2) to satisfy following properties listed in Assumption III.1.

Assumption III.1 (Penalty Function Assumptions)**.** *The penalty function is coordinate-separable such that* $\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho_\lambda(\beta_j)$ *for some scalar function* $\rho_\lambda : \mathbb{R} \mapsto \mathbb{R}$*. In addition,*

(i) the function $t \mapsto \rho_\lambda(t)$ *is symmetric around zero and* $\rho_\lambda(0) = 0$*;*

(ii) the function $t \mapsto \rho_\lambda(t)$ *is non-decreasing on* $\mathbb{R}^+$*;*

(iii) the function $t \mapsto \frac{\rho_\lambda(t)}{t}$ *is non-increasing on* $\mathbb{R}^+$*;*

(iv) the function $t \mapsto \rho_\lambda(t)$ *is differentiable for* $t \neq 0$*;*

(v) $\lim_{t \to 0^+} \rho'_\lambda(t) = \lambda$*;*

(vi) there exists $\mu > 0$ *such that the function* $t \mapsto \rho_\lambda(t) + \frac{\mu}{2} t^2$ *is convex;*

(vii) there exists $\delta \in (0, \infty)$ *such that* $\rho'_\lambda(t) = 0$ *for all* $t \geq \delta\lambda$*.*

Those properties in Assumption III.1 are related to the penalty functions studied in [LW13] and [Loh17], where ρ_λ is said to be μ-amenable if ρ_λ satisfies conditions (i)-(vi) for μ defined in (vi). If ρ_λ also satisfies condition (vii), we say that ρ_λ is (μ, δ)-amenable. Some popular choices of amenable penalty functions include Lasso [Tib96b], SCAD [FL01], and MCP [Z$^+$10] given as follows:

- The **Lasso** penalty, $\rho_\lambda(t) = \lambda|t|$, is 0-amenable but not $(0, \delta)$-amenable for any $\delta < \infty$.

- The **SCAD** penalty,

$$\rho_\lambda(t) = \begin{cases} \lambda|t| & for\ |t| \leq \lambda, \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & for\ \lambda < |t| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & for\ |t| > a\lambda, \end{cases}$$

  where $a > 2$ is a fixed parameter. The SCAD penalty is also (μ, δ)-amenable with $\mu = \frac{1}{a-1}$ and $\delta = a$.

- The **MCP** penalty,

$$\rho_\lambda(t) = sign(t)\lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

  where $b > 0$ is a fixed parameter. The MCP penalty is also (μ, δ)-amenable with $\mu = \frac{1}{b}$ and $\delta = b$.

It has been shown that the folded concave penalty, such as SCAD or MCP, possesses better variable selection properties than the convex penalty like the Lasso.

III.2.3. Loss Functions

From the linear model setting in Section III.2.1, we know $E(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}^*$. We are interested in finding a well-behaved mean-regression estimator of $\boldsymbol{\beta}^*$. Since we consider a general setting discussed in Section III.2.1, we wish to study the empirical loss function $\mathcal{L}_{\alpha,n}$ that are robust to outliers and/or heavy-tailed distribution. Let

$l_\alpha : \mathbb{R} \mapsto \mathbb{R}$ denote a residual function, or a loss function, defined on each observation pair $(\mathbf{x}_i, y_i)$. The corresponding empirical loss function for (III.2) is then given by

$$\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} l_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}). \tag{III.4}$$

With a well chosen non-quadratic function l_α, the penalized mean regression estimators from (III.2) can be robust to outliers or heavy-tailed distribution in the additive noise term ϵ_i. However, it may generate bias to the conditional mean when the conditional distribution of ϵ_i is not symmetric.

To reduce such bias induced by the non-quadratic loss, we consider a family of loss function with flexible robustness and diverging parameters satisfying (III.3) to approximate the traditional quadratic loss. In particular, we require the following approximation:

$$\textbf{Approximation Equation: } \lim_{\alpha\to\infty} l_\alpha(u) = \frac{1}{2}u^2, \quad \forall u \in \mathbb{R}. \tag{III.5}$$

The empirical loss function satisfy (III.5) is called a robust approximated quadratic loss function. The following approximations take the Huber loss, Tukey's biweight loss and Cauchy loss to robustly approximate the quadratic loss functions:

- **Huber Approximation**

$$l_\alpha(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \le \alpha, \\ \alpha|u| - \frac{\alpha^2}{2} & \text{if } |u| \ge \alpha. \end{cases}$$

- **Tukey's biweight Approximation**

$$l_\alpha(u) = \begin{cases} \frac{\alpha^2}{6}(1 - (1 - \frac{u^2}{\alpha^2})^3) & \text{if } |u| \le \alpha, \\ \frac{\alpha^2}{6} & \text{if } |u| \ge \alpha. \end{cases}$$

- **Cauchy Approximation**

$$l_\alpha(u) = \frac{\alpha^2}{2}\log(1+\frac{u^2}{\alpha^2}).$$

It is straight forward to verify that all above three loss functions satisfy equation (III.5). In addition, the Tukey's biweight loss and Cauchy loss produce redescending M-estimators. In the robust regression literature, we call an M-estimator redescending if there exists $u_0 > 0$ such that $|l'_\alpha(u)| = 0$ or decrease to 0 smoothly, for all $|u| \geq u_0$. In that case, large residuals can be downweighted. See more discussions in [Mul04] and [SMS08].

III.2.4. PRAM Estimators and the Extensions

A class of PRAM estimators takes the form:

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{III.6}$$

where the penalty function ρ_λ satisfies Assumption III.1, the loss function l_α is a scalar function satisfying equation (III.5) and $\alpha > 0$ is a robustness parameter which is allowed to diverge.

Whereas a PRAM estimator in equation (III.6) takes into account the contamination or heavy-tailed distribution in asymmetric additive error, a single outlier in $\mathbf{x}_i$ may still cause the corresponding estimator to perform arbitrarily badly. We downweight large values of $\mathbf{x}_i$ and extend the class of PRAM estimators to

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)} l_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta})v(\mathbf{x}_i)) + \rho_\lambda(\boldsymbol{\beta}) \right\}, \tag{III.7}$$

where w, v are weight functions mapping from $\mathbb{R}^p$ to $\mathbb{R}^+$. When $w \equiv v \equiv 1$, (III.7) is reduced to the PRAM class defined in (III.6). A few options for choosing the weight

functions can be found in [Mal75], [Hil77], [MS71]. Such a downweighting strategy was also adopted in [Loh17].

For the rest of the chapter, we specify the PRAM estimator with the Huber approximation, Tukey's biweight approximation and Cauchy approximation as the HA-type, TA-type and CA-type PRAM estimator, respectively. In particular, we also specify a PRAM estimator using a redescending loss function approximation (e.g. Tukey's biweight approximation and Cauchy approximation) a redescending PRAM estimator. Additionally, we classify a PRAM estimator with the Lasso penalty and MCP penalty as the Lasso-type and MCP-type PRAM estimator correspondingly.

III.3. Statistical Properties

III.3.1. Estimation Consistency

As in (III.7), we consider a class of PRAM estimators with the loss function in a general setting,

$$\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)}l_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta})v(\mathbf{x}_i)). \tag{III.8}$$

To obtain the estimation consistency, we make the following additional assumptions on l_α.

Assumption III.2 (Loss Function Assumptions)**.** *$l_\alpha : \mathbb{R} \mapsto \mathbb{R}$ is a scalar function for $\alpha > 0$ with the existence of the first derivative l'_α everywhere and the second derivative l''_α almost everywhere. In addition,*

(i) there exists a constant $0 < k_1 < \infty$ such that $|l'_\alpha(u)| \leq k_1\alpha$ for all $u \in \mathbb{R}$;

(ii) for all $\alpha > 0$, $l'_\alpha(0) = 0$ and l'_α is Lipschitz such that $|l'_\alpha(x) - l'_\alpha(y)| \leq k_2|x - y|$ for all $x, y \in \mathbb{R}$ and some $0 < k_2 < \infty$;

(iii) for some $k \geq 2$, there exists a constant $d_1 > 0$ such that $|1 - l''_\alpha(u)| \leq d_1|u|^k\alpha^{-k}$ for almost all $|u| \leq \alpha$.

Note that Assumption III.2(i) indicates that the magnitude of l'_α is bounded from above at the same rate of α so that the PRAM estimator can achieve robustness. Assumption III.2(ii) implies $|l'_\alpha(u)| \leq k_2|u|$ for all $u \in \mathbb{R}$ and $|l''_\alpha(u)| \leq k_2$ for almost every $u \in \mathbb{R}$. In particular, the loss functions we study in this chapter actually satisfy Assumption III.2(ii) with $k_2 = 1$, showing that l_α is bounded by the quadratic loss function $u^2/2$ for any α. Assumption III.2(iii) indicates that for almost all $u \in \mathbb{R}$, l''_α converges point-wisely to 1 with at least the order of α^{-k} for $k \geq 2$.

The above assumptions cover a wide range of loss functions, including the Huber loss, Hampel loss, Tukey's biweight loss and Cauchy loss.

Remark. By some simple math, we can show that $\lim_{\alpha\to\infty} l'_\alpha(u) = u$ for all $u \in \mathbb{R}$ based on Assumption III.2. Suppose in addition that $l_\alpha(0) = 0$, we can further obtain the approximation equation (III.5), indicating that Assumption 2 alone gives sufficient conditions for l_α to approximate the quadratic loss.

Remark. By dominated convergence theorem, we have

$$\begin{aligned}\lim_{\alpha\to\infty} E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] &= \lim_{\alpha\to\infty} E[w(\mathbf{x}_i)\mathbf{x}_i l'_\alpha(\epsilon_i v(\mathbf{x}_i))] \\ &= E[w(\mathbf{x}_i)\mathbf{x}_i(\epsilon_i v(\mathbf{x}_i))] = E[w(\mathbf{x}_i)\mathbf{x}_i E(\epsilon_i \mid \mathbf{x}_i)v(\mathbf{x}_i))] = \mathbf{0}.\end{aligned}$$

So under Assumption III.2, we have $\lim_{\alpha\to\infty} E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$ and thus it allows the random error to be heterogeneous and/or asymmetric.

We now make some weak assumptions on both random error ϵ and covariate vector $\mathbf{x}$ for the investigation of the approximation error.

Assumption III.3 (Error and Covariate Assumptions)**.** *For $w(\mathbf{x})$ and $v(\mathbf{x})$ given in* (III.7)*, the random error ϵ with $E[\epsilon \mid \mathbf{x}] = 0$ and covariate vector $\mathbf{x}$ with $E[\mathbf{x}] = \mathbf{0}$ satisfy:*

(i) $E[E(|\epsilon|^k \mid \mathbf{x})v(\mathbf{x})^k]^2 \leq M_k < \infty$*, for $k \geq 2$ in Assumption 2(iii);*

(ii) $\sup_{\|u\|_2=1} E[v(\mathbf{x})\mathbf{x}^T u]^{2k} = q_k < \infty$*, for $k \geq 2$ in Assumption 2(iii);*

(iii) $0 < k_l < \lambda_{min}(E[w(\mathbf{x})v(\mathbf{x})\mathbf{x}\mathbf{x}^T])$ *and* $\lambda_{max}(E[w(\mathbf{x})^2\mathbf{x}\mathbf{x}^T]) < k_u$*;*

(iv) for any $\boldsymbol{\nu} \in \mathbb{R}^p$, $w(\mathbf{x})\mathbf{x}^T\boldsymbol{\nu}$ is sub-Gaussian with parameter at most $k_0^2\|\boldsymbol{\nu}\|_2^2$.

Note that condition (i) requires only the existence of second conditional moment of ϵ, indicating that this condition is independent of the distribution of ϵ itself and can hold for heavy-tailed or skewed distribution. If $w(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$, the conditions (ii) and (iv) hold when $\mathbf{x}_i^T\boldsymbol{\nu}$ is sub-Gaussian for any $\boldsymbol{\nu} \in \mathbb{R}^p$. In this case, Assumption III.3 becomes conditions (C1-C3) in [FLW17]. If covariate $\mathbf{x}$ is contaminated or heavy-tailed distributed, conditions (ii)-(iv) nonetheless holds with some proper choices of $w(\mathbf{x})$ and $v(\mathbf{x})$ (e.g. $w(\mathbf{x})\mathbf{x}^T\boldsymbol{\nu}$ is bounded for any $\boldsymbol{\nu} \in \mathbb{R}^p$), which potentially relaxes the sub-Gaussian assumption on $\mathbf{x}$.

Let $\boldsymbol{\beta}_\alpha^*$ be a local non-penalized population minimizer under the PRAM loss,

$$\boldsymbol{\beta}_\alpha^* \in \underset{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq R_0}{\operatorname{argmin}} \left\{ E\left[\frac{w(\mathbf{x})}{v(\mathbf{x})} l_\alpha((y - \mathbf{x}^T\boldsymbol{\beta})v(\mathbf{x})) \right] \right\}, \tag{III.9}$$

for some $0 < R_0 < \infty$. Note that $\boldsymbol{\beta}_\alpha^*$ is a local minimizer of (III.9) within a neighborhood of $\boldsymbol{\beta}^*$. If the regularization parameter λ in equation (III.7) converges to 0 sufficiently fast, then $\hat{\boldsymbol{\beta}}$ is a natural unpenalized M-estimator of $\boldsymbol{\beta}_\alpha^*$ for any $\alpha > 0$. Whereas $\boldsymbol{\beta}_\alpha^*$ differs from $\boldsymbol{\beta}^*$ in general, $\boldsymbol{\beta}_\alpha^*$ is expected to converge to $\boldsymbol{\beta}^*$ when $\alpha \to \infty$,

due to the approximation equation (III.5) for PRAM. The rate of the approximation error $\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2$ is established in Theorem 1.

Theorem III.1. *Under the Assumption III.2 and III.3, there exists a universal positive constant C_1, such that $\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2 \leq 2^k C_1 k_l^{-1}\sqrt{k_u}(\sqrt{M_k} + R_0^k\sqrt{q_k})\alpha^{1-k}$. Here k, k_l, k_u, M_k, q_k appear in Assumption III.2, III.3 and R_0 appears in (III.9).*

Theorem III.1 gives an upper bound of the approximation error between the true parameter vector and the non-penalized PRAM population minimizer. The approximation error vanishes when $\alpha \to \infty$. It vanishes faster if a higher moment of $\epsilon|\mathbf{x}$ exists. In fact, Theorem 1 demonstrates that the approximation of the loss function l_α to the quadratic loss helps to reduce the bias induced by the asymmetry on ϵ. If we let l_α in equation (III.8) be the Huber loss and $w(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$, Theorem 1 gives the upper bound of the approximation error studied in [FLW17].

In order to obtain the estimation consistency for the PRAM estimator in (III.7), we also require the loss function $\mathcal{L}_{\alpha,n}$ to satisfy the following uniform Restricted Strong Convexity (RSC) condition.

Assumption III.4 (Uniform RSC condition)**.** *There exist γ, τ, $\alpha_0 > 0$ and a radius $r > 0$ such that for all $\alpha \geq \alpha_0$, the loss function $\mathcal{L}_{\alpha,n}$ in* (III.7) *satisfies*

$$\langle \nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_1) - \nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle \geq \gamma\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 - \tau\frac{\log p}{n}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1^2, \qquad \text{(III.10)}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ such that $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*\|_2 \leq r$ for $j = 1, 2$.

Note that the uniform RSC assumption is only imposed on $\mathcal{L}_{\alpha,n}$ inside the ball of radius r centered at $\boldsymbol{\beta}^*$. Thus the loss function used for robust regression can be wildly nonconvex while it is away from the origin. The radius r essentially specifies a

local ball centered around $\boldsymbol{\beta}^*$ in which stationary points of the PRAM estimator are well-behaved.

Remark. In [LW13] and [Loh17], the RSC condition were imposed on a specific loss function. Although Assumption III.4 requires that the RSC condition is satisfied uniformly over a family of loss functions generated from a range of α, this assumption is in fact not stronger: Assumption III.4 holds naturally if there exists $\alpha_0 > 0$ such that $\mathcal{L}_{\alpha_0,n}$ satisfies Assumption 2 and inequality (III.10) for some $\gamma, \tau > 0$. We further establish the uniform RSC condition in Appendix.

We present our main estimation consistency result on the PRAM estimator in the following Theorem III.2.

Theorem III.2. *Suppose the random error and covariates satisfy Assumption III.3 and $\mathcal{L}_{\alpha,n}$ in* (III.7) *satisfies Assumption III.2. Then we have the following results.*

(i) If $\max\{(\frac{2d}{R_0})^{\frac{1}{k-1}}, C_2(\frac{n}{\log p})^{\frac{1}{2(k-1)}}\} \leq \alpha \leq C_3\sqrt{\frac{n}{\log p}}$, *then with probability greater than* $1 - 2\exp(-C_4 \log p)$, $\mathcal{L}_{\alpha,n}$ *satisfies*

$$\|\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)\|_\infty \leq C_5\sqrt{\frac{\log p}{n}}. \tag{III.11}$$

(ii) Suppose $\mathcal{L}_{\alpha,n}$ also meets the uniform RSC condition in Assumption III.4. Suppose ρ_λ is μ-amenable with $\frac{3}{4}\mu < \gamma$ in Assumption III.1. Let $\hat{\boldsymbol{\beta}}$ be a local PRAM estimator in the uniform RSC region. Then for $R \geq \|\boldsymbol{\beta}^*\|_1$, $\lambda \geq \max\{4\|\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)\|_\infty, 8\tau R\frac{\log p}{n}\}$ and $n \geq C_0 r^{-2} k \log p$, $\hat{\boldsymbol{\beta}}$ exists and satisfies the bounds

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{24\lambda\sqrt{s}}{4\gamma - 3\mu} \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{96\lambda s}{4\gamma - 3\mu}.$$

The statistical consistency result of Theorem III.2 holds even when the random errors lack of symmetry and homogeneity, and the regressors lack of sub-Gaussian

assumption. It shows that with high probability one can choose $\lambda = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{s \log p}{n}}\right)$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_p\left(\sqrt{\frac{s^2 \log p}{n}}\right)$. Hence, it guarantees that when the parameter α diverges at a certain rate, a local PRAM estimator within the local region of radius r is statistically consistent at the minimax rate enjoyed by the LS-Lasso. The rate range of α stated in Theorem III.2 (i) in fact reveals that in the presence of asymmetric and heavy-tailed/contaminated data, α should diverge faster enough, for example, faster than $\mathcal{O}\left((\frac{n}{\log p})^{\frac{1}{2(k-1)}}\right)$, to reduce the bias sufficiently but meanwhile not too fast, for instance, slower than $\mathcal{O}\left((\frac{n}{\log p})^{\frac{1}{2}}\right)$, in order to preserve certain robustness of a PRAM estimator. The existence of a higher moment of $\epsilon|\mathbf{x}$ (a larger k) actually allows α to diverge at a lower rate.

Remark. The proof of Theorem III.2 in Appendix reveals that the estimation consistency result also holds for the local stationary points in program (III.2). Here $\tilde{\boldsymbol{\beta}}$ is a stationary point of the optimization in (III.2) if

$$\langle \nabla \mathcal{L}_{\alpha,n}(\tilde{\boldsymbol{\beta}}) + \nabla \rho_\lambda(\tilde{\boldsymbol{\beta}}), \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \rangle \geq 0,$$

for all feasible $\boldsymbol{\beta}$ in a neighbour of $\tilde{\boldsymbol{\beta}}$. Note that stationary points include both the interior local maxima as well as all local and global minima. Hence Theorem III.2 guarantees that all stationary points within the ball of radius r centered at $\boldsymbol{\beta}^*$ have local statistical consistency at the minimax rate enjoyed by the LS-Lasso.

III.3.2. Oracle Properties

In this section, we establish the oracle properties for the PRAM estimators in program (III.7). We first define the local oracle estimator as

$$\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^S : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r}{\operatorname{argmin}} \left\{ \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) \right\}, \tag{III.12}$$

where we set $S = \{j : \beta_j^* \neq 0\}$. Let $\beta_{\min}^* = \min_{j \in S} |\beta_j^*|$ denote a minimum signal strength on $\boldsymbol{\beta}^*$. Our oracle result shows that when the penalty ρ_λ is (μ, δ)-amenable and the assumptions stated earlier are satisfied, those stationary points of the PRAM estimator in program (III.7) within the local neighborhood of $\boldsymbol{\beta}^*$ are actually unique and agree with the oracle estimator (III.12), as stated in the following theorem.

Theorem III.3. *Suppose the penalty ρ_λ is (μ, δ)-amenable and conditions in Theorem III.2 hold. Suppose in addition that $v(\mathbf{x})\mathbf{x}_j$ is sub-Gaussian for all $j = 1, \cdots, p$, $\|\boldsymbol{\beta}^*\|_1 \leq \frac{R}{2}$ for some $R > \frac{192\lambda s}{4\gamma - 3\mu}$, $\beta_{\min}^* \geq C_6\sqrt{\frac{\log p}{ns}} + \delta\lambda$, and $n \geq C_{01} s \log p$ for a sufficiently large constant C_{01}. Suppose α satisfies $C_{22}\left(\frac{ns^2}{\log p}\right)^{\frac{1}{2(k-1)}} \leq \alpha \leq C_3\sqrt{\frac{n}{\log p}}$ and $s^2 = \mathcal{O}\left((\frac{n}{\log p})^{k-2}\right)$. Let $\tilde{\boldsymbol{\beta}}$ be a stationary point of program (III.7) in the uniform RSC region. Then with probability at least $1 - C_8 \exp(-C_{41}\frac{\log p}{s^2})$, $\tilde{\boldsymbol{\beta}}$ satisfies $supp(\tilde{\boldsymbol{\beta}}) \subseteq S$ and $\tilde{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$.*

Two most often considered (μ, δ)-amenable penalties are SCAD and MCP, as introduced in Section III.2.2. Since the Lasso penalty is not (μ, δ)-amenable, the Lasso-type PRAM estimator does not have the oracle properties. In Theorem III.3, the lower bound rate of α is higher than the one in Theorem III.2, with a ratio $\mathcal{O}\left(s^{\frac{1}{k-1}}\right)$. Thus to have the oracle properties, s cannot grow with n too fast. In particular, $s = \mathcal{O}\left((\frac{n}{\log p})^{\frac{k-2}{2}}\right)$ for $k \geq 2$. Note that the feasibility condition $\|\boldsymbol{\beta}^*\| \leq \frac{R}{2}$ instead of R in Theorem 2, is for the technical proof. It means that (III.7) is optimized in a larger neighborhood of $\boldsymbol{\beta}^*$ in order to cover $(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}, \mathbf{0}_{S^c})$ such that $\|\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 < r$.

Remark. The condition $s^2 = \mathcal{O}\left((\frac{n}{\log p})^{k-2}\right)$ shows that, if the number of non-zero parameters s is finite, Theorem III.3 requires only the existence of second moment of $\epsilon|\mathbf{x}$ $(k = 2)$; if we also allow s to grow with sample size n, the oracle result holds when at least the third moment of $\epsilon|\mathbf{x}$ exists $(k \geq 3)$.

Since $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ is essentially an s-dimensional M-estimator, to analyze the asymptotic behavior of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_S$, Theorem III.3 allows us to apply previous results in the literature concerning the asymptotic distribution of low-dimensional M-estimators. In particular, [HS00] established the asymptotic normality for a fairly general class of convex M-estimators where p is allowed to grow with n. Although the loss function we considered may be highly nonconvex, the restricted program in (III.12) can still be convex under the uniform RSC condition. Hence by applying our Theorem III.3 and the standard results for M-estimators with a diverging number of parameters in [HS00], we can obtain the following theorem concerning the asymptotic normality of any stationary point of the program (III.7). For the sake of simplicity, we only provide the result under $w(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$. The result of a weighted PRAM can be derived accordingly.

Theorem III.4. *Suppose conditions in Theorem III.3 hold and the loss function $\mathcal{L}_{\alpha,n}$ given in (III.8) is twice differentiable within the ℓ_2-ball of radius r around $\boldsymbol{\beta}^*$. Suppose for all $\alpha > 0$, l''_α is Lipschitz such that $|l''_\alpha(x) - l''_\alpha(y)| \leq k_3|x-y|$ for all $x, y \in \mathbb{R}$ and some $0 < k_3 < \infty$. Suppose in addition that $\alpha > (2C_9/k_l)^{1/k}$ and $\alpha^{1-k} = o(n^{-1/2})$. Let $\tilde{\boldsymbol{\beta}}$ be a stationary point of program (III.7) in the uniform RSC region. If $\frac{s \log^3 s}{n} \to 0$, then $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{s}{n}}\right)$. If $\frac{s^2 \log s}{n} \to 0$, then for any $\boldsymbol{\nu} \in \mathbb{R}^p$, we have*

$$\frac{\sqrt{n}}{\sigma_{\boldsymbol{\nu}}} \cdot \boldsymbol{\nu}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0,1),$$

where

$$\sigma_{\boldsymbol{\nu}}^2 = \boldsymbol{\nu}_S^T E[(\nabla^2 \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{SS}]^{-1} Var(l'_\alpha(\epsilon_i)(\mathbf{x}_i)_S) E[(\nabla^2 \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{SS}]^{-1} \boldsymbol{\nu}_S.$$

The condition $\alpha^{1-k} = o(n^{-1/2})$ indicates that α should diverge at least faster than $n^{\frac{1}{2(k-1)}}$, in addition to the rate stated in Theorem III.3. Together with the result

in Theorem III.1, it means that the approximation error $\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2$ should vanish at a rate of $o(n^{-1/2})$, in order to obtain the asymptotic normality properties. Note that the condition $\alpha > (2C_9/k_l)^{1/k}$ is required to guarantee the invertibility of matrix $E[(\nabla^2 \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{SS}]$.

Remark. To further understand the condition $\alpha^{1-k} = o(n^{-1/2})$, we take $\alpha = \mathcal{O}\left(\sqrt{\frac{n}{\log p}}\right)$ as an example, the fastest divergent rate indicated in Theorem III.3. Then the condition requires $\frac{\log p}{n} \cdot n^{\frac{1}{k-1}} \to 0$. Thus $\frac{1}{k-1} < 1$ and then $k > 2$. Therefore the asymptotic normality result holds only when at least the third moment of $\epsilon|\mathbf{x}$ exists. In particular, when $k = 3$, we obtain $n^{-\frac{1}{2}} \log p \to 0$.

III.4. Implementation of the PRAM Estimators

Note that the optimization in (III.2) may not be a convex optimization problem since we allow both loss function $\mathcal{L}_{\alpha,n}$ and ρ_λ to be non-convex. To obtain the corresponding stationary point, we use the composite gradient descend algorithm [Nes13]. Denoting $q_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 - \rho_\lambda(\boldsymbol{\beta})$ and $\bar{L}_{\alpha,n}(\boldsymbol{\beta}) = \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) - q_\lambda(\boldsymbol{\beta})$, we can rewrite the program as

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1 \le R}{\operatorname{argmin}} \left\{ \bar{L}_{\alpha,n}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\| \right\}.$$

Then the composition gradient iteration is given by

$$\boldsymbol{\beta}^{t+1} \in \underset{\|\boldsymbol{\beta}\|_1 \le R}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - (\boldsymbol{\beta}^t - \eta\nabla\bar{L}_{\alpha,n}(\boldsymbol{\beta}^t))\|_2^2 + \eta\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{III.13}$$

where $\eta > 0$ is the step size for the update and can be determined by the backtracking line search method described in [Nes13]. A simple calculation shows that the iteration in (III.13) takes the form

$$\boldsymbol{\beta}^{t+1} = S_{\eta\lambda}\left(\boldsymbol{\beta}^t - \eta\nabla\bar{L}_{\alpha,n}(\boldsymbol{\beta}^t)\right),$$

where $S_{\eta\lambda}(\cdot)$ is the soft-thresholding operator defined as

$$[S_{\eta\lambda}(\boldsymbol{\beta})]_j = sign(\beta_j)\left(|\beta_j| - \eta\lambda\right)_+ .$$

We further adopt the two-step procedure discussed in [Loh17] to guarantee the convergence to a stationary point for the non-convex optimization problem:

Step 1: Run the composite gradient descent using the convex Huber loss function with the convex Lasso penalty to get an initial PRAM estimator.

Step 2: Run the composite gradient descent on the desired high-dimensional PRAM estimator using the initial PRAM estimator from Step 1.

For tuning parameters selection, the optimal values of α and λ are chosen by a two-dimensional grid search using the cross-validation. In Particular, the searching grid is formed by partitioning a rectangle uniformly in the scale of $(\alpha, \log(\lambda))$. The optimal values are found by the combination that minimizes the cross-validated trimmed mean squared prediction error.

III.5. Simulation Studies

In this section, we assess the performance of the PRAM estimators by considering different types of loss and penalty functions through various models. The simulation setting is similar to the one in [FLW17]. The data is generated from the following model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i.$$

We choose the true regression coefficient vector as $\boldsymbol{\beta}^* = (\mathbf{3}_5^T, \mathbf{2}_5^T, \mathbf{1.5}_5^T, \mathbf{0}_{p-15}^T)^T$, where the first 15 elements consist of 5 numbers of 3, 2, 1.5 receptively and the rest are 0. In all simulation settings, we let *n=100* and *p=500*.

Example III.1. (Homogeneous case) The covariates vector $\mathbf{x}_i$s are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{I}_p$ independently. The random errors $\epsilon_i = e_i - E[e_i]$, where e_i are generated independently from the following 5 scenarios:

(a). $N(0, 4)$: Normal with mean 0 and variance 4;

(b). $\sqrt{2}t_3$: $\sqrt{2}$ times the t-distribution with degrees of freedom 3;

(c). MixN: Equal mixture of Normal distributions N(-1, 4) and N(8, 1);

(d). LogNormal: Log-normal distribution such that $e_i = \exp(1.3z_i)$, where $z_i \sim N(0, 1)$.

(e). Weibull: Weibull distribution with the shape parameter 0.3 and the scale parameter 0.15.

We consider three types of loss functions equipped with diverging parameters (the Huber loss, Tukey's biweight loss and Cauchy loss) and two types of penalty functions (the Lasso and MCP penalties). Thus it produces 6 different PRAM estimators: HA-Lasso, TA-Lasso, CA-Lasso, HA-MCP, TA-MCP and CA-MCP. Note the HA-Lasso becomes the RA-Lasso estimator in [FLW17], where the HA-Lasso has been demonstrated to perform better than the Lasso and R-Lasso, especially when the errors were asymmetric and heavy-tailed (LogNormal and Weibull). Thus in our simulation we skip those comparisons and only evaluate the performance of all those 6 PRAM estimators. Their performances on both mean estimation and variable selection under the five scenarios were reported by the following five measurements:

(1) L_2 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

(2) L_1 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$.

(3) Model size (MS), the average number of selected covariates.

(4) False positives rate (FPR), the percent of selected but unimportant covariates:

$$FPR = \frac{|\hat{S} \bigcap S^c|}{|S^c|} \times 100\%. \tag{III.14}$$

(5) False negatives rate (FNR), the percent of non-selected but important covariates:

$$FNR = \frac{|\hat{S}^c \bigcap S|}{|S|} \times 100\%. \tag{III.15}$$

Here $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ and $S = \{j : \beta_j^* \neq 0\}$. The model considered in Example III.1 is homogeneous, in which the error distribution is independent of covariate $\mathbf{x}$. We also assess the performance of PRAM estimators under heteroscedastic model in the next example.

Example III.2. (Heteroscedastic case) We generate the data from

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + c^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}^*)^2 \epsilon_i,$$

where the constant $c = \sqrt{3}\|\boldsymbol{\beta}^*\|_2^2$ makes $E[c^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}^*)^2]^2 = 1$. We also consider $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_p)$ and generate the random error ϵ from the same five scenarios described in Example III.1.

Finally, we design a simulation setting to evaluate the performance of the generalized PRAM estimators under weaker distribution assumptions on the covariates.

Example III.3. (Non-Gaussian x case) Similar to Example III.1, except that the covariate $\mathbf{x}$ in 20% of observations are first generated from independent chi-square variables with 10 degrees of freedom, and then recentered to have mean zero.

