LUO, BIN, M.A. Robust High-dimensional Data Analysis Using A Weight Shrinkage Rule. (2016)
Directed by Dr. Xiaoli Gao. 73 pp.

In high-dimensional settings, a penalized least squares approach may lose its efficiency in both estimation and variable selection due to the existence of either outliers or heteroscedasticity. In this thesis, we propose a novel approach to perform robust high-dimensional data analysis in a penalized weighted least square framework. The main idea is to relate the irregularity of each observation to a weight vector and obtain the outlying status data-adaptively using a weight shrinkage rule. By usage of ℓ_1-type regularization on both the coefficients and weight vectors, the proposed method is able to perform simultaneous variable selection and outliers detection efficiently. Eventually, this procedure results in estimators with potentially strong robustness and non-asymptotic consistency. We provide a unified link between the weight shrinkage rule 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. These theoretical results allow the number of variables to far exceed the sample size. The performance of the proposed estimator is demonstrated in both simulation studies and real examples.

ROBUST HIGH-DIMENSIONAL DATA ANALYSIS USING A WEIGHT SHRINKAGE RULE

by

Bin Luo

A Thesis 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
Master of Arts

Greensboro
2016

Approved by

Committee Chair

APPROVAL PAGE

This thesis 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

_____
Scott Richter

_____
Haimeng Zhang

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION AND BACKGROUND

1.1 High-dimensional Data Analysis

In traditional statistical methodology, we assume that there are many observations and each observation is a vector of values we measure on a few well-chosen variables. Informally, if we let n denote the number of observation and let p denote the number of variables, the traditional statistical methodologies and application has been largely limited to the 'small p, large n' scenario.

Due to the rapid development of advanced technologies over the last decades, however, it has become much cheaper to collect a large amount of data. The trend is towards more observations but even radically larger numbers of variables. Observations with curves, images or movies, along with many other variables, are frequently seen in contemporary scientific research and technological development. Therefore a single observation has dimension in the thousands or billions, while there are only tens or hundreds of instances available for study. We described this key scenario as 'large p, small n' [D$^+$00]. 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 [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].

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 square (PLS) methods have become very popular in high-dimensional linear regression analysis since the introduction of the LASSO [Tib96]. 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. Both theoretical and computational properties of PLS with LASSO-type penalties and some concave penalties have been widely investigated. See for example the LS-LASSO and its properties in [CT07, Zou06, ZY06, MB06], and the LS-SCAD in [FL01, FP$^+$04, XH09]. One can refer [ZZ14] for a complete review.

1.2 Data Contamination

Statistical inference is based on two sources of information: empirical data and assumptions which are presented in the form of statistical model. Naive interpretation of statistics derived from data sets that include data contamination may be

misleading. 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. Outliers can have many anomalous causes: changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations [Wik16b]. In some cases, the contaminated data also exhibit certain heteroscedasticity, when among sub-populations there exists different variabilities which could be quantified by the variance or any other measures of statistical dispersion [Wik16a]. 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. In wage regression in econometrics, more working experience often arises a larger variance in wage.

Outliers detection 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. In traditional statistics, most often the concepts of proximity is used to find outliers based on their relationship to the rest of the data. Due to sparsity of data in high dimensional space, however, the idea of proximity fails to maintain its meaningfulness since nearly every point can be treated as good outliers from that perspective [AY01]. Other traditional outliers detection methods include some graphical tools, such as normal probability plots and residuals plots, and some diagnostics statistics [Coo77, Pop76, VR13]. However, these methods can fail due to the occur of multiple outliers. Two phenomena had been noted in outliers detection: masking and swamping. Masking occurs if an outlier is not be detected, and swamping occurs if a good observation is considered as an outlier.

Many robust analysis tools were proposed in low-dimensional data analysis to deal with the data contamination. For example, robust regression with low breakdown values such as the least absolute deviation (LAD) estimator [RL05], robust regression with high breakdown values such as the repeated median estimate [Sie82], the least median squares (LMS) [Rou84], the least trimmed squares(LTS) [Rou84], the S-estimate [RY84], the MM-estimator [YZ88] , among many others. Here the breakdown value measures the smallest amount of contamination that can have an arbitrarily large effect on an estimator. Another line of robust analysis focuses on simultaneous outliers detection and robust estimation. For example, [SO12] proposed an adaptive approach by shrinking those additional mean shift parameters to zero under a mean shift model framework. [AR+13] and [ARC10] used the forward search (FS) to search the outliers individually.

Most above mentioned robust models have been extended in high-dimensional data settings by incorporating some LASSO-type penalties into robust regression. For example, LAD-LASSO uses an (adaptive) LASSO penalty under the ℓ_1 loss [GH10, WLJ07, Wan13], sparse LTS uses an (adaptive) LASSO penalty under the least trimmed squares loss [ACG+13], MMNNG uses a non-negative garrote penalty in MM regression [GV15]. Some other sparse robust models include penalized exponential square loss regression [WJHZ13], MM-Bridge [SY15], Robust Lars in [KVAZ07], MM-ridge or S-ridge in[SY15], and a sparse outlier shrinkage(SROS) model in [XJ13]. One can also refer [WM15] for a selective review on other robust HDDA. More detail regarding these topics would be discussed in Chapter 3.