For all three examples described above, we run 100 simulations for each scenario. In Example III.3, we consider the generalized PRAM estimators with $v(\mathbf{x}) \equiv 1$ and $w(\mathbf{x}) = \min\left\{1, \frac{4}{\|\mathbf{x}\|_\infty}\right\}$. For all six PRAM estimators, tuning parameters λ and α are chosen optimally by 10-fold cross-validation, with α ranges in $(0.1\sqrt{\frac{n}{\log p}}, 10\sqrt{\frac{n}{\log p}})$ and λ ranges in $(0.01\sqrt{\frac{\log p}{n}}, 2.5\sqrt{\frac{\log p}{n}})$. These ranges are motivated from Theorem III.2. The mean values out of 100 iterations (with standard errors in parentheses) are reported in Table III.1, III.2, III.3, respectively.

We have two findings based on results in Table III.1 and III.2. Firstly, all the MCP-type PRAM estimators largely outperform the Lasso-type estimators in all the measurements, rendering satisfactory finite sample performances under different settings. This is consistent with the oracle property of the PRAM estimators using a proper non-convex penalty stated in Theorem III.3. Secondly, for estimators with the same penalty, although all estimators perform comparably for light-tailed settings ($N(0, 4)$ and MixN), the TA-type and CA-type PRAM estimators outperform the HA-type estimators using the same penalty in heavy-tailed settings ($\sqrt{2}t_3$, LogNormal and Weibull). This is actually not surprising due to the following two facts: (1) redescending M-estimators can achieve the minimax variance sensitivity under certain global minimax criterion [SMS08]; (2) the HA-Lasso estimation is used as the initial in the optimization process of TA-type and CA-type PRAM estimators. Note that the error terms $c^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}^*)^2\epsilon_i$ in the heteroscedastic model have the same variance as those in the homogeneous model, however, their distribution possess heavier tails. Hence in the heteroscedastic model, except for a few errors being far away on tail, most of the others get even closer to the center. This fact explains why the performances in Table III.2 are consistently better than those in Table III.1.

Table III.1. Simulation Results under the Homogeneous Model with Standard Normal Covariates in Example III.1. The mean L_2 error, L_1 error, MS, FPR (%) and FNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		HA-Lasso	TA-Lasso	CA-Lasso	HA-MCP	TA-MCP	CA-MCP
N(0,4)	L_2 error	3.3 (0.9)	3.31 (0.94)	3.3 (0.9)	0.99 (0.47)	1.01 (0.53)	0.94 (0.27)
	L_1 error	17.75 (4.2)	17.79 (4.33)	17.72 (4.12)	3.29 (2.07)	3.34 (2.22)	3.06 (1.01)
	MS	67.32 (9.11)	66.99 (10.15)	66.95 (10.13)	17.21 (2.47)	16.84 (2.46)	16.71 (2.36)
	FPR, FNR	10.85, 2.13	10.8, 2.53	10.79, 2.4	0.46, 0.27	0.4, 0.53	0.35, 0.07
$\sqrt{2}t_3$	L_2 error	3.59 (0.96)	3.62 (1)	3.56 (1)	1.18 (0.9)	1.13 (0.89)	1.14 (0.93)
	L_1 error	19.09 (5.03)	19.26 (5.14)	19.04 (5.12)	3.95 (3.65)	3.78 (3.56)	3.76 (3.65)
	MS	63.72 (9.76)	64.07 (11.39)	65.13 (9.55)	16.85 (2.11)	16.7 (2.53)	16.2 (2.17)
	FPR, FNR	10.14, 3	10.23, 3.53	10.43, 3.13	0.43, 1.6	0.4, 1.53	0.31, 1.87
MixN	L_2 error	3.48 (0.78)	3.48 (0.79)	3.5 (0.8)	1.25 (0.71)	1.27 (0.73)	1.25 (0.69)
	L_1 error	18.99 (3.71)	18.99 (3.72)	19.05 (3.8)	4.2 (2.97)	4.17 (2.79)	4.11 (2.72)
	MS	68.12 (8.85)	68.14 (9.06)	67.65 (9.4)	17.52 (3.57)	17.05 (3.8)	17.06 (5.43)
	FPR, FNR	11, 1.6	11.01, 1.73	10.92, 2	0.55, 0.93	0.47, 1.47	0.46, 1.2
LogNormal	L_2 error	4.66 (1.2)	4.56 (1.13)	4.5 (1.24)	2.13 (2.05)	1.74 (1.68)	2.12 (1.97)
	L_1 error	23.84 (6.2)	23.75 (5.63)	23.44 (6.15)	7.69 (8.52)	5.88 (6.61)	7.4 (7.88)
	MS	57.16 (11.44)	60.68 (14.11)	60.64 (12.13)	16.7 (3.61)	16.03 (2.69)	16.29 (7.03)
	FPR, FNR	8.97, 8.93	9.68, 8.53	9.68, 8.73	0.62, 8.73	0.41, 6.53	0.58, 10.13
Weibull	L_2 error	3.91 (1.06)	3.63 (1.05)	3.46 (1.08)	1.35 (1.43)	0.94 (1.15)	1.03 (1.26)
	L_1 error	19.62 (5.38)	19.15 (5.36)	18.17 (5.65)	4.64 (5.73)	3.18 (4.5)	3.42 (4.69)
	MS	55.37 (11.91)	64.12 (11.96)	63.5 (8.98)	16.15 (2.47)	15.65 (1.76)	15.44 (1.65)
	FPR, FNR	8.51, 5.87	10.26, 4.13	10.09, 3.07	0.36, 4.07	0.2, 2.13	0.18, 2.87

Table III.2. Simulation Results under the Heteroscedastic Model with Standard Normal Covariates in Example III.2. The mean L_2 error, L_1 error, MS, FPR (%) and FNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		HA-Lasso	TA-Lasso	CA-Lasso	HA-MCP	TA-MCP	CA-MCP
N(0,4)	L_2 error	2.84 (0.81)	2.94 (0.91)	2.72 (0.84)	0.55 (0.35)	0.55 (0.19)	0.6 (0.21)
	L_1 error	14.74 (4.14)	15.45 (4.84)	14.13 (4.38)	1.78 (1.16)	1.73 (0.64)	1.91 (0.82)
	MS	61.56 (9.65)	63.25 (11.03)	62.11 (8.42)	15.68 (1.27)	15.28 (0.87)	15.43 (1.71)
	FPR, FNR	9.62, 0.67	9.98, 1.13	9.74, 0.73	0.14, 0.07	0.06, 0	0.09, 0
$\sqrt{2}t_3$	L_2 error	2.88 (0.94)	2.89 (0.96)	2.67 (0.96)	0.48 (0.28)	0.51 (0.16)	0.54 (0.15)
	L_1 error	14.64 (4.78)	14.87 (4.77)	13.74 (5)	1.54 (0.93)	1.61 (0.53)	1.72 (0.49)
	MS	59.54 (11.57)	61.11 (11.62)	61.39 (9.38)	15.69 (1.13)	15.34 (1.17)	15.54 (3.05)
	FPR, FNR	9.22, 1.07	9.55, 1.47	9.59, 0.93	0.14, 0	0.07, 0	0.11, 0
MixN	L_2 error	3.25 (0.87)	3.33 (0.94)	3.17 (0.93)	0.67 (0.35)	0.64 (0.22)	0.64 (0.2)
	L_1 error	16.86 (4.64)	17.53 (5.17)	16.58 (5.01)	2.16 (1.3)	2.02 (0.73)	2.04 (0.69)
	MS	61.23 (10.51)	62.36 (10.93)	62.55 (8.76)	15.87 (1.91)	15.29 (1.01)	15.24 (0.67)
	FPR, FNR	9.57, 1.27	9.82, 1.67	9.85, 1.33	0.18, 0	0.06, 0	0.05, 0
LogNormal	L_2 error	3.68 (1.05)	3.64 (1)	3.4 (1.05)	0.9 (1.03)	0.64 (0.36)	0.74 (0.77)
	L_1 error	18.76 (5.16)	19.08 (4.87)	17.72 (5.37)	2.95 (3.68)	2.03 (1.19)	2.43 (3.05)
	MS	58.63 (10.39)	63.13 (11.89)	62.99 (8.07)	15.62 (1.72)	15.26 (0.69)	15.18 (0.66)
	FPR, FNR	9.12, 4.07	10.04, 3.73	9.98, 2.67	0.19, 1.93	0.06, 0.2	0.07, 1.07
Weibull	L_2 error	3.01 (1.19)	2.83 (1.09)	2.66 (1.11)	0.75 (0.9)	0.59 (0.52)	0.64 (0.67)
	L_1 error	15.09 (6.07)	14.89 (5.68)	13.66 (5.63)	2.5 (3.48)	1.94 (2.1)	2.11 (2.67)
	MS	57.28 (11.85)	65.07 (9.4)	61.77 (7.9)	15.71 (1.21)	15.39 (0.91)	15.3 (0.98)
	FPR, FNR	8.78, 2.13	10.37, 1.53	9.69, 1.53	0.19, 1.27	0.09, 0.27	0.08, 0.67

In Example III.3, we only report results from the MCP-type PRAM estimators, since they have been shown to perform better than the Lasso-type estimators. In the

homogeneous model with non-Gaussian covariates, Table III.3 clearly indicates that the PRAM estimators with well chosen $w(\mathbf{x})$ perform better in all cases than those PRAM with $w(\mathbf{x}) = 1$. In addition, among those three weighted PRAM estimators, the weighted TA-MCP (WTA-MCP) and the weighted CA-MCP (WCA-MCP) again show advantages over the weighted HA-MCP (WHA-MCP) when the errors are heavy-tailed, which is consistent with the findings obtained in Example III.1 and III.2.

In conclusion, the PRAM estimator with a folded concave penalty (e.g. MCP penalty) render promising performances in different settings, which is consistent with our theoretical results. Our simulation study also shed some lights on how to implement robust high-dimensional M-estimators for real applications: when the data are strongly heavy-tailed or contaminated, regardless of asymmetry and/or heteroscedasticity, a redescending PRAM estimator with a concave penalty yields better performance than a convex PRAM estimator in practice.

Table III.3. Simulation Results under the Homogeneous Model with Non-Gaussian Covariates in Example III.3. The mean L_2 error, L_1 error, MS, FPR (%) and FNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		HA-MCP	WHA-MCP	TA-MCP	WTA-MCP	CA-MCP	WCA-MCP
N(0,4)	L_2 error	0.87 (0.91)	0.69 (0.61)	1.19 (1.58)	0.75 (0.83)	0.93 (1.1)	0.69 (0.6)
	L_1 error	3.65 (4.29)	2.38 (2.71)	5.14 (7.6)	2.62 (3.77)	3.89 (5.17)	2.39 (2.7)
	MS	36.92 (12.84)	17.7 (4.16)	36.88 (13.86)	17.69 (4.12)	36.07 (13.11)	17.89 (4.57)
	FPR, FNR	4.53, 0.2	0.56, 0.27	4.58, 2.27	0.58, 0.73	4.36, 0.67	0.6, 0.27
$\sqrt{2}t_3$	L_2 error	0.91 (0.72)	0.65 (0.28)	1.11 (1.55)	0.63 (0.34)	0.87 (0.85)	0.61 (0.26)
	L_1 error	3.75 (3.41)	2.12 (1.07)	4.8 (7.42)	2.07 (1.27)	3.56 (3.96)	1.98 (0.96)
	MS	36.39 (11.54)	16.77 (2.78)	35.75 (11.68)	16.56 (2.63)	35.15 (11.86)	16.48 (2.71)
	FPR, FNR	4.42, 0.27	0.36, 0	4.36, 2.67	0.32, 0	4.16, 0.33	0.31, 0
MixN	L_2 error	0.95 (0.89)	0.82 (0.71)	1.29 (1.52)	0.83 (0.75)	0.98 (0.99)	0.82 (0.74)
	L_1 error	4.08 (4.25)	2.9 (3.31)	5.71 (7.33)	2.9 (3.4)	4.14 (4.69)	2.85 (3.4)
	MS	38.22 (11.12)	18.5 (3.8)	39.42 (11.2)	17.9 (3.73)	37.65 (12.55)	17.88 (3.61)
	FPR, FNR	4.8, 0.47	0.74, 0.53	5.09, 1.8	0.61, 0.47	4.69, 0.8	0.61, 0.4
LogNormal	L_2 error	2.26 (2.19)	1.31 (1.5)	2.74 (2.44)	1.38 (1.71)	2.02 (1.94)	1.36 (1.73)
	L_1 error	10.24 (10.45)	4.8 (6.48)	12.51 (11.68)	5.19 (7.53)	9.07 (9.31)	4.85 (6.77)
	MS	40.23 (11.37)	18.03 (4.55)	43.42 (12.09)	18.11 (4.35)	41.04 (12.25)	16.6 (3.6)
	FPR, FNR	5.39, 6	0.74, 3.67	6.16, 9.67	0.79, 4.93	5.52, 4.87	0.5, 5.6
Weibull	L_2 error	1.6 (1.98)	1.11 (1.75)	1.84 (2.26)	0.92 (1.55)	1.44 (1.9)	0.85 (1.36)
	L_1 error	7.11 (9.73)	4.34 (8.28)	8.04 (10.19)	3.34 (6.36)	6.17 (8.6)	3.03 (5.55)
	MS	37.25 (11.37)	17.45 (5.57)	38.69 (13.33)	17.57 (4.88)	35.74 (10.84)	16.83 (3.61)
	FPR, FNR	4.7, 3.6	0.62, 3.8	5.08, 6.4	0.62, 2.93	4.38, 3.4	0.44, 2.13

III.6. Real Data Example

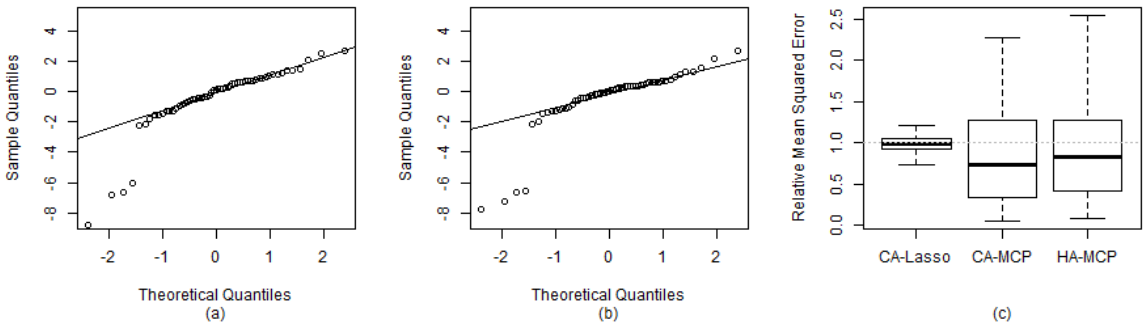
In this section, we use the NCI-60 data introducted in I.1. We perform some pre-screenings and keep only p_1 genes with largest variations and then choose p_2 genes out of them which are most correlated with the response variable. Here the final dataset is obtained by choosing $p_1 = 2000$ and $p_2 = 500$, yielding $n = 59$ and $p = 500$ for PRAM data analysis. Similar to our simulation studies, we then apply 6 PRAM estimators to select important genes, with tuning parameters α and λ chosen from the 10-fold cross validation. Since the TA-type and CA-type PRAM estimators perform similarly, we will only report results from four methods: HA-Lasso, CA-Lasso, HA-MCP and CA-MCP.

The number of selected genes from four PRAM methods are 27 (HA-Lasso), 31 (CA-Lasso), 12 (HA-MCP), 5 (CA-MCP), respectively. HA-Lasso and CA-Lasso that selected 27 and 31 genes respectively could potentially result in over selection since the total sample size is only 59. Figure III.1(a) and Figure III.1(b) show that the residual distributions generated from HA-MCP and CA-MCP both had a longer tail on the left side. It indicates that PRAM estimators with non-convex penalties can be resistant to the data contamination or data's irregularity due to the flexible robustness and nice variable selection property.

Table III.4. Selected Genes and the Corresponding Coefficient Estimation by HA-MCP and CA-MCP. Probe IDs are listed in parentheses.

HA-MCP	KRT8	NRN1	GPX3	CELF2	CELF2	LEF1
	(209008_x_at)	(218625_at)	(201348_at)	(202156_s_at)	(202157_s_at)	(221558_s_at)
	6.230	-1.505	0.031	-0.002	0.000	-0.003
	MEST	FAR2	PBX1	CLEC11A	CLEC11A	ATP2A3
	(202016_at)	(220615_s_at)	(212148_at)	(205131_x_at)	(211709_s_at)	(213036_x_at)
	0.009	-0.037	0.035	-0.036	-0.017	-0.003
CA-MCP	KRT8	NRN1	GPX3	GPNMB	ATP2A3	
	(209008_x_at)	(218625_at)	(201348_at)	(201141_at)	(213036_x_at)	
	6.122	-0.775	0.693	-0.556	-0.763	

Figure III.1. (a) The QQ Plot of the Residuals from HA-MCP. (b) The QQ Plot of the Residuals from CA-MCP. (c) The Boxplot of the Relative Mean Squared Prediction Errors.



For the sake of simplicity, we only report those selected genes and corresponding coefficient estimation by HA-MCP and CA-MCP in Table III.4. According to our analysis, genes KRT8, NRN1 and GPX3 are selected by all four methods. It is not surprising for gene KRT8 since it has the largest correlation with the response variable and has a long history of being paired with KRT18 in cancer studies for cell death and survival, cellular growth and proliferation, organismal injury and abnormalities, and so on [LZ16, WHM$^+$07]. Gene NRN1 was investigated to be involved in melanoma migration, attachment independent growth, and vascular mimicry [BSE$^+$17]. Recent studies showed that gene GPX3 plays as a tumor suppressor in lung cancer cell line [ACO$^+$18] and its down-regulation is related to pathogenesis of melanoma [CZK$^+$16]. Notice that gene ATP2A3 is also singled out by both HA-MCP and CA-MCP. This gene encodes the enzyme involved in calcium sequestration associated with muscular excitation and contraction, and was shown to act an important role in resveratrol anticancer activity in breast cancer cells [ITRMMZH17]. In addition, Table III.4 indicates that gene GPNMB is only selected by CA-MCP. The GPNMB expression

was found to be associated with reduction in disease-free and overall survival in breast cancer and its over-expression had been identified in numerous cancers [MRAS13]. Therefore, both genes (ATP2A3 and GPNMB) deserve further study in genetics research.

To further evaluate the prediction performance of those PRAM estimators, we randomly choose 6 observations as the test set and applied four methods to the rest patients to get the coefficients estimation, then compute the prediction error on the test set. We repeat the random splitting 100 times and the boxplots of the Relative Mean Squared Prediction Error (RPE) with respect to HA-Lasso are shown in Figure III.1(c). A method with RPE < 1 indicates a better performance than HA-Lasso. It is clearly seen from Figure III.1(c) that the MCP-type PRAM estimators have better predictions than those from the Lasso-type estimators, even though they select much smaller number of variables. In addition, Figure III.1(c) together with Table III.4 show that a redescending PRAM estimator with a non-convex penalty (e.g. CA-MCP) is more likely to give a more parsimonious model with better prediction performance, which is consistent with the findings from our simulation studies.

CHAPTER IV

HIGH-DIMENSIONAL M-ESTIMATION FOR BI-LEVEL VARIABLE SELECTION

IV.1. Introduction

Covariates often function group-wisely in many applications. For example, in gene expression analysis, genes from the same biological pathways may exhibit similar activities. In high-dimensional linear regression, penalized least squares approaches with penalties incorporating grouping structures have become very popular in recent decades. [YL06] proposed the group Lasso, as a nature extension of the Lasso [Tib96b], to select variables at the group level by applying the Lasso penalty on the ℓ_2 norm of coefficients associated with each group of variables in penalized least squares regression (LS-GLasso). To address the bias and inconsistency of the group Lasso estimator in high-dimensional settings, several methods have been investigated, including the adaptive group Lasso [WH10], the ℓ_2-norm MCP [HBM12], the ℓ_2-norm SCAD [GZWW15], among others. However, above approaches encourage only "all-in" or "all-out" variable selection at the group level. To further encourage the sparsity within certain groups, extensive methods have been proposed to perform bi-level variable selection. See for example the group Bridge [HMXZ09], the sparse group Lasso [FHT10, SFHT13], the concave ℓ_1-norm group penalty [JH14], the composite MCP [BH09], the group exponential Lasso [Bre15], among others. See [HBM12] for a complete review.

When the data dimensionality grows much faster than the sample size, irregular settings often appear, such as the response and a large number of variables are contaminated or heavy-tailed. It has been shown that the LS-GLasso is estimation

consistent when the random errors are sub-Gaussian [WH10]. However, the quadratic loss in LS-GLasso is non-robust to outliers and the estimator is no longer consistent if the random errors are wildly deviated from sub-Gaussian distribution. In addition, the required restricted eigenvalue condition on design matrix may not hold if the predictors are non-Gaussian.

To tackle the problem of heavy-tailed random errors in high-dimensional settings, a few robust penalized approaches have been recently studied. [Lil15] proposed the penalized least absolute deviation (LAD) estimator with the group Lasso penalty to relieve the model's sensitivity due to the existence of outliers in random errors. This method was also extended to the weighted LAD group Lasso when some predictors are contaminated or heavy-tailed. [WT16] investigated a general penalized M-estimators framework using convex loss functions and concave ℓ_2-norm penalties for the partially linear model with grouped covariates. However, those above robust methods can only select variables at group level and thus do not perform bi-level variable selection. In the examples of gene expression study, while the data may be heavy-tailed or contaminated due to the complex data generation procedures, we may be still interested in selecting important genes as well as important groups.

Additionally, the above robust methods all require the loss function to be convex. It is well known that the convex loss functions such as Huber loss and LAD loss do not downweight the very large residuals due to their convexity. [SMS08] showed that redescending M-estimators with non-convex loss function possess certain optimal robustness properties. In fact, there still lacks a systematic study of high-dimensional M-estimators that perform robust bi-level variable selection, allowing both loss and group penalty functions to be non-convex.

In this chapter, we consider high-dimensional linear regression with grouped covariates, in irregular settings that the data (random errors and/or covariates) may be contaminated or heavy-tailed. In particular, we propose a novel high-dimensional bi-level variable selection method through a two-stage penalized M-estimator framework: penalized M-estimation with a concave ℓ_2-norm penalty achieving the consistent group selection at the first stage, and a post-hard-thresholding operator to achieve the within-group sparsity at the second stage.

The remainder of this chapter is organized as follows. In Section IV.2, we introduce a basic setup for our two-stage penalized M-estimator framework. In Section IV.3, we present statistical properties of our proposed bi-level M-estimators under some sufficient conditions. We discuss the implementation of the two-stage M-estimators in Section IV.4. In Section IV.5, we conduct some simulation studies to demonstrate the performance of the proposed estimators under different settings. We also apply the proposed estimators for NCI-60 data analysis and illustrate all results in Section IV.6. All technical proofs are relegated to Appendix.

IV.2. The Two-stage M-estimator Framework

Let's consider a high-dimensional data with p covariates from J non-overlapping groups. A linear regression model can be written as

$$y_i = \sum_{j=1}^{J} \mathbf{x}_{ij}^T \boldsymbol{\beta}_j^* + \epsilon_i, \quad i = 1, \cdots, n, \tag{IV.1}$$

where ϵ_is are i.i.d random errors, $\mathbf{x}_{ij}$s are independent and identically distributed (i.i.d) d_j-dimensional covariate vectors corresponding to the jth group, $\boldsymbol{\beta}_j^*$ is the d_j-dimensional true regression coefficient vector of the jth group. Then $p = \sum_{j=1}^{J} d_j$. Let $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{iJ}^T)^T$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \cdots, \boldsymbol{\beta}_J^{*T})^T$. We assume the independence between covariates $\mathbf{x}_i$ and random errors ϵ_i for the sake of simplicity. We also assume

the group sparsity condition of the model: there exists $S \subseteq \{1, \cdots, J\}$ such that $\boldsymbol{\beta}_j^* = \mathbf{0}$ for all $j \notin S$. Note that we allow the within-group sparsity on some $\boldsymbol{\beta}_j^* \neq \mathbf{0}$ and thus there exists bi-level sparsity on the coefficient vector $\boldsymbol{\beta}^*$.

Some More Notations. We use bold symbols to denote matrices or vectors. Let β_m be the mth element of $\boldsymbol{\beta} \in \mathbb{R}^p$. For any $A \subseteq \{1, 2, \cdots, p\}$, we denote $\boldsymbol{\beta}_A = (\beta_m,\ m \in A)^T$ a coefficient sub-vector with indexes in A. Define $d_a := \max_{1 \leq j \leq J} d_j$, $d_b := \min_{1 \leq j \leq J} d_j$, $d := \sqrt{\frac{d_a}{d_b}}$. Let $I_j \subseteq \{1, 2, \cdots, p\}$ denote the index set of coefficients in group j. Then $I_S := \bigcup_{j \in S} I_j$ includes all indexes of coefficients in those important groups. Let $I_0 = \{m : \beta_m^* \neq 0, 1 \leq m \leq p\}$ and thus $I_0 \subseteq I_S$. Define $\boldsymbol{\beta}_{\min}^{*G} := \min_{j \in S} \|\boldsymbol{\beta}_j^*\|_2$ as the minimum group strength on $\boldsymbol{\beta}^*$, where $\|\cdot\|_2$ is the ℓ_2 norm. Define $\boldsymbol{\beta}_{\min}^{*I} := \min_{m \in I_0} |\beta_m^*|$ as the minimum individual signal strength on $\boldsymbol{\beta}^*$. Let $s = |S|$ and $k = |I_S|$ be the number of important groups and number of variables among all important groups, respectively. We denote $u_+ = \max(u, 0)$ for any $u \in \mathbb{R}$.

Our Proposed M-estimator Framework for Bi-level Variable Selection. To perform an efficient bi-level variable selection with potential robustness for the existence of possible data contamination or heavy-tailed distribution between ϵ_i and $\mathbf{x}_i$, we propose the following two-stage penalized M-estimator framework:

- *Group Penalization (GP) Stage.* First we perform penalized M-estimation with a group concave penalty achieving the between-group sparsity:

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda) \right\}.$$

- *Hard-thresholding (HT) Stage.* Then we apply a post-hard-thresholding operator on $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}^h(\theta) = \hat{\boldsymbol{\beta}} \cdot I(|\hat{\boldsymbol{\beta}}| \geq \theta) \tag{IV.2}$$

where "$\cdot$" and "$\geq$" in (IV.2) are elementary-wise.

Note that $\mathcal{L}_n$ is an empirical loss function may encourage a robust solution and ρ is a penalty function, which encourages the group sparsity in the solution. Here λ and θ are two tuning parameters controlling the between-group and within-group sparsity, respectively. We include the side condition $\|\boldsymbol{\beta}\|_1 \leq R$ in the Group Penalization Stage in order to guarantee the existence of local/global optima, for the case where the loss or regularizer may be non-convex. In real applications, we can choose R to be a sufficiently large number such that $\|\boldsymbol{\beta}^*\|_1 \leq R$.

Let $l : \mathbb{R} \mapsto \mathbb{R}$ denote a residual function, or a loss function, defined on each observation pair $(\mathbf{x}_i, y_i)$. Then the above Group Penalization Stage becomes

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda) \right\}. \tag{IV.3}$$

With a well chosen l, the penalized M-estimator from (IV.3) can be robust to heavy-tailed random error ϵ_i. Some typical robust loss functions l include:

- **Huber Loss**

$$l(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \leq \alpha, \\ \alpha|u| - \frac{\alpha^2}{2} & \text{if } |u| \geq \alpha. \end{cases}$$

- **Tukey's biweight Loss**

$$l(u) = \begin{cases} \frac{\alpha^2}{6}(1 - (1 - \frac{u^2}{\alpha^2})^3) & \text{if } |u| \leq \alpha, \\ \frac{\alpha^2}{6} & \text{if } |u| \geq \alpha. \end{cases}$$

- **Cauchy Loss**

$$l(u) = \frac{\alpha^2}{2} \log\left(1 + \frac{u^2}{\alpha^2}\right).$$

The derivatives of above three loss functions are bounded and thus they can mitigate the effect of larger residuals. In particular, the Tukey's biweight loss and Cauchy loss produce redescending M-estimators. From the robust regression literature, we call an M-estimator redescending if there exists $u_0 > 0$ such that $|l'(u)| = 0$ or decrease to 0 smoothly, for all $|u| \geq u_0$. In that case, strong robustness is obtained by ignoring the large outliers completely. See more discussions in [Mul04] and [SMS08].

Whereas the robust loss function in (IV.3) takes into account the contamination or heavy-tailed distribution in error ϵ_i, a single outlier in $\mathbf{x}_i$ may still cause the corresponding estimator to perform arbitrarily badly. To downweight large values of $\mathbf{x}_i$, we extend the Group Penalization Stage in (IV.3) to

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)} l((y_i - \mathbf{x}_i^T \boldsymbol{\beta}) v(\mathbf{x}_i)) + \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda) \right\}, \quad \text{(IV.4)}$$

where w, v are weight functions such that $w, v > 0$. A few options for choosing those weight functions can be found in [Mal75], [Hil77], [MS71] and [Loh17].

Since $\boldsymbol{\beta}_j^* = \mathbf{0}$ for $j \notin S$, we need the Group Penalization Stage to generate sparse solutions between groups. In particular, we require the penalty function ρ in (IV.4) to satisfy amendable properties listed in Assumption IV.1.

Assumption IV.1 (Penalty Function Assumptions)**.** *$\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is a scalar function that satisfies the following conditions:*

(i) For any fixed $t \in \mathbb{R}^+$, the function $\lambda \mapsto \rho(t, \lambda)$ is non-decreasing on $\mathbb{R}^+$.

(ii) There exists a scalar function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that for any $r \in [1, \infty)$,

$\frac{\rho(t,r\lambda)}{\rho(t,\lambda)} \leq g(r)$ *for all* $t, \lambda \in \mathbb{R}^+$.

(iii) *The function* $t \mapsto \rho(t,\lambda)$ *is symmetric around zero and* $\rho(0,\lambda) = 0$*, given any fixed* $\lambda \in \mathbb{R}$.

(iv) *The function* $t \mapsto \rho(t,\lambda)$ *is non-decreasing on* $\mathbb{R}^+$*, given any fixed* $\lambda \in \mathbb{R}$.

(v) *The function* $t \mapsto \frac{\rho(t,\lambda)}{t}$ *is non-increasing on* $\mathbb{R}^+$*, given any fixed* $\lambda \in \mathbb{R}$.

(vi) *The function* $t \mapsto \rho(t,\lambda)$ *is differentiable for* $t \neq 0$*, given any fixed* $\lambda \in \mathbb{R}$.

(vii) $\lim_{t\to 0^+} \frac{\partial \rho(t,\lambda)}{\partial t} = \lambda$*, given any fixed* $\lambda \in \mathbb{R}$.

(viii) *There exists* $\mu > 0$ *such that the function* $t \mapsto \rho(t,\lambda) + \frac{\mu}{2}t^2$*is convex, given any fixed* $\lambda \in \mathbb{R}$.

(ix) *There exists* $\delta \in (0,\infty)$ *such that* $\frac{\partial \rho(t,\lambda)}{\partial t} = 0$ *for all* $t \geq \delta\lambda$*, given any fixed* $\lambda \in \mathbb{R}$.

The properties (iii-ix) in Assumption 1 are related to the penalty functions studied in [Loh17] and [LW13]. Adopting the notation from [Loh17], we consider ρ to be μ-amenable if ρ satisfies conditions (i)-(viii). If ρ also satisfies condition (ix), we say that ρ is (μ,δ)-amenable. In particular, if ρ is μ-amenable, then $q(t,\lambda) := \lambda|t| - \rho(t,\lambda)$ is everywhere differentiable. Define the vector version $q_\lambda(\boldsymbol{\beta}) := \sum_{j=1}^J q(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda)$ accordingly. It is easy to see that there exists $\mu > 0$ such that $\frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2 - q_\lambda(\boldsymbol{\beta})$ is convex. This property is important for both computational implementation and theoretical investigation of the group selection properties.

Some popular choices of amenable penalty functions include Lasso [Tib96b], SCAD [FL01], and MCP [Z+10] given as follows:

- The **Lasso** penalty $\rho(t,\lambda) = \lambda|t|$ is 0-amenable but not $(0,\delta)$-amenable for any $\delta < \infty$.

- This **SCAD** penalty takes the form

$$\rho(t,\lambda) = \begin{cases} \lambda|t| & for\ |t| \le \lambda, \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & for\ \lambda < |t| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & for\ |t| > a\lambda, \end{cases}$$

  where $a > 2$ is fixed. The SCAD penalty is (μ,δ)-amenable with $\mu = \frac{1}{a-1}$ and $\delta = a$.

- The **MCP** penalty takes the form

$$\rho(t,\lambda) = sign(t)\lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

  where $b > 0$ is fixed. The MCP penalty is (μ,δ)-amenable with $\mu = \frac{1}{b}$ and $\delta = b$.

It has been shown that a folded concave penalty, such as the SCAD or MCP, often has better variable selection properties than the convex penalty including the Lasso.

IV.3. Statistical Properties

In this section, we present our theoretical results for the proposed two-stage penalized M-estimator framework. We begin with statistical properties of the estimator $\hat{\boldsymbol{\beta}}$ in program (IV.4) generated from the Group Penalization Stage. On the one hand, we show a general non-asymptotic bound of the estimation error and establish the local estimation consistency of $\hat{\boldsymbol{\beta}}$ at the minimax rate enjoyed by the LS-GLasso, under certain mild conditions. On the other hand, we show that the estimator $\hat{\boldsymbol{\beta}}$ in fact equals the local oracle solution with the correct group support and thus obtain the group-level oracle properties. Finally, we show that those nice statistical properties of

$\hat{\boldsymbol{\beta}}$ can be carried over during the hard-thresholding stage and we establish the bi-level variable selection consistency of $\hat{\boldsymbol{\beta}}^h$. All proofs are given in Appendix.

As introduced in (IV.4), the loss function in the two-stage penalized M-estimator framework takes the following form,

$$\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)} l((y_i - \mathbf{x}_i^T\boldsymbol{\beta})v(\mathbf{x}_i)). \tag{IV.5}$$

To obtain the estimation consistency, we make the following assumptions on the residual function l.

Assumption IV.2 (Loss Function Assumptions)**.** *$l : \mathbb{R} \mapsto \mathbb{R}$ is a scalar function with the existence of the first derivative l' everywhere and the second derivative l'' almost everywhere. In addition,*

(i) there exists a constant $0 < k_1 < \infty$ such that $|l'(u)| \leq k_1$ for all $u \in \mathbb{R}$.

(ii) l' is Lipschitz such that $|l'(x) - l'(y)| \leq k_2|x - y|$, for all $x, y \in \mathbb{R}$ and some $0 < k_2 < \infty$.

Note that Assumption IV.2(i) requires bounded derivative of the loss function, which can limit the effect of large residuals and thus achieve certain robustness. Assumption IV.2(ii) indicates that $|l''(u)| < k_2$ for all $u \in \mathbb{R}$ where $l''(u)$ exists. The above assumptions actually cover a wide range of loss functions, including Huber loss, Hampel loss, Tukey's biweight and Cauchy loss.

We now make some assumptions on both random error ϵ and covariate vector $\mathbf{x}$.

Assumption IV.3 (Error and Covariate Assumptions)**.** *For $w(\mathbf{x})$ and $v(\mathbf{x})$ given in* (III.8)*, the random error ϵ with $E[\epsilon] = 0$ and covariate vector $\mathbf{x}$ with $E[\mathbf{x}] = \mathbf{0}$ satisfy:*

(i) for any $\boldsymbol{\nu} \in \mathbb{R}^p$, $w(\mathbf{x})\mathbf{x}^T\boldsymbol{\nu}$ *is sub-Gaussian with parameter at most* $k_0^2\|\boldsymbol{\nu}\|_2^2$.

(ii) either (a) $v(\mathbf{x}) = 1$ *and* $E[w(\mathbf{x})\mathbf{x}] = 0$, *or (b)* $E[l'(v(\mathbf{x})\epsilon)|\mathbf{x}] = 0$.

Note that Assumption IV.3(i) and (ii)(a) hold when $\mathbf{x}_i^T\boldsymbol{\nu}$ is sub-Gaussian for any $\boldsymbol{\nu} \in \mathbb{R}$ and $w(\mathbf{x}) = 1$. If covariate $\mathbf{x}$ is contaminated or heavy-tailed, Assumption IV.3(i) nonetheless holds with some proper choices of $w(\mathbf{x})$ (e.g. $w(\mathbf{x})\mathbf{x}^T\boldsymbol{\nu}$ is bounded for any $\boldsymbol{\nu} \in \mathbb{R}$), which potentially relaxes the sub-Gaussian assumption on $\mathbf{x}$. Assumption IV.3(ii)(b) holds when l' is an odd function and ϵ follows a symmetric distribution. Despite the possible mild condition of symmetry, those assumptions above are independent of the distribution of ϵ, allowing the additive error ϵ to be heavy-tailed or contaminated.

In order to obtain the estimation consistency for $\hat{\boldsymbol{\beta}}$ in (IV.4), we also require the loss function $\mathcal{L}_n$ to satisfy the following local Restricted Strong Convexity (RSC) condition. This RSC condition was also investigated in [LW13] and [Loh17].

Assumption IV.4 (RSC condition)**.** *There exist* γ, $\tau > 0$ *and a radius* $r > 0$ *such that the loss function* $\mathcal{L}_n$ *in* (III.8) *satisfies*

$$\langle \nabla\mathcal{L}_n(\boldsymbol{\beta}_1) - \nabla\mathcal{L}_n(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle \geq \gamma\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 - \tau\frac{\log p}{n}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1^2, \tag{IV.6}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ *such that* $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*\|_2 \leq r$ *for* $j = 1, 2$.

Note that the RSC assumption is only imposed on $\mathcal{L}_n$ inside the ball of radius r centered at $\boldsymbol{\beta}^*$. Thus the loss function used for robust regression can be wildly nonconvex while it is away from the origin. The ball of radius r essentially specifies a local region around $\boldsymbol{\beta}^*$ in which stationary points of program (IV.4) are well-behaved. We call such region as the RSC region.

We present the estimation consistency result concerning estimator $\hat{\boldsymbol{\beta}}$ in the following Theorem IV.1.

Theorem IV.1. *Suppose the random error and covariates satisfy Assumption IV.3 and $\mathcal{L}_n$ in* (III.8) *satisfies Assumption IV.2. Then we have the following results.*

(i) It holds with probability at least $1 - C_1 \exp(-C_2 \log p)$ that $\mathcal{L}_n$ satisfies

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \leq C_0 k_0 k_1 \sqrt{\frac{\log p}{n}}. \tag{IV.7}$$

(ii) Suppose L_n satisfies the RSC condition in Assumption IV.4 with $\boldsymbol{\beta}_2 = \boldsymbol{\beta}^*$ and ρ is μ-amenable with $\frac{3}{4}\mu < \gamma$ in Assumption IV.1. Let $\hat{\boldsymbol{\beta}}$ be a local estimator in (IV.4) in the RSC region. Then for $n \geq Cr^{-2} d_a s \log p$, $R \geq \|\boldsymbol{\beta}^*\|_1$ and $\lambda \geq \max\{4\|\nabla L_n(\boldsymbol{\beta}^*)\|_\infty, 8\tau R \frac{\log p}{n}\}$, $\hat{\boldsymbol{\beta}}$ exists and satisfies the bounds

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{6\sqrt{d_a}\lambda\sqrt{s}}{4\gamma - 3\mu} \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{6(1 + 3g(d))d_a\lambda s}{4\gamma - 3\mu}.$$

The statistical consistency result of Theorem IV.1 holds even though the random errors are heavy-tailed and contaminated, and the regressors lack of the sub-Gaussian assumption. Theorem IV.1(ii) essentially gives general deterministic bounds of the estimation error, provided that the loss function $\mathcal{L}_n$ satisfies the RSC condition and the penalty function ρ is μ-amenable. In particular, Theorem IV.1 shows that with high probability one can choose $\lambda = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{d_a s \log p}{n}}\right)$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_p\left(g(d) d_a s \sqrt{\frac{\log p}{n}}\right)$. Hence if d_a is finite, the estimator $\hat{\boldsymbol{\beta}}$ at the Group Penalization Stage is statistically consistent at the minimax rate enjoyed by the LS-GLasso under the sub-Gaussian assumption.

Remark. Recall that $\tilde{\boldsymbol{\beta}}$ is a stationary point of the optimization in (IV.4) if

$$\langle \nabla \mathcal{L}_n(\tilde{\boldsymbol{\beta}}) + \nabla \rho_\lambda(\tilde{\boldsymbol{\beta}}), \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \rangle \geq 0,$$

for all feasible $\boldsymbol{\beta}$ in a neighbour of $\tilde{\boldsymbol{\beta}}$, where $\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{J} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda)$. The stationary points include the interior local maxima as well as all local and global minima. The proof in Appendix reveals that the estimation consistency result also holds for the stationary points in program (IV.4). Hence Theorem IV.1 guarantees that all stationary points within the ball of radius r centered at $\boldsymbol{\beta}^*$ have local statistical consistency at the minimax rate enjoyed by the LS-GLasso. To simplify the notation, $\hat{\boldsymbol{\beta}}$ also denotes the stationary points of program (IV.4).

Next we establish the group-level oracle properties of estimator $\hat{\boldsymbol{\beta}}$ in (IV.4). Suppose I_S is given in advance,we define the group-level local oracle estimator as

$$\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{I_S} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r}{\operatorname{argmin}} \{\mathcal{L}_n(\boldsymbol{\beta})\}. \tag{IV.8}$$

Let $\hat{\boldsymbol{\beta}}^{\mathcal{O}} := (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}, \mathbf{0}_{I_S^c})$. The next theorem shows that when the penalty ρ is (μ, δ)-amenable and conditions in Theorem IV.1 are satisfied, the stationary point from (IV.4) within the local neighborhood of $\boldsymbol{\beta}^*$ is actually unique and agree with the group oracle estimator in (IV.8).

Theorem IV.2. *Suppose the penalty ρ is (μ, δ)-amenable and conditions in Theorem IV.1 hold. Suppose in addition that $v(\mathbf{x})\mathbf{x}_j$ is sub-Gaussian for all $j = 1, \cdots, p$, $\|\boldsymbol{\beta}^*\|_1 \leq \frac{R}{2}$ for some $R > \frac{12(1+3g(d))d_a\lambda s}{4\gamma - 3\mu}$, $\boldsymbol{\beta}^{*G}_{\min} \geq C_3\sqrt{\frac{k \log k}{n}} + \sqrt{d_a}\delta\lambda$, $n \geq C_0 k \log p$ and $k^2 \log k = \mathcal{O}(\log p)$. Let $\hat{\boldsymbol{\beta}}$ be a stationary point of the program in the RSC region. Then with probability at least $1 - C_5 \exp(-C_2 \log k)$, $\hat{\boldsymbol{\beta}}$ satisfies supp$(\hat{\boldsymbol{\beta}}) \subseteq I_S$ and $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$.*

Theorem IV.2 guarantees that the Group Penalization Stage in our proposed framework can recover the true group support with high probability, when the condition of minimum group signal strength is satisfied. Two most common (μ, δ)-amenable penalties are SCAD and MCP, as introduced in Section IV.2.

It has been shown that the GP Stage can select important covariates groups and provides consistent estimation for parameter $\boldsymbol{\beta}^*$. We are now ready to establish statistical properties of $\hat{\boldsymbol{\beta}}^h$ after the HT stage in our proposed framework. We reveal in the following theorem that when the condition of minimum individual signal strength is satisfied, the estimate of the zero elements and the non-zero elements of $\boldsymbol{\beta}^*$ after the GP Stage can then be well separated. Hence, there exists some thresholds that are able to filter out those non-important covariates within the selected important groups, and thus the HT Stage can perform bi-level variable selection consistently.

Theorem IV.3. *Suppose conditions of Theorem IV.2 hold and in addition that* $\boldsymbol{\beta}^{*I}_{\min} \geq C_3\sqrt{\frac{k\log k}{n}} + \theta$ *and* $\theta > C_3\sqrt{\frac{k\log k}{n}}$. *With probability at least* $1 - C_5\exp(-C_2\log k)$, *the hard-thresholding estimator* $\hat{\boldsymbol{\beta}}^h(\theta)$ *given in (IV.2) satisfies* $\hat{\boldsymbol{\beta}}^h = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_0}, \mathbf{0}_{I_0^c})$ *and* $\|\hat{\boldsymbol{\beta}}^h - \boldsymbol{\beta}^*\|_2 \leq C_3\sqrt{\frac{k\log k}{n}}$.

Theorem IV.3 guarantees that the estimator $\hat{\boldsymbol{\beta}}^h$ in our proposed two-stage framework possesses estimation consistency and bi-level variable selection consistency, when conditions of Theorem IV.2 hold and the condition of minimum individual signal strength is satisfied. Note that such signal strength condition is fairly mild and the bound can decrease arbitrarily closed to 0 as the growth of sample size n.

IV.4. Implementation

We discuss the implementation of the proposed two-stage M-estimator framework in this section, including finding a stationary point in program (IV.4) for a fixed λ and the tuning parameters selection for both λ and θ.

Note that the optimization in (IV.4) may not be a convex optimization problem since we allow both loss function $\mathcal{L}_n$ and ρ to be non-convex. To obtain the corresponding stationary point, we use composite gradient descend algorithm [Nes13]. Recall $q_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^J \sqrt{d_j}\lambda\|\boldsymbol{\beta}_j\|_2 - \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda)$ and $\bar{L}_{\alpha,n}(\boldsymbol{\beta}) = \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) - q_\lambda(\boldsymbol{\beta})$, we can rewrite the program as

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}\|_1 \le R}{\operatorname{argmin}} \left\{ \bar{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^J \sqrt{d_j}\lambda\|\boldsymbol{\beta}_j\|_2 \right\}.$$

Then the composition gradient iteration is given by

$$\boldsymbol{\beta}^{t+1} \in \underset{\|\boldsymbol{\beta}\|_1 \le R}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - (\boldsymbol{\beta}^t - \frac{\nabla \bar{L}_n(\boldsymbol{\beta}^t)}{\eta})\|_2^2 + \sum_{j=1}^J \lambda\eta\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \right\}, \tag{IV.9}$$

where $\eta > 0$ is the step size for the update and can be determined by the backtracking line search method described in [Nes13]. A simple calculation shows that the iteration in (IV.9) takes the form

$$\boldsymbol{\beta}_j^{t+1} = S_{\lambda\eta\sqrt{d_j}}\left(\left(\boldsymbol{\beta}^t - \eta\nabla\bar{L}_n(\boldsymbol{\beta}^t)\right)_j \right),$$

where $S_{\sqrt{d_j}\lambda\eta}(\cdot)$ is the group soft-thresholding operator defined as

$$S_\delta(\boldsymbol{z}) := \left(1 - \frac{\delta}{\|\boldsymbol{z}\|_2}\right)_+ \boldsymbol{z}.$$

We adopt the following two-step procedure discussed in [Loh17] to guarantee the convergence to a stationary point for the non-convex optimization problem in (IV.4).

Step 1: Run the composite gradient descent using a Huber loss function with convex group Lasso penalty to get an initial estimator.

Step 2: Run the composite gradient descent on the program (IV.4) at the Group Penalization Stage using the initial estimator from Step 1.

As to the tuning parameters selection, the optimal values of tuning parameters λ and θ are chosen from a two-dimensional grid search using the cross-validation. In particular, the searching grid is formed by partitioning a rectangle uniformly in the scale of $(\theta, \log(\lambda))$. Motivated by conditions of Theorem IV.1 and Theorem IV.3, the range of the rectangle can be chosen as $C_{11}\sqrt{\frac{\log p}{n}} \leq \lambda \leq C_{12}\sqrt{\frac{\log p}{n}}$ and $C_{21}\sqrt{\frac{k \log k}{n}} < \theta \leq C_{22}$. The optimal values are then found by the combination that minimizes the cross-validated trimmed mean squared prediction error.

IV.5. Simulation Studies

In this section, we assess the performance of our two-stage M-estimator framework by considering different types of loss functions and penalty functions through various models. The data is generated from the following model

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

The covariates vector $\mathbf{x}_i$s are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$ independently. For covariance $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, we choose

$$\sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ (-1)^{i+j}a & \text{if } i \neq j \text{ and } i, j \text{ are in the same group,} \\ (-1)^{i+j}ab & \text{if } i \neq j \text{ and } i, j \text{ are in different groups,} \end{cases}$$

where $a = 0.8$ or 0.5 and $b = 0.8$ or 0.5. Let $\boldsymbol{\beta}^* = \boldsymbol{\phi} \cdot |\boldsymbol{\beta}^*|$, where $\boldsymbol{\phi}$ is a p-dimensional vector with the jth element being $(-1)^{j+1}$.

Example IV.1. (Group-level Sparsity) The number of observations $n = 100$ and the number of variables $p = 500$ with $J = 100$ unequal-size groups. We choose $a = 0.8$ and $b = 0.5$. The model includes only between-group sparsity with five relevant groups,

$$|\boldsymbol{\beta}_1^*| = |\boldsymbol{\beta}_2^*| = (\underbrace{3, \cdots, 3}_{4})^T = \mathbf{3_4^T}, \quad |\boldsymbol{\beta}_\mathbf{3}^*| = |\boldsymbol{\beta}_\mathbf{4}^*| = \mathbf{2_6}, \quad |\boldsymbol{\beta}_\mathbf{5}^*| = \mathbf{1.5_5},$$
$$\boldsymbol{\beta}_\mathbf{6}^* = \cdots = \boldsymbol{\beta}_\mathbf{100}^* = \mathbf{0_5}.$$

We generate random error ϵ_i from the following 3 scenarios: (a) $N(0, 1)$, (b) t_1, (c) Mix Cauchy (70% are from $N(0, 1)$ and 30% are from standard Cauchy).

We consider bi-level penalized M-estimators with different types of loss functions (the ℓ_2 loss, Huber loss, Cauchy loss) and two types of penalty functions (the Lasso and MCP penalties). In particular, we evaluate the performance of non-group estimators, one-stage estimators and two-stage estimators. Without causing any confusion, let $\hat{\boldsymbol{\beta}}$ be any estimator of $\boldsymbol{\beta}^*$. Its performances on both parameter estimation and group/variable selection were evaluated by the following eight measurements:

(1) ℓ_2 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

(2) ℓ_1 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$.

(3) Model size (MS), the average number of selected covariates.

(4) Group size (GS), the average number of selected groups.

(5) False positives rate for individual variable selection (FPR), the percent of selected covariates which are actually unimportant variables.

(6) False negatives rate for individual variable selection (FNR), the percent of non-selected covariates which are actually important variables.

(7) False positives rate for group variable selection (GFPR), the percent of selected groups which are actually unimportant groups.

(8) False negatives rate for group variable selection (GFNR), the percent of non-selected groups which are actually important groups.

Note that FPR $= \frac{|\hat{I} \bigcap I_0^c|}{|I_0^c|} \times 100\%$, FNR $= \frac{|\hat{I}^c \bigcap I_0|}{|I_0|} \times 100\%$, GFPR $= \frac{|\hat{S} \bigcap S^c|}{|S^c|} \times 100\%$ and GFNR $= \frac{|\hat{S}^c \bigcap S|}{|S|} \times 100\%$, where $\hat{I} = \{m : \hat{\beta}_m \neq 0,\ 1 \leq m \leq p\}$, $I_0 = \{m : \beta_m^* \neq 0,\ 1 \leq m \leq p\}$, $\hat{S} = \{j : \hat{\boldsymbol{\beta}}_j \neq \mathbf{0},\ 1 \leq j \leq J\}$ and $S = \{j : \boldsymbol{\beta}_j^* \neq \mathbf{0},\ 1 \leq j \leq J\}$.

The model considered in Example IV.1 contains only the between-group sparsity. We also assess the performance of the two-stage M-estimator framework under models with bi-level sparsity in the following example.

Example IV.2. (Bi-level Sparsity) The number of observations $n = 100$ and we generate the random error ϵ following the same three scenarios described in Example IV.1.

(i) The number of variables $p = 500$ with $J = 100$ unequal-size groups. We choose $a = 0.8$ and $b = 0.5$. The model includes within-group sparsity among six relevant groups,

$$|\boldsymbol{\beta}_1^*| = (1.5, 2, 0, 2.5)^T, \quad |\boldsymbol{\beta}_2^*| = (3, 2, 0, 0, 2)^T, \quad |\boldsymbol{\beta}_3^*| = (1.5, 0, 2.5, 3, 0, 0)^T,$$
$$|\boldsymbol{\beta}_4^*| = (2, 1.5, \underbrace{0, \cdots, 0}_{4})^T, \quad |\boldsymbol{\beta}_5^*| = (2.5, 0, 0, 0)^T, \quad |\boldsymbol{\beta}_6^*| = (3, 2.5, 2.5, 2, 1.5)^T,$$
$$\boldsymbol{\beta}_7^* = \cdots = \boldsymbol{\beta}_{100}^* = (\underbrace{0, \cdots, 0}_{5})^T.$$

(ii) Similar to (i) except that we choose $a = 0.5$ and $b = 0.8$.

(iii) The number of variables $p = 1000$ with $J = 100$ unequal-size groups. We choose $a = 0.8$ and $b = 0.5$. The model includes within-group sparsity in among four relevant groups,

$$|\boldsymbol{\beta}_1^*| = (3, 2, 0, 0, 0)^T, \quad |\boldsymbol{\beta}_2^*| = (1.5, 2, 2.5, 2.5, 3, \underbrace{0, \cdots, 0}_{5})^T,$$

$$|\boldsymbol{\beta}_3^*| = (1.5, 0, 2.5, 3, 0, 3, 2, 1.5, \underbrace{0, \cdots, 0}_{7})^T,$$

$$|\boldsymbol{\beta}_4^*| = (3, 3, 2.5, 2.5, 2, 2, 1.5, 1.5, 1.5, 1.5),$$

$$|\boldsymbol{\beta}_4^*| = \cdots = \boldsymbol{\beta}_{100}^* = (\underbrace{0, \cdots, 0}_{10})^T.$$

Finally, we design a simulation setting to evaluate the performance of the two-stage M-estimator framework when covariates are contaminated or not sub-Gaussian.

Example IV.3. (Contamination on x) All the settings are similar to example IV.2(i), except that we let $n = 120$ and covariates be partially contaminated after the data generation. In particular, 20% of the observations in $\mathbf{x}$ are replaced by data generated from $\chi^2(10)$ first, and then recentered to have mean zero.

We ran 100 simulations for each scenario described in Example IV.1-IV.3. While fixing $v(\mathbf{x}) \equiv w(\mathbf{x}) \equiv 1$ for Example IV.1 and IV.2, we consider the general two-stage M-estimator framework with $v(\mathbf{x}) \equiv 1$ and $w(\mathbf{x}) = \min\left\{1, \frac{4}{\|\mathbf{x}\|_\infty}\right\}$ in Example IV.3. As introduced in Section IV.4, we choose two tuning parameters λ and θ optimally with 10-fold cross-validation, with λ ranging in $(0.01\sqrt{\frac{\log p}{n}}, 10\sqrt{\frac{\log p}{n}})$ and θ ranging in $(0.01\sqrt{\frac{k \log k}{n}}, 0.5)$. The results from Example IV.1 to IV.3 are reported in Table IV.1 to IV.3, respectively. Note that we consider the one-stage estimators with the Lasso penalty as the GLasso-type estimators. For the MCP penalty, we call the corresponding non-group estimators, one-stage estimators and two-stage estimators the MCP-type, GMCP-type and GMCP-HT-type estimators, respectively.

Table IV.1. Simulation Results under the Model with Only Between-group Sparsity in Example IV.1. The mean ℓ_2 error, ℓ_1 error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		Group Lasso			Group MCP		
		LS	Huber	Cauchy	LS	Huber	Cauchy
N(0,1)	ℓ_2 error	1.27 (0.4)	1.29 (0.93)	1.39 (1.48)	0.92 (0.21)	0.93 (0.21)	0.95 (0.2)
	ℓ_1 error	6.3 (2.5)	6.59 (5.58)	6.93 (7.32)	3.75 (0.84)	3.77 (0.85)	3.85 (0.83)
	MS	55.9 (24.31)	66.21 (27.31)	66.01 (35.64)	31.43 (11.75)	33.23 (13.72)	33.09 (16.05)
	GS	11.18 (4.86)	13.24 (5.46)	13.2 (7.13)	6.29 (2.36)	6.65 (2.75)	6.62 (3.22)
	FP, FN	6.51, 0	8.69, 0.36	8.73, 1.76	1.35, 0	1.73, 0	1.7, 0
	GFP, GFN	6.51, 0	8.69, 0.4	8.73, 1.8	1.36, 0	1.74, 0	1.71, 0
t_1	ℓ_2 error	13.77 (42.15)	2.02 (1.65)	1.82 (1.47)	24.96 (53.14)	2.72 (0.85)	2.46 (2.2)
	ℓ_1 error	166.82 (711.19)	11.04 (9.68)	9.59 (7.25)	243.53 (838.07)	10.88 (3.43)	10.22 (11.69)
	MS	114.89 (78.82)	71.75 (17.21)	70.12 (19.47)	65.64 (72.87)	27.9 (9.95)	29.15 (10.18)
	GS	23 (15.78)	14.35 (3.44)	14.03 (3.89)	13.16 (14.6)	5.58 (1.99)	5.83 (2.04)
	FP, FN	19.26, 6.32	9.94, 1.8	9.59, 1.68	9.11, 10.6	0.61, 0	0.87, 0
	GFP, GFN	19.26, 6	9.94, 1.8	9.59, 1.6	9.12, 10	0.61, 0	0.87, 0
Mix Cauchy	ℓ_2 error	12.84 (64.77)	1.42 (0.87)	1.36 (1.08)	16.92 (70.15)	1.46 (0.34)	1.36 (0.34)
	ℓ_1 error	178.11 (1045.4)	7.44 (5.31)	6.92 (5.73)	225.05 (1227.29)	5.82 (1.41)	5.48 (1.44)
	MS	94.6 (84.99)	72.11 (20.8)	71.25 (27.59)	51.99 (86.46)	27.2 (8.66)	29.45 (10.32)
	GS	18.92 (16.99)	14.42 (4.16)	14.26 (5.53)	10.4 (17.29)	5.44 (1.73)	5.89 (2.06)
	FP, FN	14.75, 1.8	9.94, 0.36	9.79, 1	5.84, 3.04	0.46, 0	0.94, 0
	GFP, GFN	14.75, 1.8	9.94, 0.4	9.8, 1	5.84, 3	0.46, 0	0.94, 0

We mainly evaluate the performance of one-stage estimators in Example IV.1 since there only exists the between-group sparsity. Table IV.1 shows that with the same loss function, while the GMCP-type estimators perform comparably to the GLasso-type estimators in estimation, the former have better group/variable selection accuracy than the latter. This is consistent with the group oracle property stated in Theorem IV.2. As expected, for the estimators with the same penalty, while they behave similarly in the light-tail setting ($N(0,1)$), estimators using Huber loss and Cauchy loss largely outperform the least squares estimator for the heavy-tailed settings (t_1 and Mix Cauchy).

We compare the results of non-group estimators, one-stage estimators and two-stage estimators for Example IV.2. Note that here we only consider the MCP penalty since it has been shown to perform better than the Lasso penalty. For Example IV.2(i), Table IV.2 shows that the GMCP-type estimators outperform the MCP-type estimators in all measurements, since the former incorporates the grouping structure in

$\mathbf{x}$. By comparing the results of GMCP-type estimators and GMCP-HT-type estimators, we see that the extra hard-thresholding step in the two-stage estimators can effectively improve the estimation and group/variable selection performance. Similar to Example IV.1, the robust estimators given by the Huber loss and the Cauchy loss have more advantageous than the least squares estimators in heavy-tailed settings. In addition, estimators using the Cauchy loss further outperform the one with Huber Loss for the heavy-tailed settings, showing that the redescending estimators are more robust to outliers and more efficient for irregular settings. We observe similar patterns in the results of Example IV.2(ii)-(iii) and thus we omit those results in this chapter.

Table IV.2. Simulation Results under the Model with Bi-level Sparsity in Example IV.2.1. The mean ℓ_2 error, ℓ_1 error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		MCP			GMCP			GMCP-HT		
		LS	Huber	Cauchy	LS	Huber	Cauchy	LS	Huber	Cauchy
N(0,1)	ℓ_2 error	7.2 (2.34)	6.98 (2.42)	7.45 (2.36)	1.68 (0.35)	1.67 (0.34)	1.66 (0.32)	1.59 (0.36)	1.57 (0.36)	1.6 (0.35)
	ℓ_1 error	29.95 (11.26)	29.3 (11.74)	31.52 (11.9)	7.52 (1.59)	7.5 (1.57)	7.49 (1.5)	6.91 (1.99)	6.79 (2)	6.9 (2.03)
	MS	25.29 (15.92)	28.77 (19.08)	27.56 (22.65)	52.53 (14.73)	53.89 (15.05)	53.55 (15.78)	30.49 (13.87)	30.49 (14.99)	30.54 (15.26)
	GS	16.72 (8.7)	18.75 (10.22)	18 (11.01)	10.51 (2.95)	10.79 (3.03)	10.71 (3.16)	8.11 (3.83)	8.27 (3.99)	8.21 (4.37)
	FP, FN	2.88, 32.94	3.54, 31.41	3.37, 33.53	7.36, 0	7.64, 0	7.57, 0	2.79, 0	2.79, 0	2.8, 0
	GFP, GFN	11.51, 1.67	13.66, 1.5	12.8, 0.5	4.8, 0	5.1, 0	5.01, 0	2.24, 0	2.41, 0	2.35, 0
t_1	ℓ_2 error	33.42 (50.3)	11.31 (1.92)	11.5 (1.68)	25.2 (51.17)	4.37 (0.83)	3.75 (1.04)	25.07 (51.45)	4.34 (0.89)	3.68 (1.02)
	ℓ_1 error	262.31 (807.06)	46.5 (6.98)	47.37 (6.96)	244.67 (831.39)	19.45 (3.69)	16.68 (4.87)	243.1 (832.36)	19.28 (4.34)	16.08 (5.28)
	MS	24.74 (53.71)	12.61 (10.39)	9.85 (4.62)	79.86 (75.4)	52.57 (16.04)	47.53 (12.6)	65.81 (68.74)	34.7 (12.97)	31.44 (12.39)
	GS	17.31 (19.54)	9.99 (6.17)	8.51 (3.42)	15.93 (15.09)	10.51 (3.21)	9.5 (2.52)	14.04 (14.66)	8.28 (3.2)	7.74 (3)
	FP, FN	4.34, 77.82	1.28, 62.29	0.83, 65.53	13.81, 22.53	7.37, 0.18	6.33, 0.18	10.98, 24.82	3.72, 1.59	3.03, 1.12
	GFP, GFN	14.2, 34	4.94, 10.83	3.4, 11.5	12.37, 28.33	4.83, 0.5	3.76, 0.5	10.49, 30.33	2.51, 1.33	1.9, 0.83
Mix Cauchy	ℓ_2 error	25.18 (69.7)	8.9 (2.12)	8.91 (2.08)	18.6 (70.35)	2.47 (0.61)	2.11 (0.52)	18.18 (70.17)	2.39 (0.61)	2.03 (0.5)
	ℓ_1 error	248.19 (1192.94)	37.94 (10.19)	38.24 (10.52)	234.97 (1234.62)	11.14 (2.81)	9.42 (2.3)	231.01 (1233.91)	10.29 (3.24)	8.62 (2.67)
	MS	26.93 (53.54)	18.87 (9.61)	18.91 (12.08)	71.1 (78.13)	47.4 (14.24)	48.94 (15.85)	52.56 (74.75)	29.24 (13.63)	29.06 (12.7)
	GS	16.87 (18.63)	13.84 (6.5)	13.97 (7.42)	14.2 (15.61)	9.48 (2.85)	9.79 (3.17)	11.88 (15.54)	7.24 (2.92)	7.42 (2.8)
	FP, FN	4.08, 57.59	2.03, 46.65	2.08, 48	11.44, 6.71	6.29, 0	6.61, 0	7.63, 7.65	2.54, 0.06	2.5, 0
	GFP, GFN	12.37, 12.67	8.48, 2.17	8.59, 1.67	9.29, 8.83	3.7, 0	4.03, 0	6.83, 9	1.32, 0	1.51, 0

In Example IV.3 we only compare the performance of two-stage estimators with their weighted version. Table IV.3 indicates that the two-stage estimators with well chosen $w(\mathbf{x})$ perform better in all cases than the two-stage estimators with $w(\mathbf{x}) = 1$. Again when the errors are heavy-tailed (t_1 and Mix Cauchy), the least

squares estimator lose its efficiency and the redescending estimators produced by Cauchy loss perform the best for all scenarios.