1.3 Real Example

We introduce two real data examples in this section. One is the air pollution data collected from 60 Standard Metropolitan Statistical Areas in the United States, which is corresponding to a low-dimensional case ($p < n$); The other is the NTC-60 data, a gene expression data set collected from Affymetrix HG-U133A chip, which is corresponding to a high-dimensional case ($p > n$).

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 method is needed for regression analysis on the air pollution data.

As to the NCI-60 dataset, it 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 study, considering the possible data contamination in the dataset, the robust high-dimensional data analysis approaches are applied.

1.4 Objective

Most of above mentioned sparse robust HDDA models do not identify outliers in particular, which themselves can provide important scientific findings. For example, in e-Commerce business, one is interested in predicting the product prices in gray market. However, some sellers marked down their products dramatically compared with others. It would be interest to identify those sellers and check whether they are selling fake products. As we discussed in Section 1.2, suspected outliers could be identified using visualizing tools such as studentized-residuals plot and Cook's distance plot [Wei05]. However, when there are multiple outliers, these simple methods can fail, because of two phenomena, masking and swamping. These two phenomena were demonstrated by examining a famous artificial dataset, the Hawkins-Bradu-Kass (HBK) data [HBK84] in [SO12]. The occurrence of multiple outliers in high-dimensional setting could be more common in a big data world. For a HDDA method with separate outliers detection and variable selection process, the damage of high-dimensionality and data contamination can be intertwined. On the one hand, if some

potential outliers are masked or some normal observations are swamped such that the size of valid samples is even smaller, the variable selection results can be invalid. On the other hand, if some covariate are incorrectly selected or non-selected, then potential outliers among those covariates will be either swamped or masked. Therefore in HDDA a robust procedure with simultaneously variable selection, outliers detection and coefficient estimates is highly demanded.

Due to the above challenges, our objective is to propose a method that performs simultaneous variables selection, coefficient estimation and outliers detection. We expect this method to be computationally efficient in high-dimensional analysis and to be data-adaptive.

1.5 Propose Method Framework

In this thesis, we propose to perform robust HDDA, outliers detection and robust regression in a penalized weighted least squares framework. To be more specific, suppose we have data $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')$, where y_i is the observed response variable and $\mathbf{x}_i$ is a $p-$dimensional covariates vector. Let $\mathbf{L}(\beta, \mathbf{w}; \mathbf{y}, \mathbf{X})$ denote the weighted loss function for the data $(\mathbf{y}, \mathbf{X})$ with some $p-$dimensional parameter $\beta = (\beta_1, \ldots, \beta_p)'$ and $n-$dimensional weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, we solve

$$(\tilde{\beta}, \tilde{\mathbf{w}})(\lambda_1, \lambda_2) = \underset{\beta \in \mathbb{R}^p, 0 < w_i \leq 1}{\arg\min} \{\mathbf{L}(\beta, \mathbf{w}; \mathbf{y}, \mathbf{X}) + \mathbf{P}_{\lambda_1}(\beta) + \mathbf{P}_{\lambda_2}(\mathbf{w})\}, \tag{1.1}$$

where $\mathbf{P}_{\lambda_1}$,$\mathbf{P}_{\lambda_2}$ are called the penalty function. In (1.1), we introduce a shrinkage rule for the weight vector in a penalized weighted least squares framework to perform simultaneous outliers detection, variable selection and robust estimation. 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. 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 this model as the PAWLS method in general since the weighted least square model is considered and a penalization approach is linked to the proposed weight shrinkage rule.

The rest of the thesis is organized as follows. In chapter 2, we introduce the penalized regression method, especially the LASSO [Tib96], adaptive LASSO [Zou06] and ridge regression [HK70]. The robust regression methods are discussed in chapter 3, along with more detailed introduction on M-estimator [H$^+$64] and least trimmed squared (LTS) estimator [Rou84]. In chapter 4, we introduce the PAWLS model, including the theoretical properties, implementation and some numerical result. A brief discussion is given in chapter 6. The technical proofs are relegated to the Appendix.