Table IV.3. Simulation Results under the Model with 20% Contamination on X in Example IV.3. The mean ℓ_2 error, ℓ_1 error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed. Standard errors are listed in parentheses.

		GMCP - HT			WGMCP - HT		
		LS	Huber	Cauchy	LS	Huber	Cauchy
N(0,1)	ℓ_2 error	7.49 (0.77)	7.52 (0.87)	7.54 (1.01)	6.74 (0.83)	6 (1.1)	4.97 (1.4)
	ℓ_1 error	43.56 (6.15)	42.78 (6.51)	40.76 (8.13)	35.75 (5.95)	28.81 (7.41)	22.42 (7.18)
	MS	71.76 (24.18)	66.93 (24.61)	54.81 (23.15)	60.82 (20.98)	38.22 (16.6)	32.53 (17.64)
	GS	17.34 (4.65)	16.26 (4.78)	13.74 (5.72)	14.02 (4.72)	9.38 (4.63)	9.72 (8.52)
	FP, FN	11.64, 8.59	10.63, 8.35	8.15, 9.06	9.28, 5.82	4.6, 5.88	3.38, 4.76
	GFP, GFN	12.72, 10.33	11.59, 10.5	8.9, 10.5	9.04, 8	3.98, 6	4.33, 5.83
t_1	ℓ_2 error	125.26 (992.64)	8.46 (1.16)	8.54 (1.36)	126.92 (998.84)	6.96 (1.54)	6.43 (1.35)
	ℓ_1 error	2081.98 (17983.9)	47.63 (7.94)	46.78 (10.86)	2099.27 (18099.01)	33.87 (10.92)	30.84 (10.21)
	MS	96.16 (87.48)	61.26 (25.54)	52.46 (24.03)	86.38 (88.52)	37.42 (14.34)	35 (16.84)
	GS	22.76 (17.48)	15.2 (5.56)	14.12 (7.74)	19.26 (18.01)	9.07 (4.3)	9.56 (7.71)
	FP, FN	17.39, 28.29	9.71, 15.65	7.91, 16.12	15.35, 27.94	4.57, 9.71	4.05, 9.18
	GFP, GFN	19.76, 30.17	10.93, 17.83	9.77, 17.67	16.06, 30.67	4.02, 11.83	4.56, 12.17
Mix Cauchy	ℓ_2 error	18.52 (81.12)	7.72 (1.13)	7.66 (1.39)	18.48 (83.42)	5.97 (1.47)	5.22 (1.56)
	ℓ_1 error	211.8 (1454.22)	43.23 (6.98)	39.92 (9.08)	210.62 (1492.36)	28.68 (8.54)	24.67 (9.68)
	MS	75.62 (51.22)	64.31 (25.1)	48.28 (21.8)	63.31 (48.76)	37.8 (15.33)	35.46 (20.97)
	GS	18.17 (10.38)	15.69 (5.45)	12.39 (5.7)	13.98 (9.68)	9.41 (4.95)	10.16 (8.39)
	FP, FN	12.69, 15.76	10.19, 11.18	6.87, 11.29	10.02, 12.29	4.52, 6	4, 4.94
	GFP, GFN	14.09, 17.83	11.16, 13.33	7.59, 12.33	9.41, 14.5	4.09, 7.17	4.74, 5

In summary, our simulation studies show that in the proposed two-stage M-estimator framework, (1) the GP Stage can utilize the grouping structure to yield satisfactory parameter estimation and group variable selection results for irregular settings, if a robust loss function (e.g. Huber loss and Cauchy loss) is used; (2) the HT Stage further improve the performance by filtering out the non-important selected variable from the first stage; (3) the two-stage M-estimators with redescending loss functions (e.g. Cauchy loss) and concave penalties consistently render more satisfactory results when data are heavy-tailed or strongly contaminated (t_1 and Mix Cauchy).

IV.6. Real Data Example

In this section, we consider the NCI-60 data introduced in Section I.1. We first perform some pre-screenings by keeping only 2000 genes with largest variations and choosing 500 genes out of which are most correlated with the response variable. Then for each gene, we use B-spline with 5 bases to form a group with 5 variables. Thus our final data set has $n = 59$ samples, $p = 2500$ variables and $J = 500$ groups. Similar to our simulation studies, we apply the non-group estimators, one-stage estimators and two-stage estimators to select important genes, with tuning parameter λ and θ chosen from the 10-folded cross validation with λ ranging in $(0.01\sqrt{\frac{\log p}{n}}, 10\sqrt{\frac{\log p}{n}})$ and θ ranging in $(0.01, 1)$. In particular, we report results from six methods: Huber-MCP, Cauchy-MCP, Huber-GMCP, Cauchy-GMCP, Huber-GMCP-HT, Cauchy-GMCP-HT.

The QQ-plots of the residuals generated from those six methods are shown in Figure IV.1. It shows that each residual distribution has a longer tail on the left side, meaning that the data may be contaminated or heavy-tailed. Table IV.4 displays the important genes selected by those six methods. It shows that the number of selected genes from those methods are 5 (Huber-MCP), 8 (Huber-GMCP), 8 (Huber-GMCP-HT), 5 (Cauchy-MCP), 11 (Cauchy-GMCP) and 14 (Cauchy-GMCP-HT), respectively. It implies that the methods incorporating grouping information can potentially select more genes. Notice that the Huber-MCP and Cauchy-MCP both select the same genes, which indicates that the contamination in the data may not be strong enough to cause different selection results between these two loss functions. In addition, it is reasonable to observe that the Huber-GMCP and Huber-GMCP-HT also select exactly the same genes, since there is no sparsity within each group in the data. However, the genes found by the Cauchy-GMCP are somewhat different from those selected by the Cauchy-GMCP-HT. Such difference is possibly due to

unstable solutions induced by the concavity of Cauchy loss. For further investigation, we randomly choose 6 observations as the test set and applied those six methods to the rest patients to get the coefficients estimation, then compute the prediction error on the test set. We repeat the random splitting 100 times and the boxplots of the Mean Absolute Error of predictions are shown in Figure IV.2. It is clearly observed from Figure IV.2 that the Huber-GMCP and Cauchy-GMCP perform better than the other methods. This is not surprising since there is only between-group sparsity in the dataset. In addition, Figure IV.2 also shows that Cauchy-type estimators perform similarly to the corresponding Huber-type estimators, which indicates that when there exist only moderate contamination in the data, it may be sufficient to consider the convex Huber loss in our framework.

Table IV.4. Selected Genes by Huber-MCP, Cauchy-MCP, Huber-GMCP, Cauchy-GMCP, Huber-GMCP-HT, Cauchy-GMCP-HT

Huber-MCP	KRT8	NRN1	GAS7	EPS8L2	GPX3			
Huber-GMCP	KRT8	ANXA3	KRT19	DSP	GPX3	LEF1	TDRD7	SRPX
Huber-GMCP-HT	KRT8	ANXA3	KRT19	DSP	GPX3	LEF1	TDRD7	SRPX
Cauchy-MCP	KRT8	NRN1	GAS7	EPS8L2	GPX3			
Cauchy-GMCP	KRT8	ANXA3	KRT19	GPX3	LEF1	TDRD7	MITF	NOTCH3
	FAR2	INHBB	SIRPA					
Cauchy-GMCP-HT	KRT8	NRN1	AP1M2	ANXA3	GAS7	KRT19	EPS8L2	GPX3
	SNAI2	SPINT2	EPCAM	SFN	SLC29A2	NMU		

Figure IV.1. QQ Plots of the Residuals from Huber-MCP, Cauchy-MCP, Huber-GMCP, Cauchy-GMCP, Huber-GMCP-HT, Cauchy-GMCP-HT.

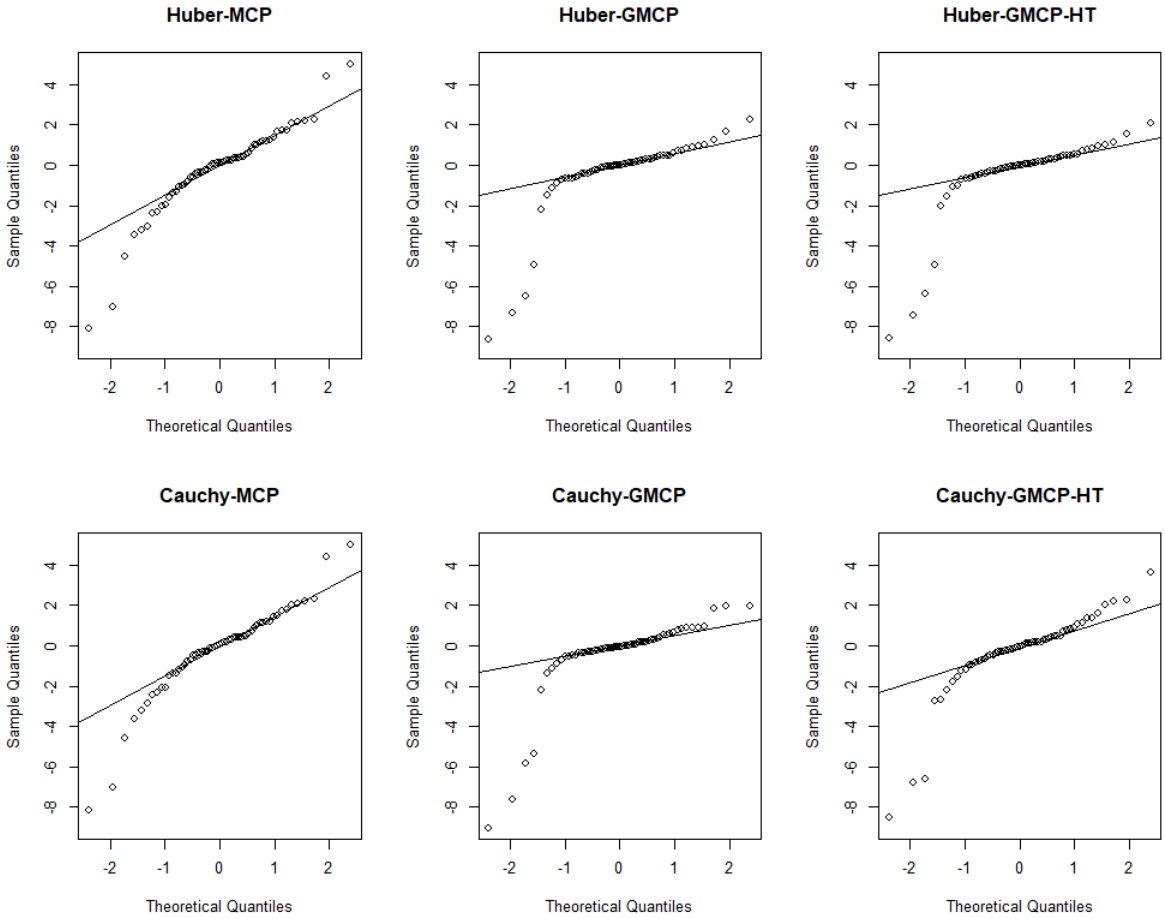
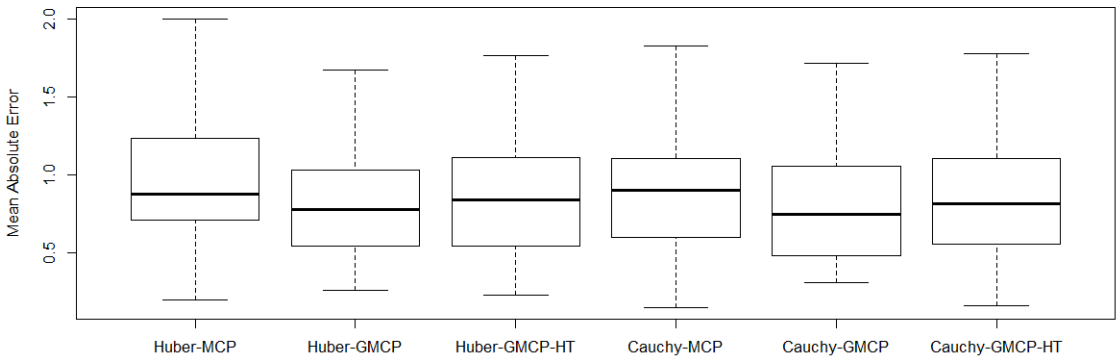


Figure IV.2. Boxplot of the Mean Absolute Error of Predictions



CHAPTER V

DISCUSSION AND FUTURE WORK

V.1. On the Penalized Weighted Least Squares Method

This thesis studies the simultaneous variable selection, outlier detection and robust estimation using an efficient weight shrinkage rule in a penalized weighted least squares framework. This approach is attractive in terms of its computation efficiency in high-dimensional settings, its Bayesian understanding, and most importantly, its united link to a regularized robust M-estimation. The Bayesian understanding justifies the rationality of the proposed PAWLS method for both outlier detection and variable selection. The joint estimation of weight and coefficient vectors and its link to M-estimation justify both of the strong robustness and estimation efficiency of this PAWLS approach under a fixed design.

[BBEKY13] studied the choice of ρ function in high-dimensional M-estimation with $p < n$ when the error distribution is assumed to be known and the ρ function is convex. The link between a weight shrinkage rule and the M-estimation studied in this thesis provides another direction on how to choose a sparse M-estimation. In particular, we can choose some sparse M-estimation with strong robustness, for example, a redescending M estimate such that ρ is not convex. If prior information or a distribution on the individual weight is provided, we can build a weight shrinkage rule based upon the *priori*. This weight shrinkage rule will be used to find the corresponding M-estimation.

Another important contribution of this thesis is the theoretical investigation of this approach when $p \gg n$. The non-asymptotic inequalities of the joint estimation

for the regression coefficients and weight parameters has been investigated in this thesis. Such a theoretical understanding advocates the use of the PAWLS for robust estimation and outlier detection. This result may also be extended to the study of regularized M-estimation in high-dimensional settings. For example, [NYWR09] establishes consistency and convergence rates for regularized M-estimators under high-dimensional settings when the ρ function satisfies a restricted strong convexity (RSC) condition. Unfortunately, the RSC condition rules out a class of redescending M-estimation in high-dimensional data analysis. The study in this thesis provides a direction of theoretic investigation of any regularized M-estimation by linking it to a specific penalized weight least squares regression model.

Currently, I am also working on the theoretical properties of the adaptive PAWLS approach. In particular, I want to provide some conditions under which the adaptive PAWLS has some nice variable selection and outlier detection properties. There are several other relevant research questions not fully addressed in this thesis. For example, the robustness of regression can also be measured by the influence function. There have been some interests concerning influence functions for high-dimensional estimators [AM14,ÖCA15]. It would be interesting to investigate the influence function of the PAWLS in high-dimensional settings. Another important issue is appropriate choices for regularization parameters with respect to both the variable selection and outlier detection. Although the thesis provides a modified BIC for tuning parameter selection in our numerical studies, there is still lack of theoretical investigations on whether this approach provides us optimal tuning parameters generating well-behaved PAWLS estimators.

In this thesis the proposed penalized weighted approach is investigated only in high-dimensional linear regression models. However, the proposed penalized weight

shrinkage rule can be generalized to deal with different types of data, such as the count data, survival data and categorical data. On the other hand, to address the problem of high colinearities among covariates, a few methods such as elastic net [ZH05] and post selection shrinkage estimation [GAF17] are proposed in high-dimensional settings. It will also be interesting to achieve the outlier detection in those methods by extending these methods to our proposed penalized weight shrinkage framework.

V.2. On the Penalized Robust Approximated Quadratic M-estimators

The irregular settings including data asymmetry, heteroscedasticity and data contamination often exist due to the data high-dimensionality. It is very important to address these irregular settings both theoretically and numerically in high-dimensional data analysis. In this thesis we have proposed a class of PRAM estimators for robust high-dimensional mean regression. The key feature of the PRAM estimators is using a family of loss functions with flexible robustness and diverging parameters to approximate the mean function produced from the traditional quadratic loss. This approximation process can reduce the bias generated by data's irregularity in high-dimensional mean regression. The proposed framework is very general and it covers a wide range of loss functions and penalty functions, allowing both functions to be non-convex.

Theoretically, we established statistical properties of PRAM in ultra high-dimensional settings when p grows with n at an almost exponential rate. Specifically, we showed its local estimation consistency at the minimax rate enjoyed by the LS-Lasso and established the oracle properties of the PRAM estimators, including both selection consistency and asymptotic normality. The theoretical result is applicable for irregular settings, including the data are contaminated by outliers, random errors and/or covariates are heavy-tailed, and random errors lack of symmetry and/or homogeneity.

One fundamental difference between our proposed PRAM estimator and the common penalized M-estimator is that we require $\lim_{\alpha\to\infty} E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$ instead of $E[\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)] = \mathbf{0}$ for every $\alpha > 0$. To establish the estimation consistency and the oracle properties, the divergent rate of α plays a crucial role. In the presence of asymmetric and heavy-tailed/contaminated data, the PRAM estimators can reduce the bias efficiently (when α diverges) and enjoy robustness (when α diverges not too fast). The divergent rate of α stated in Theorem III.2 and Theorem III.3 actually show us how α should diverge with n, in order to obtain a robust sparse PRAM estimator in high-dimensional mean regression under general irregular settings.

Additionally, our numerical studies show satisfactory finite sample performances of the PRAM estimators under irregular settings, which is consistent with our theoretical findings. Among all the possible choices of PRAM estimators, our numerical results also suggest to implement a redescending PRAM estimator with a concave penalty such as the TA-MCP and the CA-MCP, using the HA-Lasso as an initial estimator, when the data are strongly heavy-tailed or contaminated.

Our research in this thesis provides a systematic study of penalized M-estimation in high-dimensional linear regression model. However, we may not have linearity between the response and predictors in practice. Next we plan to explore the PRAM estimator with certain nonparametric regression models (e.g. an additive model). Other possible future directions of research may include devising similar theoretical guarantees for estimators with grouping structures in the covariates, or study of high-dimensional models with varying coefficients (e.g. [FMD14]) under general irregular settings.

V.3. On the High-dimensional M-estimation for Bi-level Variable Selection

Bi-level variable selection and parameter estimation are crucial when covariates function group-wisely in high dimensional settings. It has become even more challenging when data are contaminated or heavy-tailed. In this thesis, we proposed a two-stage penalized M-estimator framework for high-dimensional bi-level variable selection. This framework consists of two stages: penalized M-estimation with a concave ℓ_2-norm penalty achieving the consistent group selection at the first stage, and a post-hard-thresholding operator to achieve the within-group sparsity at the second stage. The proposed framework is very general that it covers a wide range of loss functions and penalty functions, allowing both functions to be non-convex. Thus, if the data are strongly contaminated, either in covariates or random errors, we are still able to perform bi-level variable selection efficiently through the proposed framework.

Theoretically, we established statistical properties of the proposed two-stage penalized M-estimator in ultra high-dimensional settings when p grows with n at an almost exponential rate. In particular, for the estimator at the Group Penalization Stage, we showed its local estimation consistency at the minimax rate enjoyed by LS-GLasso and established the local group selection consistency. For the the post-hard-thresholding estimator at the second stage, we showed that it naturally inherits all those nice statistical properties from the first stage and further possesses bi-level variable selection consistency. These theoretical results require weak assumptions on model settings and are applicable even though the random error and covariates are heavy-tailed or the data set is contaminated by outliers.

Our framework is computationally efficient, and is able to find a well-behaved local stationary point if a consistent initial such as Huber group Lasso is used. Our numerical studies showed satisfactory finite sample performances of the two-stage

penalized M-estimator under different irregular settings, which is consistent with our theoretical findings. In particular at the first stage, among some of the possible choices of loss and penalty functions that fit in the proposed framework, our numerical studies suggested to consider a redescending loss function, such as Cauchy loss or Tukey's biweight loss, with a group concave folded penalty, such as group MCP penalty, when the data are strongly contaminated.

Since the proposed framework relies on prior group information, it may not be reliable when the incorrect group information is used. To tackle this problem, [Gao18] proposes a class of penalized regression estimators by controlling group k-largest norm (GKAN), which is resistant to the fussy group information. However, the GKAN estimator is not robust to outliers or heavy-tailed errors. Hence, it will also be interesting to generalize the GKAN to a certain robust method by following the similar spirit of the proposed framework.

BIBLIOGRAPHY

[ACG+13] Andreas Alfons, Christophe Croux, Sarah Gelper, et al., *Sparse least trimmed squares regression for analyzing high-dimensional large data sets*, The Annals of Applied Statistics **7** (2013), no. 1, 226–248.

[ACO+18] Byung Chull An, Yoo-Duk Choi, In-Jae Oh, Ju Han Kim, Jae-Il Park, and Seung-won Lee, *Gpx3-mediated redox signaling arrests the cell cycle and acts as a tumor suppressor in lung cancer cell lines*, PloS one **13** (2018), no. 9, e0204170.

[Aka98] Hirotogu Akaike, *Information theory and an extension of the maximum likelihood principle*, Selected papers of hirotugu akaike, Springer, 1998, pp. 199–213.

[AM14] Marco Andrés Avella Medina, *Influence functions for penalized m-estimators*.

[Ars12] Olcay Arslan, *Weighted lad-lasso method for robust parameter estimation and variable selection in regression*, Computational Statistics & Data Analysis **56** (2012), no. 6, 1952–1965.

[AS03] Cynthia Anderson and Randall E Schumacker, *A comparison of five robust regression methods with ordinary least squares regression: Relative efficiency, bias, and test of the null hypothesis*, Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences **2** (2003), no. 2, 79–103.

[B+99] Sergey Bakin et al., *Adaptive regression and model selection in data mining problems*.

[BA55] George EP Box and Sigurd L Andersen, *Permutation theory in the derivation of robust criteria and the study of departures from assumption*, Journal of the Royal Statistical Society: Series B (Methodological) **17** (1955), no. 1, 1–26.

[Bac08] Francis R Bach, *Consistency of the group lasso and multiple kernel learning*, Journal of Machine Learning Research **9** (2008), no. Jun, 1179–1225.

[BBEKY13] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu, *Optimal m-estimation in high-dimensional regression*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14563–14568.

[BC64] George EP Box and David R Cox, *An analysis of transformations*, Journal of the Royal Statistical Society: Series B (Methodological) **26** (1964), no. 2, 211–243.

[BC+11] Alexandre Belloni, Victor Chernozhukov, et al., *$\ell_1$1-penalized quantile regression in high-dimensional sparse models*, The Annals of Statistics **39** (2011), no. 1, 82–130.

[BFW11] Jelena Bradic, Jianqing Fan, and Weiwei Wang, *Penalized composite quasi-likelihood for ultrahigh dimensional variable selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73** (2011), no. 3, 325–349.

[BH09] Patrick Breheny and Jian Huang, *Penalized methods for bi-level variable selection*, Statistics and its interface **2** (2009), no. 3, 369.

[Bic75] Peter J Bickel, *One-step huber estimates in the linear model*, Journal of the American Statistical Association **70** (1975), no. 350, 428–434.

[Box53] George EP Box, *Non-normality and tests on variances*, Biometrika **40** (1953), no. 3/4, 318–335.

[Bre15] Patrick Breheny, *The group exponential lasso for bi-level variable selection*, Biometrics **71** (2015), no. 3, 731–740.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov, *Simultaneous analysis of lasso and dantzig selector*, The Annals of Statistics (2009), 1705–1732.

[BSE+17] Anja Katrin Bosserhoff, Nadja Schneider, Lisa Ellmann, Lucie Heinzerling, and Silke Kuphal, *The neurotrophin neuritin1 (cpg15) is involved in melanoma migration, attachment independent growth, and vascular mimicry*, Oncotarget **8** (2017), no. 1, 1117.

[Car79] Raymond J Carroll, *On estimating variances of robust estimators when the errors are asymmetric*, Journal of the American Statistical Association **74** (1979), no. 367, 674–679.

[Col76] John R Collins, *Robust estimation of a location parameter in the presence of asymmetry*, The Annals of Statistics (1976), 68–85.

[Coo77] R Dennis Cook, *Detection of influential observation in linear regression*, Technometrics **19** (1977), no. 1, 15–18.

[CRW18] Le Chang, Steven Roberts, and Alan Welsh, *Robust lasso regression using tukey's biweight criterion*, Technometrics **60** (2018), no. 1, 36–47.

[CT07] Emmanuel Candes and Terence Tao, *The dantzig selector: statistical estimation when p is much larger than n*, The Annals of Statistics (2007), 2313–2351.

[CW88] Raymond J Carroll and Alan H Welsh, *A note on asymmetry and robustness in linear regression*, The American Statistician **42** (1988), no. 4, 285–287.

[CZK+16] Haiyue Chen, Zhenlong Zheng, Ki-Yeol Kim, Xuemei Jin, Mi Ryung Roh, and Zhehu Jin, *Hypermethylation and downregulation of glutathione peroxidase 3 are related to pathogenesis of melanoma*, Oncology reports **36** (2016), no. 5, 2737–2744.

[D+00] David L Donoho et al., *High-dimensional data analysis: The curses and blessings of dimensionality*, AMS Math Challenges Lecture (2000), 1–32.

[DCL12] Z John Daye, Jinbo Chen, and Hongzhe Li, *High-dimensional heteroscedastic regression with an application to eqtl data analysis*, Biometrics **68** (2012), no. 1, 316–326.

[DH83] David L Donoho and Peter J Huber, *The notion of breakdown point*, A festschrift for Erich L. Lehmann **157184** (1983).

[DSA18] Ying Dong, Lixin Song, and Muhammad Amin, *Scad-ridge penalized likelihood estimators for ultra-high dimensional models*, Hacettepe Journal of Mathematics and Statistics **47** (2018), no. 2, 423–436.

[EHJ+04] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al., *Least angle regression*, The Annals of statistics **32** (2004), no. 2, 407–499.

[FF93] LLdiko E Frank and Jerome H Friedman, *A statistical view of some chemometrics regression tools*, Technometrics **35** (1993), no. 2, 109–135.

[FFB14] Jianqing Fan, Yingying Fan, and Emre Barut, *Adaptive robust variable selection*, Annals of statistics **42** (2014), no. 1, 324.

[FFS11] Jianqing Fan, Yang Feng, and Rui Song, *Nonparametric independence screening in sparse ultra-high-dimensional additive models*, Journal of the American Statistical Association **106** (2011), no. 494, 544–557.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *A note on the group lasso and a sparse group lasso*, arXiv preprint arXiv:1001.0736 (2010).

[FL01] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American statistical Association **96** (2001), no. 456, 1348–1360.

[FL06] ———, *Statistical challenges with high dimensionality: Feature selection in knowledge discovery*, arXiv preprint math/0602133 (2006).

[FL08] Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 5, 849–911.

[FL10] ———, *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica **20** (2010), no. 1, 101.

[FL11] ———, *Nonconcave penalized likelihood with np-dimensionality*, IEEE Transactions on Information Theory **57** (2011), no. 8, 5467–5484.

[FLW17] Jianqing Fan, Quefeng Li, and Yuyan Wang, *Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79** (2017), no. 1, 247–265.

[FMD14] Jianqing Fan, Yunbei Ma, and Wei Dai, *Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models*, Journal of the American Statistical Association **109** (2014), no. 507, 1270–1284.

[FP+04] Jianqing Fan, Heng Peng, et al., *Nonconcave penalized likelihood with a diverging number of parameters*, The Annals of Statistics **32** (2004), no. 3, 928–961.

[FWZ+15] Kuangnan Fang, Xiaoyan Wang, Shengwei Zhang, Jianping Zhu, and Shuangge Ma, *Bi-level variable selection via adaptive sparse group lasso*, Journal of Statistical Computation and Simulation **85** (2015), no. 13, 2750–2760.

[GAF17] Xiaoli Gao, SE Ahmed, and Yang Feng, *Post selection shrinkage estimation for high-dimensional data analysis*, Applied Stochastic Models in Business and Industry **33** (2017), no. 2, 97–120.

[Gao18] Xiaoli Gao, *A flexible shrinkage operator for fussy grouped variable selection*, Statistical Papers **59** (2018), no. 3, 985–1008.

[GH10] Xiaoli Gao and Jian Huang, *Asymptotic analysis of high-dimensional lad regression with lasso*, Statistica Sinica (2010), 1485–1506.

[God60] Vidyadhar P Godambe, *An optimum property of regular maximum likelihood estimation*, The Annals of Mathematical Statistics **31** (1960), no. 4, 1208–1211.

[GPK07] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth, *Biconvex sets and optimization with biconvex functions: a survey and extensions*, Mathematical Methods of Operations Research **66** (2007), no. 3, 373–407.

[GV15] Irène Gijbels and Inge Vrinssen, *Robust nonnegative garrote variable selection in linear regression*, Computational Statistics & Data Analysis **85** (2015), 1–22.

[GZWW15] Xiao Guo, Hai Zhang, Yao Wang, and Jiang-Lun Wu, *Model selection and estimation in high dimensional regression models with group scad*, Statistics & Probability Letters **103** (2015), 86–92.

[H+64] Peter J Huber et al., *Robust estimation of a location parameter*, The Annals of Mathematical Statistics **35** (1964), no. 1, 73–101.

[Ham68] Frank R Hampel, *Contribution to the theory of robust estimation*, Ph. D. Thesis, University of California, Berkeley (1968).

[HBM12] Jian Huang, Patrick Breheny, and Shuangge Ma, *A selective review of group selection in high-dimensional models*, Statistical science: a review journal of the Institute of Mathematical Statistics **27** (2012), no. 4.

[HBMZ10] Jian Huang, Patrick Breheny, Shuangge Ma, and Cun-Hui Zhang, *The mnet method for variable selection*, Unpublished) Technical Report (2010), no. 402.

[HHM08] Jian Huang, Joel L Horowitz, and Shuangge Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, The Annals of Statistics (2008), 587–613.

[Hil77] Richard Walter Hill, *Robust regression when there are outliers in the carriers*, Ph.D. thesis, Harvard University, 1977.

[HK70] Arthur E Hoerl and Robert W Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[HMXZ09] Jian Huang, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang, *A group bridge approach for variable selection*, Biometrika **96** (2009), no. 2, 339–355.

[HMZ08] Jian Huang, Shuangge Ma, and Cun-Hui Zhang, *Adaptive lasso for sparse high-dimensional regression models*, Statistica Sinica (2008), 1603–1618.

[HS00] Xuming He and Qi-Man Shao, *On parameters of increasing dimensions*, Journal of Multivariate Analysis **73** (2000), no. 1, 120–135.

[HZ+10] Junzhou Huang, Tong Zhang, et al., *The benefit of group sparsity*, The Annals of Statistics **38** (2010), no. 4, 1978–2004.

[ITRMMZH17] Eduardo Izquierdo-Torres, Gabriela Rodríguez, Iván Meneses-Morales, and Angel Zarain-Herzberg, *Atp2a3 gene as an important player for resveratrol anticancer activity in breast cancer cells*, Molecular carcinogenesis **56** (2017), no. 7, 1703–1711.

[JH14] Dingfeng Jiang and Jian Huang, *Concave 1-norm group selection*, Biostatistics **16** (2014), no. 2, 252–267.

[Joh49] Norman L Johnson, *Systems of frequency curves generated by methods of translation*, Biometrika **36** (1949), no. 1/2, 149–176.

[JT09] Iain M Johnstone and D Michael Titterington, *Statistical challenges of high-dimensional data*, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **367** (2009), no. 1906, 4237–4253.

[JY10] Jinzhu Jia and Bin Yu, *On model selection consistency of the elastic net when p n*, Statistica Sinica (2010), 595–611.

[KBJ78] Roger Koenker and Gilbert Bassett Jr, *Regression quantiles*, Econometrica: journal of the Econometric Society (1978), 33–50.

[KBW18] Dehan Kong, Howard Bondell, and Yichao Wu, *Fully efficient robust estimation, outlier detection, and variable selection via penalized regression*, Statistica Sinica, to appear (2018).

[KCO08] Yongdai Kim, Hosik Choi, and Hee-Seok Oh, *Smoothly clipped absolute deviation on high dimensions*, Journal of the American Statistical Association **103** (2008), no. 484, 1665–1673.