CHAPTER II

PENALIZED REGRESSION METHOD

2.1 Introduction

Let $\mathbf{L}(\beta; \mathbf{y}, \mathbf{X})$ be the negative log-likelihood function of the data $(\mathbf{y}, \mathbf{X})$ with parameter vector $\beta = (\beta_1, \ldots, \beta_p)'$. The maximum likelihood estimator (MLE) is $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \{\mathbf{L}(\beta; \mathbf{y}, \mathbf{X})\}$. It is well known that MLE possesses the properties of consistency, asymptotic normality and efficiency. However, there exists certain scenarios that we would like to introduce the penalty term to the likelihood function to achieve better estimation, which is called penalized regression. For example, sometimes we are willing to reduce estimation variances by scarifying some biases. In other cases we might have lots of variables in our model, where standard regression can easily be overfitting. Dependent on the form of penalty, the penalty can help to do the variable selection as well as shrinkage of estimator. In penalized regression, we solve

$$\tilde{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \{\mathbf{L}(\beta; \mathbf{y}, \mathbf{X}) + \mathbf{P}_\lambda(\beta)\} \tag{2.1}$$

where $\mathbf{P}_\lambda(\beta)$ is called penalty function and λ is a tuning parameter in the penalty. The form of $\mathbf{P}_\lambda(\beta)$ 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 maximum likelihood objective function.

It is well known that when the random error is normal, the least square estimator is a MLE. In high-dimensional linear regression analysis, penalized least square(PLS)

methods have become very popular among all the penalized regression methods. A PLS method adopts an ℓ_2 loss function $\mathbf{L}(\beta; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n}(y_i - \mathbf{X}_i'\beta)^2$ and a penalty on the coefficient β. Perhaps the most popular penalty function used for PLS is the LASSO-type penalty [KF00], where we define penalty function $\mathbf{P}_\lambda(\beta) = \lambda \sum_{j=1}^{P} |\beta_j|^\gamma$. Hence in LASSO-type PLS we estimate β by minimizing least squares criterion

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2 + \lambda_n \sum_{j=1}^{P} |\beta_j|^\gamma, \tag{2.2}$$

where $\gamma > 0$ and λ_n is the tuning parameter. Such estimators called Bridge estimators were introduced in [FF93] as a generalization of ridge regression($\gamma = 2$). The special case when $\gamma = 1$ is related to the least absolute shrinkage and selection operator (LASSO) [Tib96], which is a very popular shrinkage method for variable selection. When $\gamma \leq 1$, the component of β in (2.2) can be shrunk to zero if λ_n is sufficiently large, thus achieving simultaneous coefficient estimation and variable selection. Considering the limiting cases of Bridge estimation as $\gamma \to 0$, since

$$\lim_{\gamma \to 0} \sum_{j=1}^{p} |\beta_j|^\gamma = \sum_{j=1}^{p} \mathbf{I}(\beta_j \neq 0), \tag{2.3}$$

it can be viewed as a model selection method that penalizes the number of variables in the model (such as AIC and BIC [LZ86]).

Since LASSO does not produce consistent variable selection results, some other concave penalties were introduced. [FL01] proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty. Unlike the LASSO penalty, SCAD penalty functions have flat tails that reduce the biases and the estimator possesses good properties: consistency of variable selection and asymptotic normality, which are also called the

oracle property. Adaptive LASSO [Zou06] is another variable selection technique that enjoys the oracle property. It adopts the weighted penalized term $\lambda \sum_{j=1}^{p} w_j|\beta_j|$, instead of the LASSO penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$. The minimax concave penalty (MCP) [Zha07] is another non-convex penalty with the oracle property.

2.2 Ridge Regression

Considering the linear regression model

$$y_i = \mathbf{x}_i'\beta + \epsilon_i, 1 \leq i \leq n, \tag{2.4}$$

where y_i and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ are the observed response variable and covariates vector, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the coefficient vector and $\epsilon_1, \ldots, \epsilon$ are i.i.d. random variables with mean 0 and variance σ^2.

The ordinary least squared OLS estimator $\hat{\beta}^{ols}$ are given by

$$\hat{\beta}^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}. \tag{2.5}$$

The solution $\hat{\beta}^{ols}$ are unbiased with variance $Var(\hat{\beta}^{ols}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

In practice, for example, when analyzing economic or medical data, the predictor covariates in the columns of $\mathbf{X}$ may have a high level of collinearity, which means there may be a nearly linear relationship among the predictor covariates. In this case, $\mathbf{X}'\mathbf{X}$ in (2.5) is nearly singular and difficult to evaluate. Meanwhile, the ill-conditioning in $\mathbf{X}'\mathbf{X}$ caused by the dependency among the columns of $\mathbf{X}$ results in large variance of OLS solutions with inflated squared lengths $\| \hat{\beta}^{ols} \|^2$ and $\hat{\beta}^{ols}$ being very sensitive to small changes in $\mathbf{X}$.

[HK70] proposed ridge regression to improve the estimates. Using the same notation in (2.2), the ridge regression estimators are the solution of

$$\hat{\beta}^{ridge}(\lambda_n) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + \lambda_n \sum_{j=1}^{P} \beta_j^2 \right\}, \tag{2.6}$$

which is given by

$$\hat{\beta}^{ridge}(\lambda_n) = (\mathbf{X}'\mathbf{X} + \lambda_n \mathbf{I}_p))^{-1}\mathbf{X}\mathbf{y}. \tag{2.7}$$

Here λ_n is a tuning parameter that controls the strength of the penalty term. From (2.6) we learn that ridge regression estimators minimizes the sum of squared residuals plus a penalty term on the squared ℓ_2 norm of the coefficient vector. Thus it shrinks all coefficients towards zero simultaneously. An equivalent statement of (2.6) is: if the squared lengths of coefficient vector β is fixed to certain amount (controlled by λ_n), then $\hat{\beta}^{ridge}(\lambda_n)$ is the value of β that gives a minimum sum of squares. Hence it prevents the length of the estimator vector from being inflated. More importantly, although the shrinkage introduced in (2.6) produces some bias on estimates, it can greatly reduce the variance, resulting in a better mean-squared error [HK70].

2.3 LASSO

Considering the linear model in (2.4), the LASSO can be specified as estimating the coefficient $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ by minimizing the residual sum of square, subject to the constraint on sum of the absolute value of the regression coefficient $\sum_{j=1}^{p} |\beta_j| \leq$

s, which is equivalent to (2.2) with $\gamma = 1$,

$$\hat{\beta}^{lasso}(\lambda_n) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2 + \lambda_n \sum_{j=1}^{P} |\beta_j|, \tag{2.8}$$

where λ_n is a nonnegative tuning parameter. The LASSO is able to continuously shrink the coefficient toward 0 as λ_n increase and some coefficients are shrunk to exactly 0 if λ_n is sufficiently large. Thus it is a regularization technique for simultaneous estimation and variable selection. Due to the bias-variance trade-off, the prediction accuracy is often improved by the continuous shrinkage method.

Fig. 1 provides some insight about why the lasso can produce coefficients that are exactly zeros, while ridge regression cannot, for the case $p = 2$. The red ellipses are the contours of the sum of residuals square. They are centered at the OLS estimates. The solid blue areas are the constraint regions, with $|\beta_1| + |\beta_2| \leq s$ for the LASSO and $\beta_1^2 + \beta_2^2 \leq s$ for the ridge regression. Fig. 1(a) indicates that the LASSO solution is the first place that the contours touch corner of the square yielding a zero coefficients; whereas there is no corner for ridge regression in 1(b) and thus zero coefficients will rarely occur.

Compared to the classical variable selection methods such as subset selection, the LASSO is more stable due to its continuity; Moreover, the LASSO is computationally feasible for high-dimensional data, while the computation in subset selection is combinatorial and not feasible in high-dimensional setting.

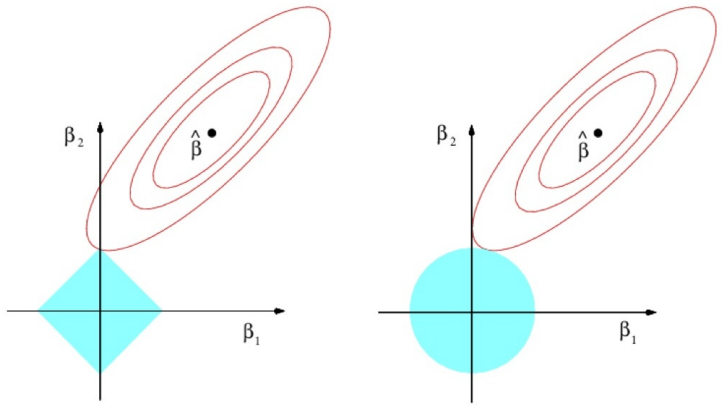
Here is a Bayesian understanding of LASSO when we consider (2.4) with $\epsilon \sim N(0, \sigma^2)$. Suppose we have independent prior distributions: $\beta_0 \propto 1, \pi(\beta_j) \propto \exp^{-\lambda|\beta_j|}$

for $1 \leq j \leq p$, $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$. Note the likelihood function can be specified as

$$p(\mathbf{y}|\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - \mathbf{x}_i'\beta)^2}{2\sigma^2} \right). \tag{2.9}$$

Thus the joint posterior distribution of the parameter is equal to

Figure 1. Estimation Picture for (a) The Lasso and (b) Ridge Regression [Tib96]



$$(\beta, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-n/2-1} \exp\{-\lambda \sum_{i=1}^{n} |\beta_j|\} \exp\left\{ - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 \right\}. \tag{2.10}$$

Then the mode, $(\hat{\beta}, \hat{\sigma}^2)$, of the above posterior distribution is

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + 2\sigma^2\lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.11}$$

$$\hat{\sigma}^2 = \frac{1}{n+2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2. \tag{2.12}$$

Letting $\lambda_n = 2\sigma^2\lambda$, minimization problem of (2.11) is equivalent to that of (2.8).