[KF00] Keith Knight and Wenjiang Fu, *Asymptotics for lasso-type estimators*, Annals of statistics (2000), 1356–1378.

[Koe04] Roger Koenker, *Quantile regression for longitudinal data*, Journal of Multivariate Analysis **91** (2004), no. 1, 74–89.

[LC14] Shan Luo and Zehua Chen, *Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space*, Journal of the American Statistical Association **109** (2014), no. 507, 1229–1240.

[Lil15] Kristin Lilly, *Robust variable selection methods for grouped data*, Ph.D. thesis, 2015.

[LLLP11] Donghwan Lee, Woojoo Lee, Youngjo Lee, and Yudi Pawitan, *Sparse partial least-squares regression and its applications to high-throughput data analysis*, Chemometrics and Intelligent Laboratory Systems **109** (2011), no. 1, 1–8.

[LLZ+11] Sophie Lambert-Lacroix, Laurent Zwald, et al., *Robust regression through the huber's criterion and adaptive lasso penalty*, Electronic Journal of Statistics **5** (2011), 1015–1053.

[LMJ07] Yoonkyung Lee, Steven N MacEachern, and Yoonsuh Jung, *Regularization of case-specific parameters for robustness and efficiency*, Dept of Statistics, The Ohio State University, Columbus, Ohio, Tech. Rep **799** (2007).

[Loh17] Loh, *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, The Annals of Statistics **45** (2017), no. 2, 866–896.

[LW13] Po-Ling Loh and Martin J Wainwright, *Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Advances in Neural Information Processing Systems, 2013, pp. 476–484.

[LW+17] Po-Ling Loh, Martin J Wainwright, et al., *Support recovery without incoherence: A case for nonconvex regularization*, The Annals of Statistics **45** (2017), no. 6, 2455–2482.

[LZ08] Youjuan Li and Ji Zhu, *L 1-norm quantile regression*, Journal of Computational and Graphical Statistics **17** (2008), no. 1, 163–185.

[LZ16] Zhi-Ling Li and Shu-Feng Zhou, *A silac-based approach elicits the proteomic responses to vancomycin-associated nephrotoxicity in human proximal tubule epithelial hk-2 cells*, Molecules **21** (2016), no. 2, 148.

[Mal75] Colin L Mallows, *On some topics in robustness*, Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ (1975).

[Mas07] Pascal Massart, *Concentration inequalities and model selection*, Springer, 2007.

[MHW+11] Shuangge Ma, Jian Huang, Fengrong Wei, Yang Xie, and Kuangnan Fang, *Integrative analysis of multiple cancer prognosis studies with gene expression measurements*, Statistics in medicine **30** (2011), no. 28, 3361–3371.

[MRAS13] Gordana Maric, April AN Rose, Matthew G Annis, and Peter M Siegel, *Glycoprotein non-metastatic b (gpnmb): A metastatic mediator and emerging therapeutic target in cancer*, OncoTargets and therapy **6** (2013), 839.

[MS71] Hyde M Merrill and Fred C Schweppe, *Bad data suppression in power system static state estimation*, IEEE Transactions on Power Apparatus and Systems **6** (1971), 2718–2725.

[Mul04] Christine Muller, *Redescending m-estimators in regression analysis, cluster analysis and image analysis*, Discussiones Mathematicae-Probability and Statistics **24** (2004), 59–75.

[MY+09] Nicolai Meinshausen, Bin Yu, et al., *Lasso-type recovery of sparse representations for high-dimensional data*, The annals of statistics **37** (2009), no. 1, 246–270.

[Nes13] Yu Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming **140** (2013), no. 1, 125–161.

[NR+08] Yuval Nardi, Alessandro Rinaldo, et al., *On the asymptotic properties of the group lasso estimator for linear models*, Electronic Journal of Statistics **2** (2008), 605–633.

[NRW+12] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al., *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statistical Science **27** (2012), no. 4, 538–557.

[NT12] Nam H Nguyen and Trac D Tran, *Robust lasso with missing and grossly corrupted observations*, IEEE transactions on information theory **59** (2012), no. 4, 2036–2058.

[NYWR09] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Advances in Neural Information Processing Systems, 2009, pp. 1348–1356.

[OBC96] Robert G Oshima, Hélène Baribault, and Carlos Caulín, *Oncogenic regulation and function of keratins 8 and 18*, Cancer and Metastasis Reviews **15** (1996), no. 4, 445–471.

[ÖCA15] Viktoria Öllerer, Christophe Croux, and Andreas Alfons, *The influence function of penalized regression estimators*, Statistics **49** (2015), no. 4, 741–765.

[PC08] Trevor Park and George Casella, *The bayesian lasso*, Journal of the American Statistical Association **103** (2008), no. 482, 681–686.

[Pop76] Allen J Pope, *The statistics of residuals and the detection of outliers*.

[Riv12] Omar Rivasplata, *Subgaussian random variables: An expository note*, Internet publication, PDF (2012).

[RL05] Peter J Rousseeuw and Annick M Leroy, *Robust regression and outlier detection*, vol. 589, John Wiley & Sons, 2005.

[Rou84] Peter J Rousseeuw, *Least median of squares regression*, Journal of the American statistical association **79** (1984), no. 388, 871–880.

[RRSY19] Bala Rajaratnam, Steven Roberts, Doug Sparks, and Honglin Yu, *Influence diagnostics for high-dimensional lasso regression*, Journal of Computational and Graphical Statistics (2019), 1–14.

[RY84] Peter Rousseeuw and Victor Yohai, *Robust regression by means of s-estimators*, Robust and nonlinear time series analysis, Springer, 1984, pp. 256–272.

[S+78] Gideon Schwarz et al., *Estimating the dimension of a model*, The annals of statistics **6** (1978), no. 2, 461–464.

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *A sparse-group lasso*, Journal of Computational and Graphical Statistics **22** (2013), no. 2, 231–245.

[SMS08] Georgy Shevlyakov, Stephan Morgenthaler, and Alexander Shurygin, *Redescending m-estimators*, Journal of Statistical Planning and Inference **138** (2008), no. 10, 2906–2917.

[SMS20] Ben Sherwood, Aaron J Molstad, and Sumanta Singha, *Asymptotic properties of concave l1-norm group penalties*, Statistics & Probability Letters **157** (2020), 108631.

[SO12] Yiyuan She and Art B Owen, *Outlier detection using nonconvex penalized regression*, Journal of the American Statistical Association (2012).

[SRN+07] Uma T Shankavaram, William C Reinhold, Satoshi Nishizuka, Sylvia Major, Daisaku Morita, Krishna K Chary, Mark A Reimers, Uwe Scherf, Ari Kahn, Douglas Dolginow, et al., *Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study*, Molecular cancer therapeutics **6** (2007), no. 3, 820–832.

[SZF19] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan, *Adaptive huber regression*, Journal of the American Statistical Association (2019), 1–24.

[Tib96a] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[Tib96b] ———, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[Ver10] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027 (2010).

[VR13] William N Venables and Brian D Ripley, *Modern applied statistics with s-plus*, Springer Science & Business Media, 2013.

[Wan09] Hansheng Wang, *Forward regression for ultra-high dimensional variable screening*, Journal of the American Statistical Association **104** (2009), no. 488, 1512–1524.

[Wan13] Lie Wang, *The l1 penalized lad estimator for high dimensional linear regression*, Journal of Multivariate Analysis **120** (2013), 135–151.

[WCL07] Lifeng Wang, Guang Chen, and Hongzhe Li, *Group scad regression analysis for microarray time course gene expression data*, Bioinformatics **23** (2007), no. 12, 1486–1494.

[WH10] Fengrong Wei and Jian Huang, *Consistent group selection in high-dimensional linear regression*, Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability **16** (2010), no. 4, 1369.

[WHM+07] Harris Gavin C. Walker, Logan C, Andrew J. Hooloway, Grant W. Mckenzie, J. Elisabeth Wells, Bridget A. Robinson, and Christine M. Morrisa, *Cytokeratin krt8/18 expression differentiates distinct subtypes of grade 3 invasive ductal carcinoma of the breast*, Cancer Genetics and Cytogenetics **178** (2007), no. 2, 94–103.

[Wil97] Michael S Williams, *A regression technique accounting for heteroscedastic and asymmetric errors*, Journal of Agricultural, Biological, and Environmental Statistics (1997), 108–129.

[WJHZ13] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang, *Robust variable selection with exponential squared loss*, Journal of the American Statistical Association **108** (2013), no. 502, 632–643.

[WL09] Yichao Wu and Yufeng Liu, *Variable selection in quantile regression*, Statistica Sinica (2009), 801–817.

[WL17] Tao Wang and Zhonghua Li, *Outlier detection in high-dimensional regression model*, Communications in Statistics-Theory and Methods **46** (2017), no. 14, 6947–6958.

[WLCL18] Tao Wang, Qun Li, Bin Chen, and Zhonghua Li, *Multiple outliers detection in sparse high-dimensional regression*, Journal of Statistical Computation and Simulation **88** (2018), no. 1, 89–107.

[WLJ07] Hansheng Wang, Guodong Li, and Guohua Jiang, *Robust regression shrinkage and consistent variable selection through the lad-lasso*, Journal of Business & Economic Statistics **25** (2007), no. 3, 347–355.

[WT16] Mingqiu Wang and Guo-Liang Tian, *Robust group non-convex estimations for high-dimensional partially linear models*, Journal of Nonparametric Statistics **28** (2016), no. 1, 49–67.

[WWL12] Lan Wang, Yichao Wu, and Runze Li, *Quantile regression for analyzing heterogeneity in ultra-high dimension*, Journal of the American Statistical Association **107** (2012), no. 497, 214–222.

[XC18] Xiaojian Xu and Xiaoyu Chen, *A practical method of robust estimation in case of asymmetry*, Journal of Statistical Theory and Practice **12** (2018), no. 2, 370–396.

[XJ13] Shifeng Xiong and V Roshan Joseph, *Regression with outlier shrinkage*, Journal of Statistical Planning and Inference **143** (2013), no. 11, 1988–2001.

[YHZ14] GuangRen Yang, Jian Huang, and Yong Zhou, *Concave group methods for variable selection and estimation in high-dimensional varying coefficient models*, Science China Mathematics **57** (2014), no. 10, 2073–2090.

[YL06] Ming Yuan and Yi Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68** (2006), no. 1, 49–67.

[YL07] ———, *On the non-negative garrotte estimator*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69** (2007), no. 2, 143–161.

[Yoh87] Victor J Yohai, *High breakdown-point and high efficiency robust estimates for regression*, The Annals of Statistics (1987), 642–656.

[Z+10] Cun-Hui Zhang et al., *Nearly unbiased variable selection under minimax concave penalty*, The Annals of statistics **38** (2010), no. 2, 894–942.

[ZFB14] Isabella Zwiener, Barbara Frisch, and Harald Binder, *Transforming rna-seq data to improve the performance of prognostic gene signatures*, PloS one **9** (2014), no. 1, e85150.

[ZH05] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.

[ZH06] Cun-Hui Zhang and Jian Huang, *Model-selection consistency of the lasso in highdimensional linear regression*, Ann. Statist **36** (2006), 1567–1594.

[ZH+08] Cun-Hui Zhang, Jian Huang, et al., *The sparsity and bias of the lasso selection in high-dimensional linear regression*, The Annals of Statistics **36** (2008), no. 4, 1567–1594.

[ZLL+13] Junlong Zhao, Chenlei Leng, Lexin Li, Hansheng Wang, et al., *High-dimensional influence measure*, The Annals of Statistics **41** (2013), no. 5, 2639–2667.

[ZLLZ11] Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu, *Model-free feature screening for ultrahigh-dimensional data*, Journal of the American Statistical Association **106** (2011), no. 496, 1464–1475.

[ZLNL19] Junlong Zhao, Chao Liu, Lu Niu, and Chenlei Leng, *Multiple influential point detection in high dimensional regression spaces*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **81** (2019), no. 2, 385–408.

[Zou06] Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American statistical association **101** (2006), no. 476, 1418–1429.

[ZRY+09] Peng Zhao, Guilherme Rocha, Bin Yu, et al., *The composite absolute penalties family for grouped and hierarchical variable selection*, The Annals of Statistics **37** (2009), no. 6A, 3468–3497.

[ZX14] Lingmin Zeng and Jun Xie, *Group variable selection via scad-l 2*, Statistics **48** (2014), no. 1, 49–66.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of lasso*, The Journal of Machine Learning Research **7** (2006), 2541–2563.

[ZY08a] Hui Zou and Ming Yuan, *Regularized simultaneous model selection in multiple quantiles regression*, Computational Statistics & Data Analysis **52** (2008), no. 12, 5296–5304.

[ZY+08b] Hui Zou, Ming Yuan, et al., *Composite quantile regression and the oracle model selection theory*, The Annals of Statistics **36** (2008), no. 3, 1108–1126.

[ZZ09] Hui Zou and Hao Helen Zhang, *On the adaptive elastic-net with a diverging number of parameters*, Annals of statistics **37** (2009), no. 4, 1733.

[ZZ10] Nengfeng Zhou and Ji Zhu, *Group variable selection via a hierarchical lasso and its oracle property*, arXiv preprint arXiv:1006.2871 (2010).

APPENDIX A

PROOF

A.1. Proof in Chapter 2

Proof of Theorem II.2

Let $\psi(\mathbf{t}) = (\psi(t_1), \cdots, \psi(t_n))'$ and $\mathbf{\Theta}(\mathbf{t}) = (\Theta(t_1), \cdots, \Theta(t_n))'$. If $\tilde{\mathbf{W}}$ is obtained at a fixed point, then

$$\tilde{\mathbf{W}}^2 = diag\{\mathbf{\Theta}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\}$$

and

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{y}.$$

Thus

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{y} = \tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y}, \tag{A.1}$$

and

$$\tilde{\mathbf{W}}^2 = diag\{\mathbf{\Theta}(\mathbf{y} - \mathbf{X}(\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{y})\} = diag\{\mathbf{\Theta}(\tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y})\},$$

where $\mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}} = \tilde{\mathbf{W}}\mathbf{X}(\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{W}}$. Let ψ and Θ satisfy (II.4). Then from (A.1),

$$\begin{aligned}
\mathbf{X}'\psi(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) &= \mathbf{X}'\psi(\tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y}) \\
&= \mathbf{X}'diag\{\mathbf{\Theta}(\tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y})\}\mathbf{W}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y} \\
&= \mathbf{X}'\tilde{\mathbf{W}}^2(\tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y} = \mathbf{0}.
\end{aligned}$$

□

A.1.1. Proof in Section II.4

To prove those lemmas and Theorem II.5 in Section II.4, we need to reformulate the model as follows. In particular, we define $r_{i,\boldsymbol{\beta}} = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ and a $n \times n$ matrix

$\mathbf{R}_{\boldsymbol{\beta}} = diag\{r_{1,\boldsymbol{\beta}}, \cdots, r_{n,\boldsymbol{\beta}}\}$. Let $\mathbf{r}_{i,\boldsymbol{\beta}}$ be the ith column vector of $\mathbf{R}_{\boldsymbol{\beta}}$. Recall the notation $\nu_i = 1 - w_i$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1 = (\beta_1, \cdots, \beta_p)'$ and $\boldsymbol{\theta}_2 = (\lambda_{2n}/\lambda_{1n})(\nu_1, \cdots, \nu_n)'$. Define $\boldsymbol{z}_{i,\boldsymbol{\beta}}' = \left(\mathbf{x}_i', (\lambda_{1n}/\lambda_{2n})\mathbf{r}_{i,\boldsymbol{\beta}}'\right)$ and $\mathbf{Z}_{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{z}_{1,\boldsymbol{\beta}}' \\ \cdots \\ \boldsymbol{z}_{n,\boldsymbol{\beta}}' \end{pmatrix} = \begin{pmatrix} \mathbf{X} & (\lambda_{1n}/\lambda_{2n})\mathbf{R}_{\boldsymbol{\beta}} \end{pmatrix}$. Then model (II.1) with true parameter values becomes

$$y_i = \mathbf{r}_{i,\boldsymbol{\beta}^*}'\boldsymbol{\nu}^* + \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i = \boldsymbol{z}_{i,\boldsymbol{\beta}^*}'\boldsymbol{\theta}^* + \epsilon_i. \tag{A.2}$$

Recall that the penalized likelihood of PAWLS in (II.3),

$$L(\boldsymbol{\beta}, \mathbf{w}) = \frac{1}{2n}\|\boldsymbol{\Omega}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\mathbf{1} - \mathbf{w}\|_1,$$

where $\boldsymbol{\Omega} = diag\{w_1, \cdots, w_n\}$ and $\mathbf{1}$ is n-dimensional vector with all elements being 1.

Notice that $\lambda_1\|\boldsymbol{\theta}\|_1 = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\nu}\|_1$. Then the above penalized likelihood becomes

$$L(\boldsymbol{\theta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{Z}_{\boldsymbol{\beta}}\boldsymbol{\theta}\|^2 + \lambda_1\|\boldsymbol{\theta}\|_1.$$

Proof of Lemma II.3

Using the definition,

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}}\|^2 + \lambda_{1n}\|\hat{\boldsymbol{\theta}}\|_1 \leq \frac{1}{2n}\|\mathbf{y} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|^2 + \lambda_{1n}\|\boldsymbol{\theta}^*\|_1.$$

Then

$$\begin{aligned} \frac{1}{2n}\|\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|^2 &\leq \frac{1}{n}\epsilon'(\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*) + \lambda_{1n}[\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1] \\ &\leq \frac{1}{n}|\epsilon'\mathbf{Z}_{\boldsymbol{\beta}^*}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)| + \frac{1}{n}|\epsilon'(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}| + \lambda_{1n}[\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1] \end{aligned} \tag{A.3}$$

Notice that

$$\begin{aligned} \mathbf{Z}_{\boldsymbol{\beta}^*}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) &= \mathbf{X}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*) + (\lambda_{1n}/\lambda_{2n})\mathbf{R}_{\boldsymbol{\beta}^*}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*) \\ &= \mathbf{X}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*) + (\lambda_{1n}/\lambda_{2n})\boldsymbol{\Omega}^{*-1}\mathbf{D}_\epsilon(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*), \end{aligned}$$

where $\mathbf{D}_\epsilon = diag(\epsilon_1, \cdots, \epsilon_n)$ is diagonal matrix consisting of ϵ. Similar notations are applied for other diagonal matrices, such as $\mathbf{D}_{\boldsymbol{\nu}}$. Then on event $\mathbb{A}_1 \cap \mathbb{A}_2$, we have

$$\begin{aligned}\frac{1}{n}|\epsilon' \mathbf{Z}^*_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)| &\leq \frac{1}{n}\|\epsilon'\mathbf{X}\|_\infty \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*_1\|_1 + \frac{\lambda_{1n}}{n\lambda_{2n}} \max_{1\leq i\leq n} \frac{\epsilon_i^2}{w_i^*} \|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*_2\|_1 \\ &\leq \frac{\lambda_{1n}}{4}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*_1\|_1 + \frac{\lambda_{1n}}{4}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*_2\|_1 \\ &\leq \frac{\lambda_{1n}}{4}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1.\end{aligned} \tag{A.4}$$

Notice that on event $\mathbb{A}_3$,

$$(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}} = (\lambda_{1n}/\lambda_{2n}) diag(\mathbf{x}_1'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \cdots, \mathbf{x}_n'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))\hat{\boldsymbol{\theta}}_2 = \mathbf{D}_{\tilde{\nu}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

Then

$$\begin{aligned}\frac{1}{n}|\epsilon'(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}| &= \frac{1}{n}|\epsilon'\mathbf{D}_{\tilde{\nu}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \\ &\leq \frac{1}{n}\|\epsilon'\mathbf{D}_{\tilde{\nu}}\mathbf{X}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \\ &\leq (\lambda_{1n}/4)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1,\end{aligned} \tag{A.5}$$

where the last "$\leq$" holds on events $\mathbb{A}_3$.

From (A.4-A.5), we obtain

$$\begin{aligned}\frac{1}{2n}\|Z_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - Z_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|^2 &\leq \frac{\lambda_{1n}}{4}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \frac{\lambda_{1n}}{4}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda_{1n}[\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1] \\ &= \frac{\lambda_{1n}}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda_{1n}[\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1] \\ &\quad + \frac{\lambda_{2n}}{4}\|\tilde{\nu} - \boldsymbol{\nu}^*\|_1 + \lambda_{2n}[\|\boldsymbol{\nu}^*\|_1 - \|\tilde{\nu}\|_1].\end{aligned} \tag{A.6}$$

Adding $\frac{\lambda_{1n}}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \frac{\lambda_{2n}}{2}\|\tilde{\nu} - \boldsymbol{\nu}^*\|_1$ on two sides,

$$\begin{aligned}&\frac{1}{2n}\|Z_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - Z_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|^2 + \frac{\lambda_{1n}}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \frac{\lambda_{2n}}{2}\|\tilde{\nu} - \boldsymbol{\nu}^*\|_1 \\ &\leq \lambda_{1n}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + [\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1]) \\ &\quad + \lambda_{2n}(\|\tilde{\nu} - \boldsymbol{\nu}^*\|_1 + [\|\boldsymbol{\nu}^*\|_1 - \|\tilde{\nu}\|_1]) \\ &\leq 2\lambda_{1n}\|\hat{\boldsymbol{\beta}}_{J_{10}} - \boldsymbol{\beta}^*_{J_{10}}\|_1 + 2\lambda_{2n}\|\tilde{\nu}_{J_{20}} - \boldsymbol{\nu}^*_{J_{20}}\|_1.\end{aligned} \tag{A.7}$$

The last "$\leq$" holds since $\|\hat{\boldsymbol{\beta}}_{J_{10}^c} - \boldsymbol{\beta}^*_{J_{10}^c}\|_1 + \|\boldsymbol{\beta}^*_{J_{10}^c}\|_1 - \|\hat{\boldsymbol{\beta}}_{J_{10}^c}\|_1 = 0$ and $\|\tilde{\nu}_{J_{20}^c} - \boldsymbol{\nu}^*_{J_{20}^c}\|_1 +$

$\|\boldsymbol{\nu}^*_{J_{20}^c}\|_1 - \|\tilde{\nu}_{J_{20}^c}\|_1 = 0$. Thus we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + (\lambda_{2n}/\lambda_{1n})\|\tilde{\nu} - \boldsymbol{\nu}^*\|_1 \leq 4\|\hat{\boldsymbol{\beta}}_{J_{10}} - \boldsymbol{\beta}^*_{J_{10}}\|_1 + 4(\lambda_{2n}/\lambda_{1n})\|\tilde{\nu}_{J_{20}} - \boldsymbol{\nu}^*_{J_{20}}\|_1.$$

Thus (II.11) holds.

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq 4\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1$$

and

$$\|\hat{\boldsymbol{\theta}}_{J_0^c} - \boldsymbol{\theta}^*_{J_0^c}\|_1 \leq 3\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1.$$

□

Proof of Lemma II.4

$$\begin{aligned}
P(\mathbb{A}_1^c) &= P\left(\|\mathbf{X}'\epsilon\|_\infty > n\lambda_{1n}/4\right) \\
&= P\left(\max_{1\leq j\leq p} |\sum_{i=1}^n x_{ij}\epsilon_i| > n\lambda_{1n}/4\right) \\
&= P\left(\max_{1\leq j\leq p} |\tau_j| > \sqrt{n}\lambda_{1n}/(4\sigma)\right) \\
&\leq pP\left(|\tau_j| > \sqrt{n}\lambda_{1n}/(4\sigma)\right) \\
&\leq 2p\exp\left\{-\frac{n\lambda_1^2}{32\sigma^2}\right\}.
\end{aligned}$$

where $\tau_j = n^{-1/2}\sum_{i=1}^n x_{ij}\epsilon_i/\sigma$ is sub-Gaussian distribution with mean with parameter 1 if $\sum_{i=1}^n x_{ij}^2 = n$. If we let $\lambda_{1n} = \sigma(c_1)^{1/2}(\ln p/n)^{1/2}$ for $c_1 > 32$, then

$$P(\mathbb{A}_1^c) \leq 2p^{1-c_1/32} \to 0.$$

We now check event $\mathbb{A}_2$. Since

$$\begin{aligned} P(\mathbb{A}_2^c) &\leq P\left(\max_{1\leq i\leq n}\epsilon_i^2 > \frac{n\lambda_{2n}a_n}{4}\right) \\ &\leq nP\left(|\epsilon_i| > \frac{\sqrt{n\lambda_{2n}a_n}}{2}\right) \\ &\leq 2n\exp\left\{-\frac{n\lambda_{2n}a_n^2}{8\sigma^2}\right\}. \end{aligned}$$

The last "$\leq$" is due to the sub-Gaussian property of ϵ_i. If we let $\lambda_{2n} = c_2\sigma^2\log(n)/(na_n^2)$ for some $c_2 > 8$, then $P(\mathbb{A}_2^c) = 2n^{1-c_2/8} \to 0$.

We now check event $\mathbb{A}_3$. For any estimation $\tilde{\nu}$, we have

$$\begin{aligned} P(\mathbb{A}_3^c) &\leq P\left(\left(\sum_{1\leq i\leq n}\epsilon_i^2\right)^{1/2}\left(\max_{1\leq j\leq p}\sum_{1\leq i\leq n}\tilde{\nu}_i^2 x_{ij}\right)^{1/2} > n\lambda_{1n}/4\right) \\ &\leq P\left(\left(\sum_{1\leq i\leq n}\epsilon_i^2\right)^{1/2} n^{1/2} > n\lambda_{1n}/4\right) \\ &\leq P\left(\sum_{1\leq i\leq n}\frac{1}{n}\frac{\epsilon_i^2}{\sigma^2} > \frac{\lambda_{1n}^2}{16\sigma^2}\right) \\ &\leq 2\exp\left\{-M_0\min\left\{\frac{n\lambda_{1n}^4}{256K^2\sigma^4}, \frac{n\lambda_{1n}^2}{16K\sigma^2}\right\}\right\}, \end{aligned} \tag{A.8}$$

where $K = \sup_{q\geq 1} q^{-1}\left[E\left(\epsilon_1^2/\sigma^2\right)^q\right]^{1/q}$ and $M_1 > 0$ is an absolute constant. This last "$\leq$" is from Bernstein-type inequality for sub-exponential random variables [Ver10]. Notice that ϵ_i^2/σ^2 is centered sub-exponential if ϵ_i/σ is subGaussian with mean 0 and scale parameter σ. If ϵ_i is normal, then $K = 1$. The rest of the proof is straightforward by plugging in the above $\lambda_{1n} = \sigma(c_1)^{1/2}(\ln p/n)^{1/2}$ for $c_1 > 32$ in (A.8). □

Proof of Theorem II.5

Define $\hat{\mathbf{\Sigma}}^* = \frac{1}{n}\mathbf{Z}'_{\boldsymbol{\beta}^*}\mathbf{Z}_{\boldsymbol{\beta}^*}$ and $\mathbf{\Sigma} = E[\hat{\mathbf{\Sigma}}^*]$. The "$\hat{\cdot}$" on $\hat{\mathbf{\Sigma}}^*$ is used to address its

stochastic property, not the estimating behavior. From the definition, we have

$$n\hat{\mathbf{\Sigma}}^* = \sum_{i=1} \boldsymbol{z}_{i,\boldsymbol{\beta}^*}\boldsymbol{z}'_{i,\boldsymbol{\beta}^*} = \begin{pmatrix} \sum_{1\le i\le n}\mathbf{x}_i\mathbf{x}'_i & (\lambda_{1n}/\lambda_{2n})\sum_{1\le i\le n}\mathbf{x}_i\mathbf{r}'_{i,\boldsymbol{\beta}^*} \\ (\lambda_{1n}/\lambda_{2n})\sum_{1\le i\le n}\mathbf{r}_{i,\boldsymbol{\beta}^*}\mathbf{x}'_i & (\lambda_{1n}/\lambda_{2n})^2\sum_{1\le i\le n}\mathbf{r}_{i,\boldsymbol{\beta}^*}\mathbf{r}'_{i,\boldsymbol{\beta}^*} \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \frac{1}{n}\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0}_{p\times n} \\ \mathbf{0}_{n\times p} & \sigma^2(\lambda_{1n}/\lambda_{2n})^2\mathbf{\Omega}^{*-2} \end{pmatrix}$$

since $E\left[\sum_{i=1}^n \mathbf{r}_{i,\boldsymbol{\beta}^*}\mathbf{r}'_{i,\boldsymbol{\beta}^*}\right] = \sigma^2\mathbf{\Omega}^{*-2} = diag\{\sigma^2/w_1^{*2},\cdots,\sigma^2/w_n^{*2}\}$. Let $\delta_n = \|\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\|_\infty$, the supremum of all absolute values. For a $n+p$ dimensional vector such that $\|\mathbf{d}_{J_0^c}\|_1 \le 3\|\mathbf{d}_{J_0}\|_1$, we have

$$|(\mathbf{d}'\hat{\mathbf{\Sigma}}^*\mathbf{d}) - (\mathbf{d}'\mathbf{\Sigma}\mathbf{d})| \le \delta_n(\|\mathbf{d}\|_1)^2 \le 16\delta_n(\|\mathbf{d}_{J_0}\|_1)^2 \le 16s\delta_n(\|\mathbf{d}_{J_0}\|)^2. \tag{A.9}$$

The last "$\le$" is from the Cauchy-Schwartz inequality. From the condition $RE(s,3)$ in (II.10) and (A.9), we have

$$\begin{aligned} \kappa(s,3)\|\mathbf{d}_{J_0}\| &\le (\mathbf{d}'\mathbf{\Sigma}\mathbf{d})^{1/2} \\ &\le (\mathbf{d}'\hat{\mathbf{\Sigma}}^*\mathbf{d})^{1/2} + (|\mathbf{d}'(\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma})\mathbf{d}|)^{1/2} \\ &\le (1/\sqrt{n})\|\mathbf{Z}_{\boldsymbol{\beta}^*}\mathbf{d}\| + 4\sqrt{s\delta_n}\|\mathbf{d}_{J_0}\|. \end{aligned}$$

Plugging in $\mathbf{d} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$, we obtain

$$\begin{aligned} &\kappa(s,3)\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\| \\ &\le (1/\sqrt{n})\|\mathbf{Z}_{\boldsymbol{\beta}^*}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| + 4\sqrt{s\delta}\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\| \\ &\le (1/\sqrt{n})\|(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}\| + (1/\sqrt{n})\|\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\| + 4\sqrt{s\delta_n}\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|. \end{aligned} \tag{A.10}$$

We will check $(1/\sqrt{n})\|(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}\|$ and $(1/\sqrt{n})\|\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|$ separately.