[FL01] proposed that a good variable selection method should satisfy the oracle properties. Suppose β^* is the true coefficient in (2.4). Let $\mathbf{A} = \{j : \beta_j^* \neq 0\}$ and further assume that $|\mathbf{A}| = p_0 < p$, which means the true model only depends on a subset of the predictors. Let $\hat{\beta}$ an estimator obtained by a fitting procedure and we call it an oracle procedure if $\hat{\beta}$ (asymptotically) owns the following oracle properties: (1) Consistency of variable selection, $\{j : \hat{\beta}_j \neq 0\} = \mathbf{A}$; (2) Asymptotic normality, $\sqrt{n}(\hat{\beta}_{\mathbf{A}} - \beta_{\mathbf{A}}^*) \to_d \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}^*)$, where $\mathbf{\Sigma}^*$ is the covariance matrix knowing the true subset model. However, [FL01] and [Zou06] show that LASSO does not follow the oracle properties. First, LASSO has been shown to only perform consistent variable selection under so-called irrepresentable condition [ZY06], which is non-trivial conditions that many data sets in practice may not satisfy; On the other hand, LASSO tends to underestimate those important variables. To fix this problem, [Zou06] proposed the adaptive LASSO, in which adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. The minimax concave penalty (MCP) [Zha07] and the Smoothly Clipped Absolute Deviation (SCAD) penalty [FL01] also possess the oracle property.

Compared with Ridge regression, one disadvantages of the LASSO is: when there exits multicollinearity among the explanatory variable, the LASSO is more likely to select only a single variable from a group of highly correlated variables. To overcome this limitation, Elastic net [ZH05] adopts a penalty function with convex combination of ℓ_1 and ℓ_2 to combine the advantages from both LASSO and Ridge regression.

2.4 Adaptive LASSO

As mentioned earlier, adaptive LASSO assign different weights to different coefficients. Suppose that $\tilde{\beta}$ is a consistent estimate of β^*, such as $\hat{\beta}^{ols}$. The adaptive

LASSO estimator is given by

$$\hat{\beta}^{alasso}(\lambda_n) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + \lambda_n \sum_{j=1}^{p} \frac{1}{|\tilde{\beta}|} |\beta_j| \right\}. \tag{2.13}$$

Note that for any fixed λ, the penalty for zero-initial estimation go to infinity, while the 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.

For fixed p, [Zou06] proved 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 obtain variable selection consistency and estimation efficiency, when the marginal regression estimators are used as the initial estimators.

Similar to (2.8) for the LASSO, (2.13) is also a convex optimization problem thus its global minimizer can be efficiently solved. Since it is an ℓ_1 penalization method, the current efficient algorithm for solving lasso can also be used to compute the adaptive LASSO estimates.

2.5 Tuning Parameter Selection

For LASSO and other penalized regression methods, a tuning parameter λ is adopted to control the size of model. Thus the selection of λ plays an important role. One of the most common criterion for selecting the tuning parameter is Cross validation (CV) [FHT01]. The goal is to find the model with best predictive performance. In CV procedure, first we randomly divide the data set into K parts with roughly

the same size m. Then we consider each single part as the validation data denoted by $\mathbf{x}_k$ and $\mathbf{y}_k$ ($k \in \{1, \ldots, K\}$), and the other $K-1$ parts as the training data denoted by $\mathbf{x}_{-k}$ and $\mathbf{y}_{-k}$. For a specific λ, we fit the model with training data and applied it to the validation set to obtain the prediction of $\mathbf{y}_k$ as $\hat{\mathbf{y}}_k(\lambda)$. The average prediction performance can be evaluated by

$$\mathbf{P_{CV}} = \frac{1}{K}\mathbf{P}(\mathbf{y}_k, \hat{\mathbf{y}}_k(\lambda)), \tag{2.14}$$

where the function $\mathbf{P}(\mathbf{y}_k, \hat{\mathbf{y}}_k(\lambda))$ is a certain metric of the prediction accuracy. The mean squared prediction error is often used as the metric of the prediction accuracy, which is defined as

$$\mathbf{P}(\mathbf{y}_k, \hat{\mathbf{y}}_k(\lambda)) = \frac{1}{m}(\mathbf{y}_k - \hat{\mathbf{y}}_k(\lambda))'(\mathbf{y}_k - \hat{\mathbf{y}}_k(\lambda)), \tag{2.15}$$

For selecting λ we aim to find an optimal parameter that minimizes the averaged prediction error in (2.14).

However, a tuning parameter chosen by the cross validation often leads to a model with too many non-zero effects [Xu07]. Therefore an alternative criterion for tuning parameter selection is Bayesian information criterion (BIC). In this case, we are interested in finding the model with both accurate predictions and identifying the true model structure. The BIC criterion tries to find λ that minimizes the following score function

$$BIC = n \log \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\hat{\beta})^2 + \log(n) df(\lambda), \tag{2.16}$$

where $df(\lambda)$ is the degree freedom of model for a specific λ. It can be estimated by the number of non-zero regression coefficients estimated from LASSO. From (2.16) we can consider BIC as a compromise between model fitting and model complexity. For high-dimensional data, [CC08] claimed that BIC still tends to select a model with many covariates and they proposed an extended Bayesian information criterion (EBIC) which can obtain more aggressive variable selection.

CHAPTER III

ROBUST METHOD

3.1 Introduction

Regression methods are widely used for prediction and theoretical explanation in education, psychology, sociology, medication, economics and others. It is an approach for modeling the relationship between one dependent variable and one or more explanatory variables. In statistics, for proper interpretation of data analysis, regression methods make a number of assumptions about the predictors, the response variables and their relationship. Perhaps the most popular statistical regression methods is ordinary least square (OLS) regression, of which the assumptions include normality, equal variance and independence of random errors. Problems occur when these assumptions are not satisfied. When there exists data contamination, such as heavy-tailed errors or outliers in datasets, the assumptions of normality and equal variance of errors may be violated. In these situations OLS approach can produce unstable prediction estimates and yield sensitive results. Outliers can occur in the x-axis direction (called leverage points), the y-axis direction, or both axes directions simultaneously. The impact of the outliers on estimation of regression coefficients can be varying depending on where the outliers occur. Compared with outliers in y-axis, an outliers in x-axis direction may exhibit more influence.

To deal with data contamination, it is nature to detect and remove the outliers before fitting any classical regression models. Some statistical methods for outliers diagnostics were developed in the last decades. One statistical method for detect-

ing outliers in multivariate case is to compute Mahalanobis distance [DSP66], also known as 'diagonal of the hat matrix'. Although Mahalanobis distance is a common measure of leverage in regression, it fails when the outliers are at y-axis, or leverage outliers are masked by effect of other leverage point in data. Other statistical methods for outliers 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]. [RL05] points out that these statistics do not work well in locating the joint influences of multivariate outliers. Another way for detecting outliers are graphical methods based on plotting residuals [RL05] . However, although graphical diagnostics procedures can be helpful in certain situation, it is less helpful when x-axis outliers occur.

Instead of deleting the outliers before fitting statistical model, many robust regression has been proposed to accommodate them. In low dimension setting, for example, least absolute deviation (LAD) estimator [RL05] estimates the coefficients by minimizing the sum of absolute value of residuals. It can be useful when OLS fail to produce a reliable estimator in presence of outliers. However, LAD estimator is neither a bounded-influence nor a high breakdown point estimator [AS03]. Here the breakdown value measures the smallest amount of contamination that can have an arbitrarily large effect on an estimator. The least median squares (LMS) [Rou84] estimator can be considered as being similar to OLS except that the median value instead of the mean value if used. LMS is a high breakdown procedure but without high relative efficiency. The least trimmed squares(LTS) [Rou84] is similar to OLS except that the largest squared residuals are exclude from the summation. LTS is considered

to be a high breakdown method and it can be very efficients in certain situations. M-estimators [H$^+$64] is developed based on the idea of replacing the squared residuals in OLS with another function of residuals. M-estimators is statistically more efficient than LAD and is robust against y-axis outliers. However, it is not robust to x-axis outleirs. Other robust regression methods with high breakdown value include the repeated median estimate [Sie82], the S-estimate [RY84], the MM-estimator [YZ88], among many others.

Another line of robust analysis focuses on simultaneous outlier detection and robust estimation. For example, [SO12] proposed an adaptive approach by shrinking those additional mean shift parameters to zero under a mean shift model framework. [AR$^+$13] and [ARC10] used the forward search (FS) to search the outliers individually.

Most above mentioned robust models have been extended in high-dimensional data settings by incorporating some LASSO-type penalties into the robust regression. For example, LAD-LASSO uses an (adaptive) LASSO penalty under the ℓ_1 loss [GH10], [WLJ07], [Wan13], sparse LTS uses an (adaptive) LASSO penalty under the least trimmed squares loss [ACG$^+$13], MMNNG uses a non-negative garrote penalty in MM regression [GV15]. Some other sparse robust models include penalized exponential square loss regression [WJHZ13], MM-Bridge [SY15], Robust Lars in [KVAZ07], MM-ridge or S-ridge in[SY15], and a sparse outlier shrinkage(SROS) model in [XJ13]. One can also refer [WM15] for a selective review on other robust HDDA.

A more detail discuss regarding M-estimator and LTS are provided in the rest of this chapter.

3.2 M-estimator

Let r_i denote the residual of the i^{th} observation, the difference between the observed value and the fitted value. For example, in (2.4) we have $r_i = y_i - \mathbf{x}_i'\beta$. The ordinary least square (OLS) method aims to minimize $\sum_{i=1}^{n} r_i^2$, which produces unstable result if outliers occur in the data. [H$^+$64] proposed to replace the squared residuals r_i^2 by another function of residuals, in order to reduce the effect of outliers, yielding

$$\hat{\beta}^m = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \rho(r_i) \right\}, \tag{3.1}$$

where ρ is a symmetric, positive-definite function with a unique minimum at zero, and usually is chosen to be less increasing than square [Zha]. We call the solution in (3.1) an M-estimator. Note that an OLS method takes $\rho(t) = t^2$; a LAD method takes $\rho(t) = |t|$. Therefore these are all special cases of M-estimators.