First, from the proof in Lemma II.3, we know

$$\begin{aligned} (1/2n)\|\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\|^2 &\le (\lambda_{1n}/2)\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \lambda_{1n}[\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1] \\ &\le \lambda_{1n}\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1 \\ &\le \lambda_{1n}\sqrt{s}\|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|. \end{aligned}$$

Then

$$(1/\sqrt{n})\|\mathbf{Z}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\theta}} - \mathbf{Z}_{\boldsymbol{\beta}^*}\boldsymbol{\theta}^*\| \le (2\lambda_{1n})^{1/2}s^{1/4}\|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|^{1/2}. \tag{A.11}$$

On the other hand,

$$\begin{aligned}(1/n)\|(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}\|^2 &= (1/n)\textstyle\sum_{i=1}^n \left[\widehat{\nu}_i \mathbf{x}_i'(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\right]^2 \\ &\le (1/n)\textstyle\sum_{i=1}^n \left[\widehat{\nu}_i^2 \max_{1\le j\le p} x_{ij}^2 (\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1)^2\right] \\ &\le (\widehat{s}_{2n}/n) b_n^2 (\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1)^2 \\ &\le (\widehat{s}_{2n}/n) b_n^2 4 (\|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|_1)^2 \\ &\le (\widehat{s}_{2n}/n) b_n^2 4s (\|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|)^2,\end{aligned}$$

where $\widehat{s}_{2n} = \sum_{i=1}^n \widehat{\nu}_i$. Then

$$(1/\sqrt{n})\|(\mathbf{Z}_{\boldsymbol{\beta}^*} - \mathbf{Z}_{\hat{\boldsymbol{\beta}}})\hat{\boldsymbol{\theta}}\| \le 2s^{1/2}(\widehat{s}_{2n}/n)^{1/2} b_n \|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|. \tag{A.12}$$

In fact, as what we will verify in Lemma A.1 and A.2, if $\lambda_{1n}/\lambda_{2n} \le O(1)$, then for any $\zeta > 0$, we have

$$P\left((s\delta_n)^{1/2} > \kappa(s,3)/16\right) \to 0$$

and

$$P\left(b_n(s\widehat{s}_{2n}/n)^{1/2} > \kappa(s,3)/8\right) \to 0.$$

Thus from (A.10-A.12), we have

$$\begin{aligned}\kappa(s,3)\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\| &\le 2s^{1/2}(\widehat{s}_{2n}/n)^{1/2} b_n \|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\| \\ &\quad + (2\lambda_{1n})^{1/2}s^{1/4}\|\boldsymbol{\theta}^*_{J_0} - \hat{\boldsymbol{\theta}}_{J_0}\|^{1/2} + 4(s\delta_n)^{1/2}\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|.\end{aligned}$$

Then

$$\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\| \le \frac{2\lambda_{1n}s^{1/2}}{[\kappa(s,3) - (2s^{1/2}(\widehat{s}_{2n}/n)^{1/2}b_n + 4(s\delta_n)^{1/2})]^2} \le \frac{8\lambda_{1n}s^{1/2}}{\kappa^2(s,3)}.$$

Thus

$$\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\|_1 \le s^{1/2}\|\hat{\boldsymbol{\theta}}_{J_0} - \boldsymbol{\theta}^*_{J_0}\| \le \frac{8\lambda_{1n}s}{\kappa^2(s,3)}.$$

□

Lemma A.1. *Suppose (A1) and (A2) hold. Then under* $\lambda_{1n}/\lambda_{2n} \leq O(1)$, $s\delta_n = o_P(1)$. *Specifically, for any* $\xi > 0$, *we have*

$$P\left(s\delta_n > \zeta\right) \leq \frac{3\sigma}{\sqrt{\zeta}} \frac{\lambda_{1n}\sqrt{s}}{\sqrt{n}\lambda_{2n}a_n} \sqrt{1+\log(2n)} + \frac{3\sigma}{\sqrt{2\zeta}} \frac{s\lambda_{1n}b_n}{n\lambda_{2n}a_n} \sqrt{1+\log(2n)} \to 0. \quad \text{(A.13)}$$

Proof of Lemma A.1

Notice that $E[\mathbf{R}_{\boldsymbol{\beta}^*}] = \mathbf{0}$ and $E[\mathbf{R}^2_{\boldsymbol{\beta}^*}] = \sigma^2\boldsymbol{\Omega}^{*-2}$. Then

$$\hat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma} = (1/n) \begin{pmatrix} \mathbf{0}_{p\times p} & (\lambda_{1n}/\lambda_{2n})\mathbf{X}'\mathbf{R}_{\boldsymbol{\beta}^*} \\ (\lambda_{1n}/\lambda_{2n})\mathbf{X}\mathbf{R}'_{\boldsymbol{\beta}^*} & (\lambda_{1n}/\lambda_{2n})^2(\mathbf{R}^2_{\boldsymbol{\beta}^*} - \sigma^2\boldsymbol{\Omega}^{*-2}) \end{pmatrix}.$$

Then $s\|\hat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\|_\infty = \max\{(1/n)(\lambda_{1n}/\lambda_{2n})^2 s\|\mathbf{R}^2_{\boldsymbol{\beta}^*} - \sigma^2\boldsymbol{\Omega}^{*-2}\|_\infty, (1/n)(\lambda_{1n}/\lambda_{2n})s\|\mathbf{X}'\mathbf{R}_{\boldsymbol{\beta}^*}\|_\infty\}$. We will check $\frac{s\lambda^2_{1n}}{n\lambda^2_{2n}}\|\mathbf{R}^2_{\boldsymbol{\beta}^*} - \sigma^2\boldsymbol{\Omega}^{*-2}\|_\infty \to 0$ and $(1/n)(s\lambda_{1n}/\lambda_{2n})\|\mathbf{X}'\mathbf{R}_{\boldsymbol{\beta}^*}\|_\infty \to 0$ with probability separately. For any $\zeta > 0$,

$$\begin{aligned}
&P\left((1/n)(\lambda_{1n}/\lambda_{2n})^2 s\|\mathbf{R}^2_{\boldsymbol{\beta}^*} - \sigma^2\boldsymbol{\Omega}^{*-2}\|_\infty > \zeta\right) \\
&\leq P\left(\max_{1\leq i\leq n} |\epsilon_i^2/\sigma^2 - 1| > (n\zeta\lambda^2_{2n}a_n^2)/(s\lambda^2_{1n}\sigma^2)\right) \\
&\leq P\left(\max_{1\leq i\leq n} \epsilon_i^2/\sigma^2 > (n\zeta\lambda^2_{2n}a_n^2)/(s\lambda^2_{1n}\sigma^2) - 1\right) \\
&\leq P\left(\max_{1\leq i\leq n} \epsilon_i^2/\sigma^2 > (n\zeta\lambda^2_{2n}a_n^2)/(4s\lambda^2_{1n}\sigma^2)\right) \\
&\leq P\left(\max_{1\leq i\leq n} |\epsilon_i/\sigma| > (\sqrt{\zeta}/(2\sigma))(\sqrt{n}\lambda_{2n}a_n/(\sqrt{s}\lambda_{1n}))\right) \\
&\leq (2\sigma)/(\sqrt{\zeta})(\lambda_{1n}\sqrt{s}/(\sqrt{n}\lambda_{2n}a_n))E\left[\max_{1\leq i\leq n} |\epsilon_i/\sigma|\right] \\
&\leq (3\sigma)/(\sqrt{\zeta})(\lambda_{1n}\sqrt{s}/(\sqrt{n}\lambda_{2n}a_n))\sqrt{1+\log(2n)}.
\end{aligned} \quad \text{(A.14)}$$

The third "$\leq$" holds from A2(ii) and $\lambda_{1n}/\lambda_{2n} = O(1)$. In fact, if $\lambda_{1n}/\lambda_{2n} = O(1)$ and A2(ii) hold, we also have

$$(\lambda_{1n}/\lambda_{2n})\left(\sqrt{s\log(n)}/(\sqrt{n}a_n)\right) \leq \left(\sqrt{s\log(n)}/(\sqrt{n}a_n)\right) \to 0.$$

Thus $s\lambda_{1n}^2/(n\lambda_{2n}^2)\|\mathbf{R}_{\boldsymbol{\beta}^*}^2 - \sigma^2\boldsymbol{\Omega}^{*-2}\|_\infty \to 0$. Similarly for $\forall\zeta > 0$,

$$\begin{aligned}
&P\left((1/n)(s\lambda_{1n}/\lambda_{2n})\|\mathbf{X}'\mathbf{R}_{\boldsymbol{\beta}^*}\|_\infty > \zeta\right) \\
&\leq P\left(\max_{1\leq i\leq n}|x_{ij}r_{i,\boldsymbol{\beta}^*}| > \zeta n\lambda_{2n}/(s\lambda_{1n})\right) \\
&\leq P\left(\max_{1\leq i\leq n}|\epsilon_i| > \zeta n\lambda_{2n}a_n/(s\lambda_{1n}b_n)\right) \\
&\leq (s\lambda_{1n}b_n)/(\zeta n\lambda_{2n}a_n)E\left[\max_{1\leq i\leq n}|\epsilon_i|\right] \\
&\leq (3\sigma s\lambda_{1n}b_n\sqrt{1+\log(2n)})/(2\zeta n\lambda_{2n}a_n).
\end{aligned} \tag{A.15}$$

Notice that $s\lambda_{1n}b_n\sqrt{\log(n)})/(n\lambda_{2n}a_n) = (\lambda_{1n}/\lambda_{2n})(sb_n/\sqrt{n})(\log(n)/(a_n^2 n))^{1/2} \to 0$ from (A2) (i-ii) and $\lambda_{1n}/\lambda_{2n} = O(1)$. The expression of h_4 and h_5 in Theorem II.5 are obtained by replacing ζ by $(\kappa(s,3)/16)^2$ in (A.14) and (A.15). □

Lemma A.2. *Suppose (A1), (A2-i) and (A3) hold. Then under* $\lambda_{1n}/\lambda_{2n} \leq O(1)$,

$$P\left(b_n(s\widehat{s}_{2n}/n)^{1/2} > \kappa(s,3)/8\right) \to 0. \tag{A.16}$$

Proof of Lemma A.2

From (A.7),

$$\begin{aligned}
\frac{\lambda_{2n}}{2}\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\|_1 &\leq \lambda_{1n}\|\boldsymbol{\beta}^*\|_1 + \lambda_{2n}\|\widehat{\boldsymbol{\nu}}_{J_{20}} - \boldsymbol{\nu}^*_{J_{20}}\|_1 \\
&\leq s_1\|\boldsymbol{\beta}^*\|_\infty\lambda_{1n} + \lambda_{2n}\|\widehat{\boldsymbol{\nu}}_{J_{20}} - \boldsymbol{\nu}^*_{J_{20}}\|_1 \\
&\leq Ms_1 + 2s_2\lambda_{2n},
\end{aligned} \tag{A.17}$$

where $s_1 = |J_{10}|$ and $s_2 = |J_{20}|$. The last "$\leq$" is from (A3). Thus,

$$\begin{aligned}
\textstyle\sum_{i=1}^n \widehat{\nu}_i &\leq \|\boldsymbol{\nu}^*\|_1 + \|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\|_1 \\
&\leq s_2 + \|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\|_1 \\
&\leq 5s_2 + 2Ms_1(\lambda_{1n}/\lambda_{2n}).
\end{aligned}$$

If we $\lambda_{1n}/\lambda_{2n} \leq O(1)$, under (A2-i), we have

$$\sqrt{s\|\widehat{\boldsymbol{\nu}}\|_1 b_n^2/n} \leq O\left((sb_n^2/n)^{1/2}(5s_2 + 2Ms_1)^{1/2}\right) \leq O\left(sb_n/n^{1/2}\right) \to 0.$$

□

Proof of Corollary II.6

We only need to verify that A2(ii) holds when $\lambda_{1n} = \lambda_{2n}$ and $s = o\left(n^{(1-\alpha)/2}\right)$. If $p = O\left(\exp(n^{\alpha})\right)$ for $1/2 < \alpha < 1$, then $a_n^2 n^{(\alpha+1)/2} = (c_2\sigma/c_1^{1/2})\log(n)$ for $\lambda_{1n} = \lambda_{2n}$. Thus

$$\frac{s\log(n)}{na_n^2} = \frac{c_1^{1/2}}{c_2\sigma}\frac{s}{n^{(1-\alpha)/2}} \to 0.$$

Then from Theorem II.5, we get

$$\|\hat{\boldsymbol{\beta}}_{S_{10}} - \boldsymbol{\beta}^*_{S_{10}}\|_1 + \|\hat{\mathbf{w}}_{S_{20}} - \mathbf{w}^*_{S_{20}}\|_1 \leq \frac{8\lambda_{1n}s}{\kappa^2(s,3)}.$$

and

$$\|\hat{\boldsymbol{\beta}}_{S_{10}} - \boldsymbol{\beta}^*_{S_{10}}\|_2^2 + \|\hat{\mathbf{w}}_{S_{20}} - \mathbf{w}^*_{S_{20}}\|_2^2 \leq \left(\frac{8\lambda_{1n}s^{1/2}}{\kappa^2(s,3)}\right)^2.$$

Thus using the Cauchy-Schwarz inequality again,

$$\|\hat{\boldsymbol{\beta}}_{S_{10}} - \boldsymbol{\beta}^*_{S_{10}}\|_2 + \|\hat{\mathbf{w}}_{S_{20}} - \mathbf{w}^*_{S_{20}}\|_2 \leq \sqrt{2}\frac{8\lambda_{1n}s^{1/2}}{\kappa^2(s,3)}$$

□

A.2. Proof in Chapter 3

Establishing the uniform RSC condition

Let $\varepsilon_T = E\left[P\left(|\epsilon_i| \geq \frac{T}{2}|\mathbf{x}\right)\right]$ be the expected tail probability. Below we establish some sufficient conditions where an unweighted $\mathcal{L}_{\alpha,n}$ ($w(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$) satisfies the uniform RSC condition in Assumptions III.4 with high probability. The uniform RSC condition for weighted loss can be established accordingly.

Theorem A.3. *Suppose $\mathcal{L}_{\alpha,n}$ satisfies Assumption III.2 and the covariate $\mathbf{x}$ satisfies*

Assumption III.3. If $n \geq C_{10} s \log p$, *then with probability at least* $1 - C_{11}\exp(-C_{12}\log p)$, *the loss function* $\mathcal{L}_{\alpha,n}$ *satisfies the Uniform RSC condition in Assumption III.4 with*

$$\gamma = \frac{k_l}{32}, \qquad \tau = \frac{C_{13}(3+2k_2)^2 k_0^2 T_0^2}{2r^2} \quad \textit{and} \quad \alpha_0 = \max\{(2d_1)^{\frac{1}{k}}, 1\} \cdot T_0,$$

where $T_0 > 0$ *is a sufficiently large constant that satisfies*

$$C_{14} k_0^2 \left(\sqrt{\varepsilon_{T_0}} + \exp\left(-\frac{C_{15} T_0^2}{k_0^2 r^2} \right) \right) < \frac{k_l}{2 + 4k_2}. \tag{A.18}$$

Theorem A.3 guarantees that the loss function $\mathcal{L}_{\alpha,n}$ satisfies the uniform RSC condition with probability converging to 1. Note that the left hand side of inequality (A.18) is monotonically decreasing on T_0, meaning that inequality (A.18) is always satisfied for a sufficiently large T_0. In addition, while keeping inequality (A.18) satisfied, a larger T_0 (thus larger α_0) actually allows a larger radius r of local ball around $\boldsymbol{\beta}^*$ and a more contaminated distribution of ϵ. Theorem $A.3$ implies that the Huber loss, Hampel loss, Tukey's biweight loss and Cauchy loss satisfy Assumption III.4 with high probability.

Proof of Theorem III.1

Let $l(x) = \frac{1}{2}x^2$. Observe that

$$\begin{aligned} E[\nabla \frac{w(\mathbf{x})}{v(\mathbf{x})} l((y - \mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))] &= E[w(\mathbf{x})v(\mathbf{x})(y - \mathbf{x}^T\boldsymbol{\beta}^*)(-\mathbf{x})] \\ &= E[w(\mathbf{x})v(\mathbf{x})\epsilon(-\mathbf{x})] \\ &= E[E[\epsilon|\mathbf{x}]w(\mathbf{x})v(\mathbf{x})(-\mathbf{x})] \\ &= \mathbf{0}, \end{aligned}$$

where the last equality follows from $E[\epsilon|\mathbf{x}] = 0$. Hence $\boldsymbol{\beta}^*$ is the minimizer of

$E[\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta})v(\mathbf{x}))]$. Then it follows from Assumption III.3(iii) that

$$\begin{aligned}
&E[\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\\
&=E\{w(\mathbf{x})v(\mathbf{x})[l(y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)-l(y-\mathbf{x}^T\boldsymbol{\beta}^*)]\}\\
&=\frac{1}{2}(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)^TE[w(\mathbf{x})v(\mathbf{x})\mathbf{x}\mathbf{x}^T](\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)\geq\frac{1}{2}k_l\|\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*\|_2^2
\end{aligned}\tag{A.19}$$

Let $g_\alpha(x)=l(x)-l_\alpha(x)$. Since $\boldsymbol{\beta}^*_\alpha$ is the minimizer of $E[\frac{w(\mathbf{x})}{v(\mathbf{x})}l_\alpha((y-\mathbf{x}^T\boldsymbol{\beta})v(\mathbf{x}))]$ within a neighbour of $\boldsymbol{\beta}^*$, we have

$$\begin{aligned}
&E[\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\\
=&E\{\frac{w(\mathbf{x})}{v(\mathbf{x})}[l((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-l_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))]\}+\\
&E\{\frac{w(\mathbf{x})}{v(\mathbf{x})}[l_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-l_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\}+\\
&E\{\frac{w(\mathbf{x})}{v(\mathbf{x})}[l_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))-l((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\\
\leq&E[\frac{w(\mathbf{x})}{v(\mathbf{x})}g_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))]-E[\frac{w(\mathbf{x})}{v(\mathbf{x})}g_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]
\end{aligned}\tag{A.20}$$

It follows from mean value theorem that

$$\begin{aligned}
&E[\frac{w(\mathbf{x})}{v(\mathbf{x})}g_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-\frac{w(\mathbf{x})}{v(\mathbf{x})}g_\alpha((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\\
=&E[w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)(z-l'_\alpha(z))]\\
\leq&E[|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)||z-l'_\alpha(z)|]
\end{aligned}\tag{A.21}$$

where $z=(y-\mathbf{x}^T\tilde{\boldsymbol{\beta}})v(\mathbf{x})$ and $\tilde{\boldsymbol{\beta}}$ is a vector lying between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^*_\alpha$. Notice $l'_\alpha(0)=0$ in Assumption III.2(ii). By taking integral on each side of inequality in Assumption III.2(iii), we have

$$|u-l'_\alpha(u)|\leq\frac{d_1}{k+1}|u|^{k+1}\alpha^{-k},\tag{A.22}$$

for all $|u| \leq \alpha$. Observe that

$$\begin{aligned} E[|z - l'_\alpha(z)||\mathbf{x}] =& E[|z - l'_\alpha(z)|1(|z| \leq \alpha)|\mathbf{x}] + E[|z - l'_\alpha(z)|1(|z| > \alpha)|\mathbf{x}] \\ =& I_1 + I_2. \end{aligned} \tag{A.23}$$

From (A.22) we have

$$\begin{aligned} I_1 =& E[|z - l'_\alpha(z)|1(|z| \leq \alpha)|\mathbf{x}] \\ \leq& \frac{d_1 \alpha^{-k}}{k+1} E[|z|^{k+1} 1(|z| \leq \alpha)|\mathbf{x}] \\ \leq& \frac{d_1 \alpha^{-k}}{k+1} E[\frac{\alpha}{|z|}|z|^{k+1}|\mathbf{x}] \\ =& \frac{d_1 \alpha^{1-k}}{k+1} E[|z|^k|\mathbf{x}]. \end{aligned} \tag{A.24}$$

Also observe that

$$\begin{aligned} I_2 =& E[|z - l'_\alpha(z)|1(|z| > \alpha)|\mathbf{x}] \\ \leq& E[|z|1(|z| > \alpha)|\mathbf{x}] + E[|l'_\alpha(z)|1(|z| > \alpha)|\mathbf{x}] \\ <& \frac{1}{\alpha^{k-1}} E[|z|^k|\mathbf{x}] + k_1 \alpha E[1(|z| > \alpha)|\mathbf{x}] \\ =& \alpha^{1-k} E[|z|^k|\mathbf{x}] + k_1 \alpha^{1-k} E[|z|^k|\mathbf{x}] \\ =& (1 + k_1)\alpha^{1-k} E[|z|^k|\mathbf{x}], \end{aligned} \tag{A.25}$$

where the second inequality follows from Assumption III.2(i). Combining (A.23), (A.24) and (A.25), we obtain

$$E[|z - l'_\alpha(z)||\mathbf{x}] \leq (\frac{d_1}{k+1} + 1 + k_1)\alpha^{1-k} E[|z|^k|\mathbf{x}] = C_1 \alpha^{1-k} E[|z|^k|\mathbf{x}] \tag{A.26}$$

where $C_1 = \frac{d_1}{k+1} + 1 + k_1$ and k is the constant that stated in Assumption III.2(iii), Assumption III.3(i) and 3(ii).

Combining inequalities (A.20), (A.21) and (A.26), we obtain

$$\begin{aligned}
&E[\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))-\frac{w(\mathbf{x})}{v(\mathbf{x})}l((y-\mathbf{x}^T\boldsymbol{\beta}^*)v(\mathbf{x}))]\\
\leq& C_1\alpha^{1-k}E\{|y-\mathbf{x}^T\tilde{\boldsymbol{\beta}}|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|\}\\
=& C_1\alpha^{1-k}E\{|\epsilon+\mathbf{x}^T(\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}})|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|\}\\
\leq& C_1(2/\alpha)^{k-1}\{E[|\epsilon|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|]+\\
&E[|\mathbf{x}^T(\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}})|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|]\},
\end{aligned} \tag{A.27}$$

where the last inequality follows from Minkowski inequality. Note that

$$\begin{aligned}
E[|\epsilon|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|] =& E[E(|\epsilon|^k|\mathbf{x})v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|]\\
\leq& \{E[E(|\epsilon|^k|\mathbf{x})v(\mathbf{x})^k]^2\}^{\frac{1}{2}}\{E[w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)]^2\}^{\frac{1}{2}}\\
\leq& \sqrt{M_k k_u}\|\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*\|_2,
\end{aligned} \tag{A.28}$$

where the first inequality follows from Hölder inequality and the last inequality follows from Assumption III.3(i) and (iii). Observe that,

$$\begin{aligned}
E[|\mathbf{x}^T(\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}})|^k v(\mathbf{x})^k|w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)|] \leq& \{E[v(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}})]^{2k}\}^{\frac{1}{2}}\{E[w(\mathbf{x})\mathbf{x}^T(\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*)]^2\}^{\frac{1}{2}}\\
\leq& R_0^k\sqrt{q_k k_u}\|\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*\|_2,
\end{aligned} \tag{A.29}$$

where R_0 is defined in (III.9) and the last inequality follows from Assumption III.3(ii) and III.3(iii). By inequalities (A.19), (A.27), (A.28), (A.29) we have

$$\|\boldsymbol{\beta}^*_\alpha-\boldsymbol{\beta}^*\|_2 \leq 2^k C_1 k_l^{-1}\sqrt{k_u}(\sqrt{M_k}+R_0^k\sqrt{q_k})\alpha^{1-k}.$$

□

Proof of Theorem III.2

The gradient of $\mathcal{L}_{\alpha,n}$ is

$$\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i) l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))\mathbf{x}_i. \tag{A.30}$$

Recall $\boldsymbol{\beta}^*_\alpha$ is the minimizer of $E[\frac{w(\mathbf{x})}{v(\mathbf{x})} l_\alpha((y - \mathbf{x}^T\boldsymbol{\beta})v(\mathbf{x}))]$ within a neighbour of $\boldsymbol{\beta}^*$ defined in (III.9). When $\alpha \geq (\frac{2d}{R_0})^{\frac{1}{k-1}}$ where $d = 2^k C_1 k_l^{-1}\sqrt{k_u}(\sqrt{M_k} + R_0^k\sqrt{q_k})$, we have $\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2 \leq \frac{R_0}{2} < R_0$ under the result of Theorem III.1. Hence $\boldsymbol{\beta}^*_\alpha$ is an interior point of program (III.9). Then we have $E[w(\mathbf{x})l'_\alpha((y - \mathbf{x}^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))\mathbf{x}] = \mathbf{0}$. Observe that

$$\begin{aligned}
E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}] =& E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}] - \\
& E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)v(\mathbf{x}))x_{ij}] \\
\leq& k_2 E[|v(\mathbf{x}_i)\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)||w(\mathbf{x}_i)x_{ij}|] \\
\leq& k_2\{E|v(\mathbf{x}_i)\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)|^2\}^{\frac{1}{2}}\{E|w(\mathbf{x}_i)x_{ij}|^2\}^{\frac{1}{2}} \\
\leq& k_2\sqrt{q_1}\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2\sqrt{k_0^2 + d_2^2} \\
\leq& d_3\alpha^{1-k},
\end{aligned} \tag{A.31}$$

where $\max_{1\leq j\leq p}|E[w(\mathbf{x}_i)\mathbf{x}_{ij}]| < d_2 < \infty$ and $d_3 = 2^k k_2\sqrt{q_1(k_0^2 + d_2^2)k_u}C_1 k_l^{-1}(\sqrt{M_k} + 2^k R_0^k\sqrt{q_k})$. Note that the first inequality is from Assumption III.2(ii) and the third inequality follows from Assumption III.3(ii) and (iv). And the last inequality is from Theorem III.1.

Let $\mu_j = E[w(\mathbf{x}_i)\mathbf{x}_{ij}]$, $j = 1, 2, \ldots, p$. Then we have

$$\begin{aligned} E|w(\mathbf{x}_i)\mathbf{x}_{ij}|^m =& E|w(\mathbf{x}_i)\mathbf{x}_{ij} - \mu_j + \mu_j|^m \\ \leq& E[2^{m-1}(|w(\mathbf{x}_i)\mathbf{x}_{ij} - \mu_j|^m + |\mu_j|^m)] \\ \leq& 2^{m-1}[E|w(\mathbf{x}_i)\mathbf{x}_{ij} - \mu_j|^m + d_2^m] \\ \leq& 2^{m-1}[m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}) + d_2^m], \end{aligned} \tag{A.32}$$

where the last inequality follows from Assumption III.3(iv), by which $w(\mathbf{x}_i)x_{ij}$ is sub-Gaussian hence for $m > 0$([Riv12])

$$E|w(\mathbf{x}_i)x_{ij} - \mu_j|^m \leq m(\sqrt{2})^m k_0^m \Gamma(\frac{m}{2}).$$

Next we bound the $E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}]^m$ from the above. For $m \geq 2$, by Assumption III.2 and III.3(i) we have

$$\begin{aligned} E|w(\mathbf{x}_i)l'_\alpha(\epsilon_i v(\mathbf{x}_i))x_{ij}|^m \leq& E[(k_1\alpha)^{m-2}(k_2\epsilon_i v(\mathbf{x}_i))^2|w(\mathbf{x}_i)x_{ij}|^m] \\ \leq& k_1^{m-2}\alpha^{m-2}k_2^2 E[(\epsilon_i v(\mathbf{x}_i))^2|w(\mathbf{x}_i)x_{ij}|^m] \\ \leq& k_1^{m-2}\alpha^{m-2}k_2^2\{E[E(\epsilon_i^2|\mathbf{x}_i)v(\mathbf{x}_i)^2]^2\}^{1/2}\{E[(w(\mathbf{x}_i)x_{ij})^m]^2\}^{1/2} \\ \leq& k_1^{m-2}\alpha^{m-2}k_2^2\sqrt{M_2}\{E[(w(\mathbf{x}_i)x_{ij})^m]^2\}^{1/2}. \end{aligned} \tag{A.33}$$

By taking $m = 2$ in (A.33), we have

$$\begin{aligned} E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}]^2 \leq& k_2^2\sqrt{M_2}\{E[(w(\mathbf{x}_i)x_{ij})^2]^2\}^{1/2} \\ \leq& k_2^2\sqrt{M_2}(128k_0^4 + 8d_2^4)^{\frac{1}{2}} \\ \leq& d_4, \end{aligned} \tag{A.34}$$

where $d_4 = \sqrt{2}k_2^2\sqrt{M_2}(8k_0^2 + 2d_2^2)$ and the second inequality follows from (A.32).

For $m \geq 3$, by replacing m by $2m$ in (A.32), we obtain

$$\begin{aligned}
\{E|w(\mathbf{x}_i)\mathbf{x}_{ij}|^{2m}\}^{\frac{1}{2}} \leq & \{2^{2m-1}(2m)2^m k_0^{2m}\Gamma(m) + 2^{2m-1}d_2^{2m}\}^{\frac{1}{2}} \\
\leq & 2^{\frac{3m}{2}}k_0^m\sqrt{m!} + 2^{m-\frac{1}{2}}d_2^m \\
= & (2^{\frac{3m}{2}}k_0^m\frac{2}{\sqrt{m!}} + \frac{2^{m+\frac{1}{2}}d_2^m}{m!})\frac{m!}{2} \\
\leq & (2^{\frac{3m}{2}}k_0^m + 2^{m-1}d_2^m)\frac{m!}{2} \\
= & [(2^{\frac{3}{2}}k_0)^{m-2}\cdot(2^{\frac{3}{2}}k_0)^2 + (2d_2)^{m-2}\cdot 2d_2^2]\frac{m!}{2} \\
\leq & \frac{m!}{2}(2^{\frac{3}{2}}k_0 + 2d_2)^{m-2}(8k_0^2 + 2d_2^2).
\end{aligned} \tag{A.35}$$

Combining inequality (A.33) and (A.35), we have

$$\begin{aligned}
E|w(\mathbf{x}_i)l'_\alpha(\epsilon_i v(\mathbf{x}_i))x_{ij}|^m \leq & k_1^{m-2}\alpha^{m-2}k_2^2\sqrt{M_2}[\frac{m!}{2}(2^{\frac{3}{2}}k_0 + 2d_2)^{m-2}(8k_0^2 + 2d_2^2)] \\
< & \frac{m!}{2}(4(k_0+d_2)k_1\alpha)^{m-2}(k_2^2\sqrt{M_2}(8k_0^2 + 2d_2^2)) \\
< & \frac{m!}{2}(4(k_0+d_2)k_1\alpha)^{m-2}d_4,
\end{aligned}$$

By Bernstein inequality (Proposition 2.9 of [Mas07]) we have

$$\begin{aligned}
& P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij} - \tfrac{1}{n}\textstyle\sum_{i=1}^n E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}]|\right. \\
& \qquad \left. \geq \sqrt{\tfrac{2d_4t}{n}} + \tfrac{4(k_0+d_2)k_1\alpha t}{n}\right) \\
& \leq 2\exp(-t).
\end{aligned}$$

It implies that

$$\begin{aligned}
& P\left((|\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}|\right. \\
& \qquad \left. \geq \sqrt{\tfrac{2d_4t}{n}} + \tfrac{4(k_0+d_2)k_1\alpha t}{n} + |\tfrac{1}{n}\textstyle\sum_{i=1}^n E[w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}]|\right) \\
& \leq 2\exp(-t).
\end{aligned}$$

By the bound in (A.31),

$$P(|\tfrac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}| \geq \sqrt{\tfrac{2d_4t}{n}} + \tfrac{4(k_0+d_2)k_1\alpha t}{n} + d_3\alpha^{1-k}) \leq 2\exp(-t). \tag{A.36}$$

Let k_λ be a constant such that $2C^2d_4 < k_\lambda^2$ and $k_\lambda^{\frac{k-2}{k-1}} \leq \frac{C(C-8)d_4}{16(8d_3d_5^{k-2})^{\frac{1}{k-1}}(k_0+d_2)k_1}$, C is a sufficiently large constant to guarantee such k_λ exists and d_5 be an universal constant such that $\sqrt{\frac{\log p}{n}} \leq d_5$. Let $\lambda_n = k_\lambda\sqrt{\frac{\log p}{n}}$ and $t = \frac{\lambda_n^2 n}{2C^2d_4}$. Then

$$\sqrt{\frac{2d_4t}{n}} = \frac{\lambda_n}{C}. \tag{A.37}$$

Consider α that satisfies

$$\left(\frac{8d_3}{\lambda_n}\right)^{\frac{1}{k-1}} \leq \alpha \leq \frac{C(C-8)d_4}{16(k_0+d_2)k_1\lambda_n}. \tag{A.38}$$

Note that together with $\lambda_n = k_\lambda\sqrt{\frac{\log p}{n}}$ we obtain $C_2(\frac{n}{\log p})^{\frac{1}{2(k-1)}} \leq \alpha \leq C_3\sqrt{\frac{n}{\log p}}$, where $C_2 = (\frac{8d_3}{k_\lambda})^{\frac{1}{k-1}}$ and $C_3 = \frac{C(C-8)d_4}{16(k_0+d_2)k_1k_\lambda}$. By $\alpha \geq \left(\frac{8d_3}{\lambda_n}\right)^{\frac{1}{k-1}}$ we have

$$d_3\alpha^{1-k} \leq \frac{\lambda_n}{8}. \tag{A.39}$$

By $\alpha \leq \frac{C(C-8)d_4}{16(k_0+d_2)k_1\lambda_n}$ we have

$$\frac{4(k_0+d_2)k_1\alpha t}{n} \leq \frac{C(C-8)d_4t}{4n\lambda_n} = \frac{C(C-8)d_4}{4n\lambda_n}\cdot\frac{\lambda_n^2 n}{2C^2d_4} = \frac{\lambda_n(C-8)}{8C}$$

Together with (A.37) and (A.39), we obtain

$$\sqrt{\frac{2d_4t}{n}} + \frac{4(k_0+d_2)k_1\alpha t}{n} + d_3\alpha^{1-k} \leq \frac{\lambda_n}{4}.$$

Hence by (A.36), it gives

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i)l'_\alpha((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)v(\mathbf{x}_i))x_{ij}| \geq \frac{\lambda_n}{4}\right) \leq 2\exp\left(-\frac{n\lambda_n^2}{2C^2d_4}\right). \tag{A.40}$$

It then follows from union inequality that

$$P\left(\|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)\|_\infty \geq C_5\sqrt{\frac{\log p}{n}}\right) \leq 2\exp\left(-\frac{n\lambda_n^2}{2C^2d_4}+\log p\right) \leq 2\exp(-C_4\log p), \tag{A.41}$$

where $C_4 = \frac{k_\lambda^2}{2C^2d_4} - 1$ and $C_5 = \frac{k_\lambda}{4}$. Note that $C_4 > 0$ by $2C^2d_4 < k_\lambda^2$. This complete the proof for equation (III.11). And the rest of the result follows immediately from the Theorem 1 in Loh(2017).