The function ρ, or its derivative, denoted by ψ, can be chosen in a way such that (1) when the underling distribution is truly the same as the assumed one, it provides the estimator with desirable properties (in terms of bias and efficiency); (2) when the data are from a model which is different from the assumed distribution, it provides an estimator with 'not bad' behavior.

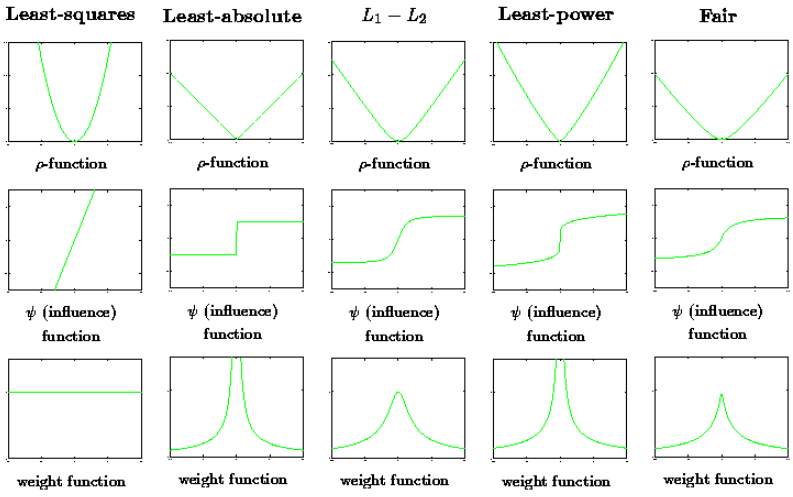
The influence function of M-estimator is proportional to its defining function $\psi(t)$. It measure the influence of an observation on the value of estimated parameter. A robust M-estimate should have a bounded influence function that reduce the influence of large errors and a convex ρ function that grantees the unique minimum. Table 1 list a few example of M-estimator and they are graphically depicted in Fig. 2.

For example, $\psi(t) = t$ in ℓ_2-type M-estimator indicates non-robustness of the OLS estimate, since the influence of an observation increase linearly with the size of its error. We can also learn that ℓ_1-type M-estimator reduces the influence of large errors; $\ell_1 - \ell_2$ estimators combine the advantage from both ℓ_1 and ℓ_2 estimators that it reduces the influence of large errors and meanwhile it is convex; ℓ_p estimators produces good estimate when $v = 1.2$ [Rey12] while it may encounter many difficulties in computation.

Table 1. A Few Example of M-estimators [Zha]

Type	$\rho(t)$	$\psi(t)$	$w(t)$
L_2	$\frac{t^2}{2}$	t	1
L_1	$\lvert t\rvert$	$sgn(t)$	$\frac{1}{\lvert t\rvert}$
$L_1 - L2$	$2(\sqrt{1+t^2/2}-1)$	$\frac{t}{\sqrt{1+t^2/2}}$	$\frac{1}{\sqrt{1+t^2/2}}$
L_p	$\frac{\lvert t\rvert^v}{v}$	$sgn(t)\lvert t\rvert^{v-1}$	$\lvert t\rvert^{v-2}$
'Fair'	$C^2[\frac{\lvert t\rvert}{c} - log(1+\frac{\lvert t\rvert}{c})]$	$\frac{t}{1+\lvert t\rvert/c}$	$\frac{1}{1+\lvert t\rvert/c}$

Figure 2. Graphic Representation of A Few Examples of M-estimators [Zha]



The M-estimate of β based on $\rho(r_i)$ is the solution of the following p equations:

$$\sum_{i}^{n} \psi(r_i) \frac{\partial r_i}{\partial \beta_j} = 0, \tag{3.2}$$

for $j = 1, \ldots, m$. Then we can define a weight function

$$w(t) = \frac{\psi(t)}{t}, \tag{3.3}$$

thus the equation (3.2) becomes

$$\sum_{i}^{n} w(r_i) r_i \frac{\partial r_i}{\partial \beta_j} = 0, \tag{3.4}$$

for $j = 1, \ldots, m$. We will obtain (3.4) if we solve the following iterated reweighted least-squares problem

$$\min \left\{ \sum_{i}^{n} w(r_i^{(s-1)}) r_i^2 \right\}, \tag{3.5}$$

where weight $w(r_i^{(s-1)})$ is computed by using residual r_i obtained from $(s-1)^{th}$ iteration.