Remark. By side conditions $\|\boldsymbol{\beta}^*\|_1 \leq R$ and $\|\hat{\boldsymbol{\beta}}\|_1 \leq R$ introduced in (III.7), we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 2R$. Thus if $\mathcal{L}_{\alpha,n}$ satisfies the uniform RSC condition with some $r \geq 2R$, which by Theorem A.3 is achievable with high probability for a sufficiently large α, then $\hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$ and thus a well-behaved PRAM estimator $\hat{\boldsymbol{\beta}}$ in Theorem III.2(ii) is attainable.

□

To prove Theorem III.3, we need the following result adopted directly from the Lemma 1 in [Loh17].

Lemma A.4. *Suppose $\mathcal{L}_{\alpha,n}$ satisfies the local RSC condition (III.4) and $n \geq \frac{2\tau}{\gamma}s\log p$. Then $\mathcal{L}_{\alpha,n}$ is strongly convex over the region $S_r = \{\boldsymbol{\beta} \in \mathbb{R}^p : supp(\boldsymbol{\beta}) \subseteq S, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$.*

Proof. The proof is similar to the proof of Lemma 1 in [Loh17]. □

Proof of Theorem III.3

The proof is an adaptation of the arguments of Theorem 2 in the paper [Loh17]. We follow the three steps of the primal-dual witness (PDW) construction described in that paper:

(i) Optimize the restricted program

$$\hat{\boldsymbol{\beta}}_S \in \underset{\boldsymbol{\beta}\in\boldsymbol{\beta}^S:\|\boldsymbol{\beta}\|_1\leq R}{\operatorname{argmin}} \left\{\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta})\right\}, \tag{A.42}$$

and establish that $\|\hat{\boldsymbol{\beta}}_S\|_1 < R$.

(ii) Recall $q_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 - \rho_\lambda(\boldsymbol{\beta})$ defined in Section III.4. Define $\hat{\boldsymbol{z}}_S \in \partial\|\hat{\boldsymbol{\beta}}_S\|_1$, and choose $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_S, \hat{\boldsymbol{z}}_{S^c})$ to satisfy the zero-subgradient condition

$$\nabla\mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}) + \lambda\hat{\boldsymbol{z}} = \mathbf{0}, \tag{A.43}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \mathbf{0}_{S^c})$. Show that $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ and establish strict dual feasibility: $\|\hat{\boldsymbol{z}}_{S^c}\|_\infty < 1$.

(iii) Verify via second order conditions that $\hat{\boldsymbol{\beta}}$ is a local minimum of the program (III.7) and conclude that all stationary points $\tilde{\boldsymbol{\beta}}$ satisfying $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$ are supported on S and agree with $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$.

Proof of Step (i) : By applying Theorem III.2 to the restricted program (A.91), we have

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 \leq \frac{96\lambda s}{4\gamma - 3\mu},$$

and thus

$$\|\hat{\boldsymbol{\beta}}_S\|_1 \leq \|\boldsymbol{\beta}^*\|_1 + \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 \leq \frac{R}{2} + \frac{96\lambda s}{4\gamma - 3\mu} < R,$$

under the assumption of the theorem. This complete step (i) of the PDW construction. □

To prove step (ii), we need the following Lemma A.10 and A.11:

Lemma A.5. *Under the conditions of Theorem III.3, we have the bound*

$$\|\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}_S^*\|_2 \leq C_6\sqrt{\frac{\log p}{ns}}$$

and $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ *with probability at least* $1 - 2\exp(-C_{41}\frac{\log p}{s^2})$.

Proof. Recall $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = (\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}, \mathbf{0}_{S^c})$. By the optimality of the oracle estimator in (III.12), we have

$$\mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}^{\mathcal{O}}) \leq \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*). \tag{A.44}$$

When $n \geq \frac{2\tau}{\gamma} s\log p$, by Lemma A.9, $\mathcal{L}_{\alpha,n}(\boldsymbol{\beta})$ is strongly convex over restricted region $S_r = \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$. Hence,

$$\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*) + \langle \nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\rangle + \frac{\gamma}{4}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 \leq \mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}^{\mathcal{O}}). \tag{A.45}$$

Together with inequality (A.93) we obtain

$$\begin{aligned}\tfrac{\gamma}{4}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 &\leq \langle \nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathcal{O}}\rangle \leq \|\nabla(\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_S\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_1 \\ &\leq \sqrt{s}\|\nabla(\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_S\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2,\end{aligned}$$

implying that

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \leq \frac{4\sqrt{s}}{\gamma}\|\nabla(\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_S\|_\infty. \tag{A.46}$$

Following the similar argument of equations (A.38) , (A.40) and (A.41) in Theorem 2, we have

$$P(\|\nabla(\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_S^*))\|_\infty) \geq \frac{\lambda_n}{4}) \leq 2\exp(-\frac{n\lambda_n^2}{2C^2d_4} + \log s),$$

for $C_{21}\lambda_n^{-\frac{1}{k-1}} \leq \alpha \leq C_{31}\lambda_n^{-1}$. Let $\lambda_n = C_{51}\sqrt{\frac{\log p}{ns^2}}$, we obtain

$$\|(\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_S\|_\infty = \|\nabla(\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_S^*))\|_\infty \leq \frac{1}{4}C_{51}\sqrt{\frac{\log p}{ns^2}} \tag{A.47}$$

with probability at least $1 - 2\exp(-C_{41}\frac{\log p}{s^2})$, where we require $s^2 \log s = \mathcal{O}(\log p)$. Then α satisifies

$$C_{22}(\frac{ns^2}{\log p})^{\frac{1}{2(k-1)}} \leq \alpha \leq C_{32}\sqrt{\frac{ns^2}{\log p}}. \tag{A.48}$$

Combining inequality (A.95) and (A.96), we obtain

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \leq C_6\sqrt{\frac{\log p}{ns}} \tag{A.49}$$

as desired, where $C_6 = C_{51}/\gamma$.

Next we show $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$. When $n > \frac{C_6^2}{r^2}\frac{\log p}{s}$, we have $\|\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}_S^*\|_2 < r$ and thus $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ is an interior point of the oracle program in (III.12), implying

$$\nabla\mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}) = \mathbf{0}. \tag{A.50}$$

By assumption that $\beta^*_{\min} \geq C_6\sqrt{\frac{\log p}{ns}} + \delta\lambda$ and inequality (A.97), we have

$$\begin{aligned} |\hat{\beta}_j^{\mathcal{O}}| \geq |\beta_j^*| - |\hat{\beta}_j^{\mathcal{O}} - \beta_j^*| &\geq \beta^*_{\min} - \|\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}_S^*\|_\infty \\ &\geq (C_6\sqrt{\tfrac{\log p}{ns}} + \delta\lambda) - C_6\sqrt{\tfrac{\log p}{ns}} \\ &= \delta\lambda. \end{aligned}$$

for all $j \in S$. Together with the assumption that ρ_λ is (μ, δ)-amenable, that is, Assumption III.2(vii), we have

$$\nabla q_\lambda(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}) = \lambda\text{sign}(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}) = \lambda\hat{\boldsymbol{z}}_S^{\mathcal{O}}, \tag{A.51}$$

where $\hat{\boldsymbol{z}}_S^{\mathcal{O}} \in \partial\|\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}\|_1$. Combining equation (A.98) and (A.99), we obtain

$$\nabla\mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}) + \lambda\hat{\boldsymbol{z}}_S^{\mathcal{O}} = \mathbf{0}. \tag{A.52}$$

Hence $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ satisifes the zero-subgradient condition on the restricted program (A.91). By step (i) $\hat{\boldsymbol{\beta}}_S$ is an interior point of the program (A.91), then it must also satisfy the

zero-subgradient condition on the restricted program. Using the strict convexity from Lemma A.11, we obtain $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$. $\square$

The following lemma guarantees that the program in (A.91) is strictly convex:

Lemma A.6. *Suppose $\mathcal{L}_{\alpha,n}$ satisfies the uniform RSC condition (III.4) and ρ_λ is μ-amenable. Suppose in addition the sampel size satisifies $n > \frac{2\tau}{\gamma-\mu} s \log p$, then the restricted program in (A.91) is strictly convex.*

We omit the proof since it is similar to the proof of Lemma 2 in [LW+17]. $\square$

Proof of step (ii) : We rewrite the zero-subgradient condition (A.92) as

$$\left(\nabla \mathcal{L}_{\alpha,n}(\hat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)\right) + \left(\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\right) + \lambda \hat{\boldsymbol{z}} = \mathbf{0}.$$

Let $\hat{Q}$ be a $p \times p$ matrix $\hat{Q} = \int_0^1 \nabla^2 \mathcal{L}_{\alpha,n}\left(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right) dt$. By the zero-subgradient condition and the fundamental theorem of calculus, we have

$$\hat{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \left(\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\right) + \lambda \hat{\boldsymbol{z}} = \mathbf{0}.$$

And its block form is

$$\begin{bmatrix} \hat{Q}_{SS} & \hat{Q}_{SS^c} \\ \hat{Q}_{S^cS} & \hat{Q}_{S^cS^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^* \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)_S - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_S) \\ \nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)_{S^c} - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) \end{bmatrix} + \lambda \begin{bmatrix} \hat{\boldsymbol{z}}_S \\ \hat{\boldsymbol{z}}_{S^c} \end{bmatrix} = \mathbf{0}. \tag{A.53}$$

The selection property implies $\nabla q_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) = \mathbf{0}$. Plugging this result into equation (A.101) and performing some algebra, we conclude that

$$\hat{\boldsymbol{z}}_{S^c} = \frac{1}{\lambda}\left\{\hat{Q}_{S^cS}(\boldsymbol{\beta}_S^* - \hat{\boldsymbol{\beta}}_S) - (\nabla \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{S^c}\right\}. \tag{A.54}$$

Therefore,

$$\begin{aligned}\|\hat{\boldsymbol{z}}_{S^c}\|_\infty &\leq \tfrac{1}{\lambda}\|\hat{Q}_{S^cS}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_\infty + \tfrac{1}{\lambda}\|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{S^c}\|_\infty \\ &\leq \tfrac{1}{\lambda}\left\{\max_{j\in S^c}\|e_j^T\hat{Q}_{S^cS}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_2\right\} + \tfrac{1}{\lambda}\|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{S^c}\|_\infty \\ &\leq \tfrac{1}{\lambda}\left\{\max_{j\in S^c}\|e_j^T\hat{Q}_{S^cS}\|_2\right\}\|(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_2 + \tfrac{1}{\lambda}\|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{S^c}\|_\infty.\end{aligned} \tag{A.55}$$

Observe that

$$\begin{aligned}[(e_j^T\hat{Q}_{S^cS})_m]^2 &\leq [\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)\mathbf{x}_{ij}v(\mathbf{x}_i)\mathbf{x}_{im}\int_0^1 l''((y_i - \mathbf{x}_i^T\boldsymbol{\beta}^* - t(\mathbf{x}_i\hat{\boldsymbol{\beta}} - \mathbf{x}_i\boldsymbol{\beta}^*))v(\mathbf{x}_i))\,\mathrm{d}t]^2 \\ &\leq k_2^2[\tfrac{1}{n}\textstyle\sum_{i=1}^n w(\mathbf{x}_i)\mathbf{x}_{ij}\cdot v(\mathbf{x}_i)\mathbf{x}_{im}]^2,\end{aligned}$$

for all $j \in S^c$ and $m \in S$, where the second inequality follows from Assumption III.2(ii). By condition of Theorem III.3, the variables $w(\mathbf{x}_i)\mathbf{x}_{ij}$ and $v(\mathbf{x}_i)\mathbf{x}_{im}$ are both sub-Gaussian. Using standard concentration results for i.i.d sums of products of sub-Gaussian variables, we have

$$P([(e_j^T\hat{Q}_{S^cS})_m]^2 \leq c_1) \geq 1 - c_2\exp(-c_3 n).$$

It then follows from union inequality that

$$P(\max_{j\in S^c}\|e_j^T\hat{Q}_{S^cS}\|_2 \leq \sqrt{c_1 s}) \geq 1 - c_2\exp(-c_3 n + \log(s(p-s))) \geq 1 - c_2\exp(-\frac{c_3}{2}n), \tag{A.56}$$

where $n \geq \frac{2}{c_3}\log(s(p-s))$. By Lemma A.10 we obtain

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_2 \leq C_6\sqrt{\frac{\log p}{ns}}. \tag{A.57}$$

Furthermore, Theorem III.2 gives

$$\|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))_{S^c}\|_\infty \leq \|\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*))\|_\infty \leq C_5\sqrt{\frac{\log p}{n}}, \tag{A.58}$$

Combining inequality (A.103), (A.104), (A.105) and (A.106), we have

$$\|\hat{\boldsymbol{z}}_{S^c}\|_\infty \leq \frac{1}{\lambda} C_7 \sqrt{\frac{\log p}{n}}.$$

with probability at least $1 - C_8 \exp(-C_{41} \frac{\log p}{s^2})$. Note that α is required to satisfy both ranges in Theorem III.2 and (A.48). Combing these two ranges we have

$$C_{22}(\frac{ns^2}{\log p})^{\frac{1}{2(k-1)}} \leq \alpha \leq C_3 \sqrt{\frac{n}{\log p}},$$

where $s^2 = \mathcal{O}\left((\frac{n}{\log p})^{k-2}\right)$. In paticular, for $\lambda > C_7 \sqrt{\frac{\log p}{n}}$, we conclude at last that the strict dual feasibility condition $\|\hat{\boldsymbol{z}}_{S^c}\|_\infty < 1$ holds, completing step (ii) of the PDW construction.

Proof of step (iii) : Since the proof for this step is almost identical to the proof in Step (iii) of Theorem 2 in [Loh17], except for the slightly different notation, we refer the reader to the arguments provided in that paper. □

To prove Theorem III.4, we need to generalized the asymptotic normality results for lower dimensional non-penalized M-estimator from [HS00] to the following Lemma:

Lemma A.7. *Suppose $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n \in \mathbb{R}^p$ are independent observations from probability distribution $F_{i,\boldsymbol{\beta}}$, $i = 1, 2, \ldots, n$, with a common parameter $\boldsymbol{\beta} \in \mathbb{R}^s$. And s may increase with the sample size n. Suppose $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \rho(\boldsymbol{z}_i, \boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}^*$ and has a unique local minimizer $\hat{\boldsymbol{\beta}}$. Define $\psi(\boldsymbol{z}_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \rho(\boldsymbol{z}_i, \boldsymbol{\beta})$ and $\eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \psi(\boldsymbol{z}_i, \tilde{\boldsymbol{\beta}}) - \psi(\boldsymbol{z}_i, \boldsymbol{\beta}) - E\psi(\boldsymbol{z}_i, \tilde{\boldsymbol{\beta}}) + E\psi(\boldsymbol{z}_i, \boldsymbol{\beta})$ and $B_s = \{\boldsymbol{\nu} \in \mathbb{R}^s : \|\boldsymbol{\nu}\|_2 = 1\}$.*

Suppose $\boldsymbol{\beta}^* \in \mathbb{R}^s$ *such that*

$$\|\sum_{i=1}^{n} \psi(\boldsymbol{z}_i, \boldsymbol{\beta}^*)\|_2 = \mathcal{O}_p((ns)^{1/2}). \tag{A.59}$$

Assume the following conditions are satisfied:

(i) $\|\sum_{i=1}^{n} \psi(\boldsymbol{z}_i, \hat{\boldsymbol{\beta}})\|_2 = o_p(n^{1/2})$.

(ii) There exist C *and* $r \in (0, 2]$ *such that* $\max_{i \le n} E_{\boldsymbol{\beta}} \sup_{\tilde{\boldsymbol{\beta}}: \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le d} \|\eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})\|_2^2 \le n^C d^r$, *for* $0 < d \le 1$.

(iii) There exists a sequence of s *by* s *matrices* $\mathbf{D}_n$ *with* $\liminf_{n \to \infty} \lambda_{min}(\mathbf{D}_n) > 0$ *such that for any* $K > 0$ *and uniformly in* $\boldsymbol{\nu} \in B_s$,

$$\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le K(s/n)^{1/2}} |\boldsymbol{\nu}^T \sum_{i=1}^{n} E_{\boldsymbol{\beta}^*}(\psi(\boldsymbol{z}_i, \boldsymbol{\beta}) - \psi(\boldsymbol{z}_i, \boldsymbol{\beta}^*)) - n\boldsymbol{\nu}^T \mathbf{D}_n(\boldsymbol{\beta} - \boldsymbol{\beta}^*)| = o((ns)^{1/2}).$$

(iv) $\sup_{\tilde{\boldsymbol{\beta}}: \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le K(s/n)^{1/2}} \sum_{i=1}^{n} E_{\boldsymbol{\beta}} |\boldsymbol{\nu}^T \eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})|^2 = \mathcal{O}(A(n, s))$ *for any* $\boldsymbol{\beta} \in \mathbb{R}^s$, $\boldsymbol{\nu} \in B_s$ *and* $K > 0$.

(v) $\sup_{\boldsymbol{\nu} \in S_s} \sup_{\tilde{\boldsymbol{\beta}}: \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le K(s/n)^{1/2}} \sum_{i=1}^{n} (\boldsymbol{\nu}^T \eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}))^2 = \mathcal{O}_p(A(n, s))$ *for any* $\boldsymbol{\beta} \in \mathbb{R}^s$ *and* $K > 0$.

If $A(n, s) = o(n/\log n)$, *we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = \mathcal{O}_p(s/n).$$

Furthermore, if $A(n, s) = o(n/(s \log n))$, *then for any unit vector* $\boldsymbol{\nu} \in \mathbb{R}^s$, *we have*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = -n^{-1} \sum_{i=1}^{n} \mathbf{D}_n^{-1} \psi(\boldsymbol{z}_i, \boldsymbol{\beta}^*) + \mathbf{r}_n,$$

with $\|\mathbf{r}_n\|_2 = o_p(n^{-1/2})$.

Proof. Our proof is similar to the proof of Theorem 1 and 2 in [HS00]. Note that in that paper, $\boldsymbol{\beta}^*$ is defined to be the solution of $\sum_{i=1}^n E_{\boldsymbol{\beta}}\psi(x_i, \boldsymbol{\beta}) = \mathbf{0}$, in addition to the condition in equation (A.59). However, a careful inspection of the proofs in that paper reveals that the results still holds if we only require $\boldsymbol{\beta}^*$ to satisfied equation (A.59). □

Proof of Theorem III.4

We then apply the result to the oracle estimator $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ defined in equation (III.12) with $w(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$. Although Lemma A.7 requires $\mathcal{L}_n$ to be convex, a throughout examination of the proofs in [HS00] shows that the results still hold if we restrict our attention to a subset of $\mathbb{R}^p$ on which $\mathcal{L}_n$ is convex and $\hat{\boldsymbol{\beta}}$ is the unique minimizer. Since $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$ is s-dimensional vector without sparsity, we denote $\mathbf{x}_i$, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ all as s-dimensional vectors for the rest of our proof. We also denote $\boldsymbol{\beta}_\alpha^*$ as $(\boldsymbol{\beta}_\alpha^*)_S$. Let $\boldsymbol{z}_i = (\mathbf{x}_i, y_i)$ and we rewrite $\rho(\boldsymbol{z}_i, \boldsymbol{\beta})$ as $l_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$, with $\mathcal{L}_{\alpha,n}$ taking the place of $\mathcal{L}_n$. Then $\psi(\boldsymbol{z}_i, \boldsymbol{\beta}) = -l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i$.

We start with verifying equation (A.59), which can be rewrited as

$$\|\sum_{i=1}^n l'_\alpha(\epsilon_i)\mathbf{x}_i\|_2 = \mathcal{O}_p((ns)^{1/2}). \tag{A.60}$$

Observe that for any $\boldsymbol{\nu} \in B_s$,

$$\begin{aligned} P(|\sum_{i=1}^n \boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i - \sum_{i=1}^n E[\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i]| > t) &\leq nVar(\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i)t^{-2} \\ &\leq nE|\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i|^2 t^{-2} \\ &\leq nE\|l'_\alpha(\epsilon_i)\mathbf{x}_i\|_2^2 t^{-2} \\ &\leq nsd_4 t^{-2}, \end{aligned} \tag{A.61}$$

where the last inequality follows from inequality (A.34). We then have

$$P(|\sum_{i=1}^{n} \boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i| > t + \sum_{i=1}^{n} |E[\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i]|) \leq nsd_4 t^{-2}. \tag{A.62}$$

Observe that

$$\begin{aligned} |E[\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i]| &= |E[l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\boldsymbol{\nu}^T\mathbf{x}_i]| \\ &= |E[l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\boldsymbol{\nu}^T\mathbf{x}_i] - E[l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*_\alpha)\boldsymbol{\nu}^T\mathbf{x}_i]| \\ &\leq k_2 E[|\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)||\boldsymbol{\nu}^T\mathbf{x}_i|] \\ &\leq k_2\{E|\mathbf{x}_i^T(\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*)|^2\}^{\frac{1}{2}}\{E|\boldsymbol{\nu}^T\mathbf{x}_i|^2\}^{\frac{1}{2}} \\ &\leq k_0^2 k_2 \|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2, \end{aligned} \tag{A.63}$$

where the first and last inequalities follow from Assumption III.2(ii) and Assumption III.3(iv) respectively. Together with the results in Theorem III.1 and condition $\alpha^{1-k} = o(n^{-1/2})$, we obtain

$$E[\boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i] = o(n^{-1/2}). \tag{A.64}$$

Thus by inequality (A.62) and (A.64) we have $\sum_{i=1}^{n} \boldsymbol{\nu}^T l'_\alpha(\epsilon_i)\mathbf{x}_i = \mathcal{O}_p((ns)^{1/2})$ for any $\boldsymbol{\nu} \in B_s$. It then implies equation (A.60).

Next we verify the conditions (i)-(v). Since the $\mathcal{L}_{\alpha,n}$ is differentiable, the left hand side of condition (i) is 0 and thus it is satisfied. By definition of η_i, we have

$$\eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) = l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i - l'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\mathbf{x}_i - El'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i + El'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\mathbf{x}_i.$$

Observe that

$$\begin{aligned} \|\eta_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})\|_2 &\leq \|l'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\mathbf{x}_i - l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i\|_2 + \|El'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\mathbf{x}_i - El'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i\|_2 \\ &\leq k_2|\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \cdot \|\mathbf{x}_i\|_2 + k_2\|E\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{x}_i\|_2 \\ &\leq k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2\|\mathbf{x}_i\|_2^2 + k_2 E\|\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\mathbf{x}_i\|_2 \\ &\leq k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2\|\mathbf{x}_i\|_2^2 + k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 E\|\mathbf{x}_i\|_2^2, \end{aligned}$$

where the second and third inequality follow from Assumption III.2(ii) and Jensen's inequality, respectively. We then obtain

$$\max_{i\le n} E_{\boldsymbol{\beta}} \sup_{\tilde{\boldsymbol{\beta}}:\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\le d} \|\eta_i(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})\|_2^2 \le \max_{i\le n} 4k_2^2 d^2 E\|\mathbf{x}_i\|_2^4.$$

Since Assumption III.3(iv) implies $E\|\mathbf{x}_i\|_2^4 = \mathcal{O}(s^2)$ for $i = 1, \cdots, n$, condition(ii) holds with $r = 2$ and if $s = \mathcal{O}(n^{r_1})$ for $r_1 > 0$.

Similarly, for any $\boldsymbol{\nu} \in B_s$, we have

$$\begin{aligned}
\boldsymbol{\nu}^T\eta_i(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta}) &\le |l'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}) - l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})||\boldsymbol{\nu}^T\mathbf{x}_i| + E[|l'_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}) - l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})||\boldsymbol{\nu}^T\mathbf{x}_i|] \\
&\le k_2|\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})||\boldsymbol{\nu}^T\mathbf{x}_i| + k_2 E[|\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})||\boldsymbol{\nu}^T\mathbf{x}_i|] \\
&\le k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2|\tilde{\boldsymbol{\nu}}^T\mathbf{x}_i||\boldsymbol{\nu}^T\mathbf{x}_i| + k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2\{E|\tilde{\boldsymbol{\nu}}^T\mathbf{x}_i|^2\}^{1/2}E\{|\boldsymbol{\nu}^T\mathbf{x}_i|^2\}^{1/2} \\
&\le k_2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2(|\tilde{\boldsymbol{\nu}}^T\mathbf{x}_i||\boldsymbol{\nu}^T\mathbf{x}_i| + k_0^2),
\end{aligned}$$

where $\tilde{\boldsymbol{\nu}} = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})/\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$. The second and last inequalities follow from Assumption III.2(ii) and Assumption III.3(iv) respectively. It then gives

$$|\boldsymbol{\nu}^T\eta_i(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})|^2 \le k_2^2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2(|\tilde{\boldsymbol{\nu}}^T\mathbf{x}_i|^2|\boldsymbol{\nu}^T\mathbf{x}_i|^2 + 2k_0^2|\tilde{\boldsymbol{\nu}}^T\mathbf{x}_i||\boldsymbol{\nu}^T\mathbf{x}_i| + k_0^4).$$

Together with Assumption III.3(iv), we obtain

$$E|\boldsymbol{\nu}^T\eta_i(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})|^2 = \mathcal{O}(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2) \tag{A.65}$$

and

$$|\boldsymbol{\nu}^T\eta_i(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})|^2 = \mathcal{O}_p(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2). \tag{A.66}$$

Hence condition (iv) and (v) hold with $A(n, s) = s$.

Finally we show condition (iii). Let $\mathbf{D}_{\alpha,n} = E[\nabla^2 \mathcal{L}_{\alpha,n}(\boldsymbol{\beta}^*)]$ and thus it is an s by s matrix. Observe that

$$\begin{aligned} E[l''_\alpha(\epsilon_i)|\mathbf{x}_i] &= E[l''_\alpha(\epsilon_i)1(|\epsilon_i| \le \alpha)|\mathbf{x}_i] + E[l''_\alpha(\epsilon_i)1(|\epsilon_i| > \alpha)|\mathbf{x}_i] \\ &\ge E[(1 - d_1|\epsilon_i|^k\alpha^{-k})1(|\epsilon_i| \le \alpha)|\mathbf{x}_i] + E[l''_\alpha(\epsilon_i)1(|\epsilon_i| > \alpha)|\mathbf{x}_i] \\ &\ge P(|\epsilon_i| \le \alpha|\mathbf{x}_i) - d_1\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i] - k_2\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i] \\ &\ge 1 - \alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i] - d_1\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i] - k_2\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i] \\ &= 1 - (d_1 + k_2 + 1)\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i], \end{aligned}$$

where the first and second inequalities follow from Assumption III.2(iii) and (ii), respectively. Thus for any $\boldsymbol{\nu} \in B_s$, we have

$$\begin{aligned} \boldsymbol{\nu}^T\mathbf{D}_{\alpha,n}\boldsymbol{\nu} &= E[l''_\alpha(\epsilon_i)\boldsymbol{\nu}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\nu}] \\ &\ge E[(1 - (d_1 + k_2 + 1)\alpha^{-k}E[|\epsilon_i|^k|\mathbf{x}_i])\boldsymbol{\nu}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\nu}] \\ &= \boldsymbol{\nu}^T E[\mathbf{x}_i\mathbf{x}_i^T]\boldsymbol{\nu} - (d_1 + k_2 + 1)\alpha^{-k}E[E(|\epsilon_i|^k|\mathbf{x}_i)(\boldsymbol{\nu}\mathbf{x}_i)^2] \\ &\ge k_l - (d_1 + k_2 + 1)\alpha^{-k}\{E[E(|\epsilon_i|^k|\mathbf{x}_i)]^2\}^{1/2}\{E[(\boldsymbol{\nu}\mathbf{x}_i)^4]\}^{1/2} \\ &\ge k_l - C_9\alpha^{-k}, \end{aligned}$$

where second inequality follows from Assumption III.3(i) and C_9 is a constant that only depends on k_0, k_2, d_1, M_k. Hence if $\alpha > (2C_9/k_l)^{1/k}$, we have $\lambda_{min}(\mathbf{D}_{\alpha,n}) > k_l/2$. It then implies $\liminf_{n\to\infty} \lambda_{min}(\mathbf{D}_{\alpha,n}) > 0$. Observe that

$$\begin{aligned} & |\boldsymbol{\nu}^T \textstyle\sum_{i=1}^n E_{\boldsymbol{\beta}^*}(\psi(\mathbf{x}_i, \boldsymbol{\beta}) - \psi(\mathbf{x}_i, \boldsymbol{\beta}^*)) - n\boldsymbol{\nu}^T\mathbf{D}_{\alpha,n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)| \\ = \; & |\boldsymbol{\nu}^T \textstyle\sum_{i=1}^n E_{\boldsymbol{\beta}^*}\{(l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{x}_i - l'_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i - l''_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{x}_i\mathbf{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}| \\ = \; & |\boldsymbol{\nu}^T \textstyle\sum_{i=1}^n E_{\boldsymbol{\beta}^*}\{(l''_\alpha(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})\mathbf{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\mathbf{x}_i - l''_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{x}_i\mathbf{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\}| \\ \le \; & |\boldsymbol{\nu}^T \textstyle\sum_{i=1}^n E_{\boldsymbol{\beta}^*}\{(k_3|\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||\mathbf{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\mathbf{x}_i|\}| \\ \le \; & k_3\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \textstyle\sum_{i=1}^n E_{\boldsymbol{\beta}^*}\{|\mathbf{x}_i^T\tilde{\boldsymbol{\nu}}|^2|\mathbf{x}_i^T\boldsymbol{\nu}|\}, \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is a vector lying between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\nu}} = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})/\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$. Note that the second equation follows from mean value theorem and the first inequality follows from the condition that l''_α is Lipschitz. By Assumption III.3 (iv) we have $\sum_{i=1}^n E_{\boldsymbol{\beta}^*}\{|\mathbf{x}_i^T\tilde{\boldsymbol{\nu}}|^2\mathbf{x}_i^T\boldsymbol{\nu}|\} = \mathcal{O}(n)$. Hence condition (iii) holds if $s/n \to 0$.

We conclude that the desired results hold for the oracle estimator $\hat{\boldsymbol{\beta}}_S^{\mathcal{O}}$. In particular, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}^* =& n^{-1}\sum_{i=1}^{n}\mathbf{D}_{\alpha,n}^{-1}l'_\alpha(\epsilon_i)\mathbf{x}_i + \mathbf{r}_n \\ =& n^{-1}\sum_{i=1}^{n}\{\mathbf{D}_{\alpha,n}^{-1}l'_\alpha(\epsilon_i)\mathbf{x}_i - E[\mathbf{D}_{\alpha,n}^{-1}l'_\alpha(\epsilon_i)\mathbf{x}_i]\} + E[\mathbf{D}_{\alpha,n}^{-1}l'_\alpha(\epsilon_i)\mathbf{x}_i] + \mathbf{r}_n,\end{aligned} \tag{A.67}$$

with $\|\mathbf{r}_n\|_2 = o_p(n^{-1/2})$. Observe that

$$\begin{aligned}\|E[\mathbf{D}_{\alpha,n}^{-1}l'_\alpha(\epsilon_i)\mathbf{x}_i]\|_2 &= \|\mathbf{D}_{\alpha,n}^{-1}E[l'_\alpha(\epsilon_i)\mathbf{x}_i]\|_2 \\ &= \|\mathbf{D}_{\alpha,n}^{-1}\tilde{\boldsymbol{\nu}}\|_2\|E[l'_\alpha(\epsilon_i)\mathbf{x}_i]\|_2 \\ &\leq [\lambda_{\min}(\mathbf{D}_{\alpha,n})]^{-1}\|E[l'_\alpha(\epsilon_i)\mathbf{x}_i]\|_2 \\ &= o(n^{-1/2}),\end{aligned} \tag{A.68}$$

where the last equality follows from equation (A.64). By equations (A.67) and (A.68), we obtain

$$\frac{\sqrt{n}}{\sigma_{\boldsymbol{\nu}}}\cdot\boldsymbol{\nu}^T(\hat{\boldsymbol{\beta}}_S^{\mathcal{O}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0,1), \tag{A.69}$$

where $\sigma_{\boldsymbol{\nu}}^2 = \boldsymbol{\nu}^T\mathbf{D}_{\alpha,n}^{-1}Var(l'_\alpha(\epsilon_i)\mathbf{x}_i)\mathbf{D}_{\alpha,n}^{-1}\boldsymbol{\nu}$. By Theorem III.3, the asymptotic result in (A.69) is also applicable for the stationary point $\tilde{\boldsymbol{\beta}}$. □

To prove Theorem A.3, we need the following result:

Lemma A.8. *Suppose covariate* $\mathbf{x}$ *satisfies Assumption III.3(iv) and* $l''_\alpha(u)$ *satisfies Assumption III.2(ii). For any fixed* $\alpha > 0$*, suppose the bound* $C_{14}k_0^2\left(\sqrt{\varepsilon_T} + \exp\left(-\frac{C_{15}T^2}{k_0^2r^2}\right)\right) < \frac{\gamma_{\alpha,T}}{\gamma_{\alpha,T}+k_2}\cdot\frac{k_l}{2}$ *holds, where* $\gamma_{\alpha,T} = \min_{|u|\leq T} l''_\alpha(u) > 0$*. Suppose in addition that the sample size satisfies* $n \geq C_{10}s\log p$*. With probability at least* $1 - C_{11}\exp(-C_{12}\log p)$*, the loss function* $\mathcal{L}_{\alpha,n}$ *satisfies*

$$\langle\nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_1) - \nabla\mathcal{L}_{\alpha,n}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle \geq \gamma_\alpha\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 - \tau_\alpha\frac{\log p}{n}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1^2, \tag{A.70}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ *such that* $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*\|_2 \leq r$ *for* $j = 1, 2$ *with*

$$\gamma_\alpha = \frac{\gamma_{\alpha,T} k_l}{16} \quad \textit{and} \quad \tau_\alpha = \frac{C_{13}(\gamma_{\alpha,T} + k_2)^2 k_0^2 T^2}{r^2}. \tag{A.71}$$

Here the constants $C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}$ *do not depend on* α.

Proof. The proof is similar to the proof of Proposition 2 in [Loh17]. Note that in that paper, it assumes $\mathbf{x}_i \perp\!\!\!\perp \epsilon_i$. However, a careful inspection of the proofs reveals that the result stills holds if we allow ϵ_i to depend on $\mathbf{x}_i$. We refer the reader to the arguments provided in that paper. □

Proof of Theorem A.3

Recall $\gamma_{\alpha,T} = \min_{|u| \leq T} l''_\alpha(u)$. By Assumption III.2(iii) , $\alpha \geq \alpha_0$ and $\alpha_0 = \max\{(2d_1)^{\frac{1}{k}}, 1\}$· T_0 we have

$$\gamma_{\alpha,T_0} = \min_{|u| \leq T_0} l_\alpha(u) \geq \min_{|u| \leq T_0} (1 - d_1 |u|^k \alpha^{-k}) \geq 1 - d_1 |T_0|^k \alpha_0^{-k} \geq \frac{1}{2}. \tag{A.72}$$

And

$$\gamma_{\alpha,T_0} = \min_{|u| \leq T_0} l_\alpha(u) \leq \min_{|u| \leq T_0} (1 + d_1 |u|^k \alpha^{-k}) \leq 1 + d_1 |T_0|^k \alpha_0^{-k} \leq \frac{3}{2}. \tag{A.73}$$

By equation (A.72), we obtain

$$\frac{\gamma_{\alpha,T_0}}{\gamma_{\alpha,T_0} + k_2} \cdot \frac{k_l}{2} \geq \frac{\frac{1}{2}}{\frac{1}{2} + k_2} \cdot \frac{k_l}{2} \geq \frac{k_l}{2 + 4k_2}.$$

Together with condition $C_{14} k_0^2 \left(\sqrt{\varepsilon_{T_0}} + \exp\left(-\frac{C_{15} T_0^2}{k_0^2 r^2} \right) \right) < \frac{k_l}{2+4k_2}$, we have

$$c_{14} k_0^2 \left(\sqrt{\varepsilon_{T_0}} + \exp\left(-\frac{c_{15} T_0^2}{k_0^2 r^2} \right) \right) < \frac{\gamma_{\alpha,T_0}}{\gamma_{\alpha,T_0} + k_2} \cdot \frac{k_l}{2}. \tag{A.74}$$

By equation (A.72), (A.73), (A.74) and Lemma A.8 we complete the proof. □

A.3. Proof in Chapter 4

Proof of Theorem IV.1

Since the proof of Theorem IV.1(i) is similar to the proof of Proposition 1 in [Loh17], we refer the reader to the arguments provided in that paper. Here we focus on the proof of (ii). We first suppose the existence of stationary points in the local RSC region and will establish this fact at the end of the proof. Suppose $\hat{\boldsymbol{\beta}}$ is a stationary point of program (IV.4) such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$. Since $\hat{\boldsymbol{\beta}}$ is a stationary point and $\hat{\boldsymbol{\beta}}$ is feasible, we have the inequality

$$\langle \nabla L_n(\hat{\boldsymbol{\beta}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}) + \lambda \mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle \geq \mathbf{0}, \tag{A.75}$$

where $\mathbf{D} := \mathrm{diag}((\sqrt{d_1}\mathbf{1}_{d_1}^T, \cdots, \sqrt{d_J}\mathbf{1}_{d_J}^T)^T)$ denotes a $p \times p$ diagonal matrix, $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{z}}_1^T, \cdots, \tilde{\boldsymbol{z}}_J^T)^T$ and $\tilde{\boldsymbol{z}}_j \in \partial \|\hat{\boldsymbol{\beta}}_j\|_2$. Recall

$$\partial \|\hat{\boldsymbol{\beta}}_j\|_2 = \begin{cases} \frac{\hat{\boldsymbol{\beta}}_j}{\|\hat{\boldsymbol{\beta}}_j\|_2} & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 \neq 0, \\ \{\boldsymbol{z} : \|\boldsymbol{z}\|_2 \leq 1, \boldsymbol{z} \in \mathbb{R}^{d_j}\} & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 = 0, \end{cases}$$

for $j = 1, 2, \cdots, J$. By the convexity of $\frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2 - q_\lambda(\boldsymbol{\beta})$, we have

$$\langle \nabla q_\lambda(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle \geq q_\lambda(\boldsymbol{\beta}^*) - q_\lambda(\hat{\boldsymbol{\beta}}) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2. \tag{A.76}$$

So together with inequality (A.75) we obtain

$$\langle \nabla L_n(\hat{\boldsymbol{\beta}}) + \lambda \mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle \geq q_\lambda(\boldsymbol{\beta}^*) - q_\lambda(\hat{\boldsymbol{\beta}}) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2.$$

Since $\langle \lambda \mathbf{D}\tilde{\boldsymbol{z}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle \leq \sum_{j=1}^{J} \sqrt{d_j}\lambda \|\boldsymbol{\beta}_j^*\|_2 - \sum_{j=1}^{J} \sqrt{d_j}\lambda \|\hat{\boldsymbol{\beta}}_j\|_2$, this means

$$\langle \nabla L_n(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle \geq \rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*) - \frac{\mu}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2. \tag{A.77}$$

Let $\tilde{\boldsymbol{\nu}} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. From the RSC inequality (IV.6), we have

$$\langle \nabla L_n(\hat{\boldsymbol{\beta}}) - \nabla L_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \geq \gamma \|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau \frac{\log p}{n} \|\tilde{\boldsymbol{\nu}}\|_1^2. \tag{A.78}$$

Combining inequality (A.78) with inequality (A.77), we then have

$$(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau \frac{\log p}{n} \|\tilde{\boldsymbol{\nu}}\|_1^2 + (\rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*)) \leq \langle \nabla L_n(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle. \tag{A.79}$$

So by Holder's inequality, we conclude that

$$(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 - \tau \frac{\log p}{n} \|\tilde{\boldsymbol{\nu}}\|_1^2 + (\rho_\lambda(\hat{\boldsymbol{\beta}}) - \rho_\lambda(\boldsymbol{\beta}^*)) \leq \|\nabla L_n(\boldsymbol{\beta}^*)\|_\infty \|\tilde{\boldsymbol{\nu}}\|_1. \tag{A.80}$$

Assume $\lambda \geq 4\|\nabla L_n(\boldsymbol{\beta}^*)\|_\infty$ and $\lambda \geq 8\tau R \frac{\log p}{n}$, we have

$$\begin{aligned}
(\gamma - \frac{\mu}{2})\|\tilde{\boldsymbol{\nu}}\|_2^2 &\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + (2R\tau \frac{\log p}{n} + \|L_n(\boldsymbol{\beta}^*)\|_\infty)\|\tilde{\boldsymbol{\nu}}\|_1 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \sum_{j=1}^{J} \sqrt{d_j}(2R\tau \frac{\log p}{n} + \|L_n(\boldsymbol{\beta}^*)\|_\infty)\|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \frac{1}{2}\sum_{j=1}^{J} \sqrt{d_j}\lambda \|\tilde{\boldsymbol{\nu}}_j\|_2 \\
&\leq (\rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}})) + \frac{1}{2}(\rho_\lambda(\tilde{\boldsymbol{\nu}}) + \frac{\mu}{2}\|\tilde{\boldsymbol{\nu}}\|_2^2),
\end{aligned}$$

implying that

$$0 \leq (\gamma - \frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \leq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}). \tag{A.81}$$

Recall $S \subseteq \{1, \cdots, J\}$ includes all indexes of important groups and $|S| = s$. By the assumption IV.1 for ρ, we have

$$\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) = \rho_\lambda(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_S) \geq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}_S),$$

where $\hat{\boldsymbol{\beta}}_S$ denotes the zero-padded vector in $\mathbb{R}^p$ with zeros on groups in S^c. Then starting from inequality (A.81), we have

$$\begin{aligned}
0 &\leq (\gamma - \frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \\
&\leq \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&= \rho_\lambda(\boldsymbol{\beta}^*) - \rho_\lambda(\hat{\boldsymbol{\beta}}_S) - \rho_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&\leq \rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}) \\
&= \frac{3}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) + \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) \\
&= \frac{3}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \frac{1}{2}\rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}).
\end{aligned} \tag{A.82}$$

Let A denote the group index set of the first s groups of $\boldsymbol{\nu}$ with largest ℓ_2 norm. Recall $d_a = \max_{1\leq j\leq J} d_j$, $d_b = \min_{1\leq j\leq J} d_j$, $d = \sqrt{\frac{d_a}{d_b}}$. By assumption IV.1(i) and (iv) we have

$$\begin{aligned}
0 \leq 3\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c}) &\leq 3\sum_{j\in S}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda) - \sum_{j\in S^c}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda) \\
&\leq 3\sum_{j\in A}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda) - \sum_{j\in A^c}\rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda).
\end{aligned} \tag{A.83}$$

Let $c := \max_{j\in A^c} \|\tilde{\boldsymbol{\nu}}_j\|_2$ and define $f(t,\lambda) := \frac{t\lambda}{\rho(t,\lambda)}$ for $t,\lambda > 0$. By assumption on ρ, for any fixed $\lambda \in \mathbb{R}^+$, function $t \mapsto f(t,\lambda)$ is non-decreasing on $\mathbb{R}^+$. Thus

$$\begin{aligned}\sum_{j\in A} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda)\cdot f(c,\sqrt{d_a}\lambda) &\le \sum_{j\in A} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_a}\lambda)\cdot f(\|\tilde{\boldsymbol{\nu}}_j\|_2,\sqrt{d_a}\lambda)\\ &\le \sum_{j\in A}\sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2.\end{aligned} \tag{A.84}$$

Similarly we also obtain

$$\begin{aligned}\sum_{j\in A^c} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda)\cdot f(c,\sqrt{d_b}\lambda) &\ge \sum_{j\in A^c} \rho(\|\tilde{\boldsymbol{\nu}}_j\|_2, \sqrt{d_b}\lambda)\cdot f(\|\tilde{\boldsymbol{\nu}}_j\|_2,\sqrt{d_b}\lambda)\\ &\ge \sum_{j\in A^c}\sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2.\end{aligned} \tag{A.85}$$

Combining inequality (A.83) with (A.84) and (A.85) we have

$$\begin{aligned}0 &\le 3\rho_\lambda(\tilde{\boldsymbol{\nu}}_S) - \rho_\lambda(\tilde{\boldsymbol{\nu}}_{S^c})\\ &\le \frac{1}{f(c,\sqrt{d_a}\lambda)}(3\sum_{j\in A}\sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{f(c,\sqrt{d_a}\lambda)}{f(c,\sqrt{d_b}\lambda)}\sum_{j\in A^c}\sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2)\\ &\le 3\sum_{j\in A}\sqrt{d_a}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{f(c,\sqrt{d_a}\lambda)}{f(c,\sqrt{d_b}\lambda)}\sum_{j\in A^c}\sqrt{d_b}\lambda\|\tilde{\boldsymbol{\nu}}_j\|_2\\ &= \sqrt{d_a}\lambda(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - \frac{\rho(c,\sqrt{d_b}\lambda)}{\rho(c,\sqrt{d_a}\lambda)}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2)\\ &\le \sqrt{d_a}\lambda(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - g(d)^{-1}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2),\end{aligned} \tag{A.86}$$

where the third inequality follows from

$$f(c,\sqrt{d_a}\lambda) \ge \lim_{r\to 0^+} f(r,\sqrt{d_a}\lambda) = \lim_{r\to 0^+}\frac{(r-0)\sqrt{d_a}\lambda}{\rho(r,\sqrt{d_a}\lambda)-\rho(0,\sqrt{d_a}\lambda)} = 1,$$

and the last inequality follows from assumption IV.1(ii). Hence,

$$3g(d)\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \ge \sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2,$$

Implying that

$$\begin{aligned}\|\tilde{\boldsymbol{\nu}}\|_1 &\leq \sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_1 + \sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_1 \\ &\leq \sum_{j\in A}\sqrt{d_a}\|\tilde{\boldsymbol{\nu}}_j\|_2 + \sum_{j\in A^c}\sqrt{d_a}\|\tilde{\boldsymbol{\nu}}_j\|_2 \\ &\leq \sqrt{d_a}(1+3g(d))\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \\ &\leq \sqrt{d_a s}(1+3g(d))\|\tilde{\boldsymbol{\nu}}\|_2.\end{aligned} \tag{A.87}$$

Combing inequalities (A.82) and (A.86) then gives

$$(\gamma-\frac{3\mu}{4})\|\tilde{\boldsymbol{\nu}}\|_2^2 \leq \frac{1}{2}\sqrt{d_a}\lambda(3\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 - g(d)^{-1}\sum_{j\in A^c}\|\tilde{\boldsymbol{\nu}}_j\|_2) \leq \frac{3}{2}\sqrt{d_a}\lambda\sum_{j\in A}\|\tilde{\boldsymbol{\nu}}_j\|_2 \leq \frac{3}{2}\sqrt{d_a s}\lambda\|\tilde{\boldsymbol{\nu}}\|_2,$$

from which we conclude that

$$\|\tilde{\boldsymbol{\nu}}\|_2 \leq \frac{6\sqrt{d_a}\lambda\sqrt{s}}{4\gamma-3\mu} \tag{A.88}$$

as wanted. Combining the ℓ_2-bound with inequality (A.87) then yields the ℓ_1 bound

$$\|\tilde{\boldsymbol{\nu}}\|_1 \leq \frac{6(1+3g(d))d_a\lambda s}{4\gamma-3\mu}. \tag{A.89}$$

Finally, in order to establish the existence of local stationary points, we simply define $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that

$$\hat{\boldsymbol{\beta}} \in \underset{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq r, \|\boldsymbol{\beta}\|_1 < R}{\operatorname{argmin}} \left\{L_n(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta})\right\}. \tag{A.90}$$

Then $\hat{\boldsymbol{\beta}}$ is a stationary point of program (A.90). So by the argument just provided, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{\frac{d_a s\log p}{n}}.$$

Provided $n > Cr^{-2}d_a s\log p$, the point $\hat{\boldsymbol{\beta}}$ will lie in the interior of the sphere of radius r around $\boldsymbol{\beta}^*$. Hence, $\hat{\boldsymbol{\beta}}$ is also a stationary point of the original program (IV.4) ,

guaranteeing the existence of such local stationary points. □

To prove Theorem IV.2, we need the following result adopted directly from the Lemma 1 in [Loh17].

Lemma A.9. *Suppose $\mathcal{L}_n$ satisfies the local RSC condition (IV.4) and $n \geq \frac{2\tau}{\gamma} k \log p$. Then $\mathcal{L}_n$ is strongly convex over the region $S_r := \{\boldsymbol{\beta} \in \mathbb{R}^p : supp(\boldsymbol{\beta}) \subseteq I_S, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$.*

Proof. The proof is similar to the proof of Lemma 1 in [Loh17]. □

Proof of Theorem IV.2

The proof is an adaptation of the arguments of Theorem 2 in the paper [Loh17]. We use the following three steps of the primal-dual witness (PDW) construction:

(i) Optimize the restricted program

$$\hat{\boldsymbol{\beta}}_{I_S} \in \underset{\boldsymbol{\beta} \in \boldsymbol{\beta}^{I_S} : \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j \in S} \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j}\lambda) \right\}, \tag{A.91}$$

and establish that $\|\hat{\boldsymbol{\beta}}_{I_S}\|_1 < R$.

(ii) Recall $q_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^J \sqrt{d_j}\lambda \|\boldsymbol{\beta}_j\|_2 - \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2 \sqrt{d_j}\lambda)$ defined in Section IV.2. Define $\hat{\boldsymbol{z}}_j \in \partial \|\hat{\boldsymbol{\beta}}_j\|_2$ and let $\hat{\boldsymbol{z}}_{I_S} = (\hat{\boldsymbol{z}}_j^T, j \in S)^T$, and choose $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_{I_S}^T, \hat{\boldsymbol{z}}_{I_S^c}^T)^T$ to satisfy the zero-subgradient condition

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}) + \lambda \mathbf{D} \hat{\boldsymbol{z}} = \mathbf{0}, \tag{A.92}$$

where $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_{I_S}, \mathbf{0}_{I_S^c})$ and $\mathbf{D} = \operatorname{diag}((\sqrt{d_1}\mathbf{1}_{d_1}^T, \cdots, \sqrt{d_J}\mathbf{1}_{d_J}^T)^T)$. Show that $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}_{I_S}^{\mathcal{O}}$ and establish strict dual feasibility: $\max_{j \in S^c} \|\hat{\boldsymbol{z}}_j\|_2 < 1$.

(iii) Verify via second order conditions that $\hat{\boldsymbol{\beta}}$ is a local minimum of the program (IV.4) and conclude that all stationary points $\hat{\boldsymbol{\beta}}$ satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$ are supported on I_S and agree with $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$.

Proof of Step (i) : By applying Theorem IV.1 to the restricted program (A.91), we have

$$\|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \leq \frac{6(1+3g(d))d_a\lambda s}{4\gamma - 3\mu},$$

and thus

$$\|\hat{\boldsymbol{\beta}}_{I_S}\|_1 \leq \|\boldsymbol{\beta}^*\|_1 + \|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \leq \frac{R}{2} + \|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_1 \leq \frac{R}{2} + \frac{6(1+3g(d))d_a\lambda s}{4\gamma - 3\mu} < R,$$

under the assumption of the theorem. This complete step (i) of the PDW construction. □

To prove step (ii), we need the following Lemma A.10 and A.11:

Lemma A.10. *Under the conditions of Theorem IV.2, we have the bound*

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 \leq C_3\sqrt{\frac{k\log k}{n}}$$

and $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$ *with probability at least* $1 - C_1\exp(-C_2\log k)$.

Proof. Recall $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}, \mathbf{0}_{I_S^c})$. By the optimality of the oracle estimator, we have

$$\mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}) \leq \mathcal{L}_n(\boldsymbol{\beta}^*). \tag{A.93}$$

Recall $n \geq \frac{2\tau}{\gamma}k\log p$. By Lemma A.9 $\mathcal{L}_n(\boldsymbol{\beta})$ is strongly convex over restricted region S_r. Hence,

$$\mathcal{L}_n(\boldsymbol{\beta}^*) + \langle\nabla\mathcal{L}_n(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\rangle + \frac{\gamma}{4}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 \leq \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}). \tag{A.94}$$

Together with inequality (A.93) we obtain

$$\begin{aligned}\tfrac{\gamma}{4}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2^2 &\leq \langle \nabla \mathcal{L}_n(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathcal{O}} \rangle \leq \|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_1 \\ &\leq \sqrt{k}\|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty \cdot \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2,\end{aligned}$$

implying that

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \leq \frac{4\sqrt{k}}{\gamma}\|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*))_{I_S}\|_\infty. \tag{A.95}$$

By applying Theorem IV.1 to the restricted program (A.91), we have

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S}\|_\infty = \|\nabla(\mathcal{L}_n(\boldsymbol{\beta}^*_{I_S}))\|_\infty \leq C_0 k_0 k_1 \sqrt{\frac{\log k}{n}} \tag{A.96}$$

with probability at least $1 - C_1 \exp(-C_2 \log k)$. Combining inequality (A.95) and (A.96), we obtain

$$\|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \leq C_3 \sqrt{\frac{k \log k}{n}} \tag{A.97}$$

as desired, where $C_3 = 4C_0 k_0 k_1 / r$.

Next we show $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$. When $n > C_3^2 k \log k / r^2$, we have $\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 < r$ and thus $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$ is an interior point of the oracle program in (III.12), implying

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) = \mathbf{0}. \tag{A.98}$$

By assumption that $\boldsymbol{\beta}^{*G}_{\min} \geq C_3 \sqrt{\frac{k \log k}{n}} + \sqrt{d_a}\delta\lambda$ and inequality (A.97), we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j\|_2 \geq \|\boldsymbol{\beta}^*_j\|_2 - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j - \boldsymbol{\beta}^*_j\|_2 &\geq \boldsymbol{\beta}^{*G}_{\min} - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}} - \boldsymbol{\beta}^*\|_2 \\ &\geq (C_3\sqrt{\tfrac{k\log k}{n}} + \sqrt{d_a}\delta\lambda) - C_3\sqrt{\tfrac{k\log k}{n}} \\ &= \sqrt{d_a}\delta\lambda.\end{aligned}$$

for all $j \in S$. Together with the assumption that ρ is (μ, δ)-amenable, we have

$$\nabla q_\lambda(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) = \lambda \mathbf{D}_{I_S I_S} \hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S}, \tag{A.99}$$

where $\hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S} = ((\hat{\boldsymbol{z}}^{\mathcal{O}}_j)^T, j \in S)^T$ and $\hat{\boldsymbol{z}}^{\mathcal{O}}_j \in \partial\|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_j\|_2$.

Combining equation (A.98) and (A.99), we obtain

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) - \nabla q_\lambda(\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}) + \lambda \mathbf{D}_{I_S I_S} \hat{\boldsymbol{z}}^{\mathcal{O}}_{I_S} = \mathbf{0}. \tag{A.100}$$

Hence $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$ satisfies the zero-subgradient condition on the restricted program (A.91). By step (i) $\hat{\boldsymbol{\beta}}_{I_S}$ is an interior point of the program (A.91), then it must also satisfy the zero-subgradient condition on the restricted program. Using the strict convexity from Lemma A.11, we obtain $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}$. □

The following lemma guarantees that the program in (A.91) is strictly convex:

Lemma A.11. *Suppose $\mathcal{L}_n$ satisfies the local RSC condition (IV.4) and ρ is μ-amenable with $\gamma > \mu$. Suppose in addition the sample size satisfies $n > \frac{2\tau}{\gamma - \mu} k \log p$, then the restricted program in (A.91) is strictly convex.*

Proof. This is almost identical to the proof of Lemma 2 in [LW+17]. We refer the reader to the arguments provided in that paper. □

Proof of step (ii) : We rewrite the zero-subgradient condition (A.92) as

$$\Big(\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\Big) + \Big(\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\Big) + \lambda \mathbf{D} \hat{\boldsymbol{z}} = \mathbf{0}.$$

Let $\hat{Q}$ be a $p \times p$ matrix $\hat{Q} = \int_0^1 \nabla^2 \mathcal{L}_n \Big(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\Big)\, dt$. By the zero-subgradient condition and the fundamental theorem of calculas, we have

$$\hat{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \Big(\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) - \nabla q_\lambda(\hat{\boldsymbol{\beta}})\Big) + \lambda \mathbf{D} \hat{\boldsymbol{z}} = \mathbf{0},$$

And its block form is

$$\begin{bmatrix} \hat{Q}_{I_S I_S} & \hat{Q}_{I_S I_S^c} \\ \hat{Q}_{I_S^c I_S} & \hat{Q}_{I_S^c I_S^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S} - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S}) \\ \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c} - \nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S^c}) \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{D}_{I_S I_S} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{I_S^c I_S^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{z}}_{I_S} \\ \hat{\boldsymbol{z}}_{I_S^c} \end{bmatrix} = \mathbf{0}. \tag{A.101}$$

The selection property implies $\nabla q_\lambda(\hat{\boldsymbol{\beta}}_{I_S^c}) = \mathbf{0}$. Plugging this result into equation (A.101) and performing some algebra, we conclude that

$$\hat{\boldsymbol{z}}_{I_S^c} = \frac{1}{\lambda} \mathbf{D}^{-1}_{I_S^c I_S^c} \left\{ \hat{Q}_{I_S^c I_S} (\boldsymbol{\beta}^*_{I_S} - \hat{\boldsymbol{\beta}}_{I_S}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c} \right\}. \tag{A.102}$$

Therefore,

$$\begin{aligned} \max_{j \in S^c} \|\hat{\boldsymbol{z}}_j\|_2 &\leq \max_{j \in S^c} \sqrt{d_j} \|\hat{\boldsymbol{z}}_j\|_\infty \\ &= \|\mathbf{D}_{I_S^c I_S} \hat{\boldsymbol{z}}_{I_S^c}\|_\infty \\ &= \tfrac{1}{\lambda} \|\hat{Q}_{I_S^c I_S} (\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty \\ &\leq \tfrac{1}{\lambda} \|\hat{Q}_{I_S^c I_S} (\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S})\|_\infty + \tfrac{1}{\lambda} \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty \\ &\leq \tfrac{1}{\lambda} \left\{ \max_{j \in I_S^c} \|\mathbf{e}_j^T \hat{Q}_{I_S^c I_S}\|_2 \right\} \|(\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S})\|_2 + \tfrac{1}{\lambda} \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty, \end{aligned} \tag{A.103}$$

where $\mathbf{e}_j$ is a standard unit vector with jth element being 1. Observe that

$$\begin{aligned} [(\mathbf{e}_j^T \hat{Q}_{I_S^c I_S})_m]^2 &\leq [\tfrac{1}{n} \textstyle\sum_{i=1}^n w(\mathbf{x}_i) \mathbf{x}_{ij} v(\mathbf{x}_i) \mathbf{x}_{im} \int_0^1 l''((y_i - \mathbf{x}_i^T \boldsymbol{\beta}^* - t(\mathbf{x}_i \hat{\boldsymbol{\beta}} - \mathbf{x}_i \boldsymbol{\beta}^*)) v(\mathbf{x}_i)) dt]^2 \\ &\leq k_2^2 [\tfrac{1}{n} \textstyle\sum_{i=1}^n w(\mathbf{x}_i) \mathbf{x}_{ij} \cdot v(\mathbf{x}_i) \mathbf{x}_{im}]^2, \end{aligned}$$

for all $j \in I_S^c$ and $m \in I_S$, where the last inequality follows from assumption III.2(ii). By conditions of Theorem IV.2, the variables $w(\mathbf{x}_i)\mathbf{x}_{ij}$ and $v(\mathbf{x}_i)\mathbf{x}_{im}$ are both sub-Gaussian. Using standard concentration results for i.i.d sums of products of sub-Gaussian variables, we have

$$P([(e_j^T \hat{Q}_{I_S^c I_S})_m]^2 \leq C_3') \geq 1 - C_2' \exp(-C_3' n).$$

It then follows from union inequality that

$$P(\max_{j\in I_S^c}\|e_j^T\hat{Q}_{I_S^c I_S}\|_2 \leq \sqrt{C_3' k}) \geq 1-C_2'\exp(-C_3' n+\log(k(p-k))) \geq 1-C_2'\exp(-\frac{C_3'}{2}n), \tag{A.104}$$

where $n \geq \frac{2}{C_3'}\log(k(p-k))$. By Lemma A.10 we obtain

$$\|\hat{\boldsymbol{\beta}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 \leq C_3\sqrt{\frac{k\log k}{n}}. \tag{A.105}$$

Furthermore, Theorem IV.1 gives

$$\|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)_{I_S^c}\|_\infty \leq \|\nabla\mathcal{L}_n(\boldsymbol{\beta}^*))\|_\infty \leq C_1\sqrt{\frac{\log p}{n}}. \tag{A.106}$$

Combining inequality (A.103), (A.104), (A.105) and (A.106), we have

$$\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2 \leq \frac{1}{\lambda}C_4\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - C_5\exp(-C_2\log k)$, under the assumption that $k^2\log k = \mathcal{O}(\log p)$. In particular, for $\lambda > C_4\sqrt{\frac{\log p}{n}}$, we conclude at last that the strict dual feasibility condition $\max_{j\in S^c}\|\hat{\boldsymbol{z}}_j\|_2 < 1$ holds, completing step (ii) of the PDW construction.

Step (iii) : Since the proof for this step is almost identical to the proof in Step (iii) of Theorem 2 in [Loh17], except for the slightly different notations. We refer the reader to the arguments provided in that paper. □

Proof of Theorem IV.3

By the condition that $\boldsymbol{\beta}^{*I}_{\min} \geq C_3\sqrt{\frac{s\log s}{n}} + \theta$, we have

$$\begin{aligned} |\hat{\beta}^{\mathcal{O}}_j| \geq |\beta^*_j| - |\hat{\beta}^{\mathcal{O}}_j - \beta^*_j| \quad &\geq \boldsymbol{\beta}^{*I}_{\min} - \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_\infty \\ &\geq (C_3\sqrt{\tfrac{k\log k}{n}} + \theta) - C_3\sqrt{\tfrac{k\log k}{n}} \\ &= \theta. \end{aligned} \tag{A.107}$$

for all $j \in I_0$, where the second inequality follows from Lemma A.10. For $j \in I_S - I_0$,

$$|\hat{\beta}^{\mathcal{O}}_j| \leq \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_\infty \leq C_3\sqrt{\frac{k\log k}{n}} < \theta, \tag{A.108}$$

where the second inequality follows from Lemma A.10 and the last inequality follows from the condition in Theorem IV.3. Recall $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S}, \mathbf{0}_{I_S^c})$. By Theorem IV.2 we have $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\mathcal{O}}$ with probability at least $1 - C_5\exp(-C_2\log k)$. Together with (A.107) and (A.108), we have

$$\hat{\boldsymbol{\beta}}^h(\theta) = \hat{\boldsymbol{\beta}} \cdot I(|\hat{\boldsymbol{\beta}}| \geq \theta) = \hat{\boldsymbol{\beta}}^{\mathcal{O}} \cdot I(|\hat{\boldsymbol{\beta}}^{\mathcal{O}}| \geq \theta) = (\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_0}, \mathbf{0}_{I_0^c}),$$

as desired. It then gives the result

$$\|\hat{\boldsymbol{\beta}}^h(\theta) - \boldsymbol{\beta}^*\|_2 \leq \|\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{I_S} - \boldsymbol{\beta}^*_{I_S}\|_2 \leq C_3\sqrt{\frac{k\log k}{n}},$$

where the last inequality follows from Lemma A.10. □