3.3 Least Trimmed Square Regression

Least trimmed square(LTS) [Rou84] is one of the robust regression methods that fits a model to a set of data without sensitively affected by the occurrence of outliers. Unlike the OLS method that minimizes the sum of squared residuals, the LTS approach tries to minimize the sum of squared residuals over a subset with size h. Denote the vectors of squared residuals by $\mathbf{r}^2(\beta) = (r_1^2, \ldots, r_n^2)'$ with $r_i^2 = (y_i - \mathbf{x}_i'\beta)^2$.

Then the LTS estimator is given by

$$\hat{\beta}^{lts}(h) = \arg\min_{\beta} \left\{ \sum_{i=1}^{h} (\mathbf{r}^2(\beta))_{i:n} \right\}, \tag{3.6}$$

where $h \leq n$ and $(\mathbf{r}^2(\beta))_{1:n} \leq \cdots \leq (\mathbf{r}^2(\beta))_{n:n}$ are the order statistics of the squared residuals. This method aims to find the subset of h observations that produces the smallest sum of squared residuals. By excluding the largest squared residuals from the summation, the LTS allows those outliers data points to be excluded completely.

There are some good properties of LTS method. In term of the breakdown points, LTS is considered to be a high breakdown method with a breakdown value $(n-h)/n$. Besides, the LTS estimate is asymptotically normal [Hös94]. Thus it is suitable to be used as a starting point for two-step estimators such as MM-estimator [YZ88] and generlaized M-estimators [SRC92]. Considering the implementation, the LTS approach is simple to implement and quite fast to compute.

To implement the LTS approach, [RVD06] proposed an algorithm called FAST-LTS based on concentrating steps or C-steps. Define an objective function $Q(\mathbf{H}, \beta) = \sum_{i \in \mathbf{H}} \mathbf{r}_i^2(\beta)$ where $\mathbf{H} \subseteq \{1, \ldots, n\}$ with $|\mathbf{H}| = h$. Let $\hat{\beta}_{\mathbf{H}} = \arg\min_{\beta}\{Q(\mathbf{H}, \beta)\}$, the estimate from OLS method over subset $\mathbf{H}$. At iteration k, the C-step consists of computing the OLS solution based on the current subset $\mathbf{H}_k$, with $|\mathbf{H}_k| = h$, and constructing the next subset $\mathbf{H}_{k+1}$ from the observation corresponding to the h smallest residuals. It was proved that a C-step method results in a decrease of the LTS objective function, $Q(\mathbf{H}_{k+1}, \hat{\beta}_{\mathbf{H}_{k+1}}) \leq Q(\mathbf{H}_{k+1}, \hat{\beta}_{\mathbf{H}_k}) \leq Q(\mathbf{H}_k, \hat{\beta}_{\mathbf{H}_k})$. Thus a sequence of C-steps yields a local minimum in a finite number of steps [RVD06].

CHAPTER IV

PENALIZED WEIGHTED LEAST SQUARE METHOD

4.1 Motivation

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 square (PLS) methods have become very popular in high-dimensional linear regression analysis since the introduction of the Lasso [Tib96]. 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 setting, most of them do not identify outliers in particular, which themselves can provide important scientific findings. Most of existed 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 thesis, 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 square 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. Our contribution in this thesis can be summarized as follows. First, we provide an efficient robust approach for simultaneous outliers 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 setting; 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.

The rest of this chapter is organized as follows. In Section 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 3. Some non-asymptotic oracle inequalities of the PAWLS estimation error for both the weights and coefficients vectors are discussed in detail in Section 4. In Section 5, we conduct some numerical studies including some simulation studies and real data analysis under both $p < n$ and $p \gg n$ settings.

4.2 Weight Shrinkage

Consider a weighted linear regression

$$y_i = \mathbf{x}_i'\beta^* + \eta_i, \quad 1 \leq i \leq n, \tag{4.1}$$

where y_i and $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ip})'$ are the observed response variable and covariates vector, $\beta^* = (\beta_0^*, \beta_1^*, \cdots, \beta_p^*)'$ 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 $0 \leq \sigma < \infty$. 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 manuscript.

If the weight vector $\mathbf{w} = (w_1, \cdots, w_n)'$ in (4.1) is given or represented as *a priori*, then we can obtain a sparse estimation of β by minimizing a penalized weighted least squares loss with a penalty on β (no penalty on intercept),

$$\tilde{\beta}(\lambda_{1n}, \mathbf{w}) = \underset{\beta \in \mathcal{R}^p}{\arg\min} \{ \frac{1}{2n} \sum_{i=1}^{n} w_i^2 (y_i - \mathbf{x}_i' \beta)^2 + P_{\lambda_{1n}}(\beta) \}. \tag{4.2}$$

For example, an LAD-Lasso takes $w_i = |y_i - \mathbf{x}_i' \beta|^{-1/2}$ and $P_{\lambda_{1n}}(\beta) = \lambda_{1n} \sum_{j=1}^{p} |\beta_j|$ [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 β. In the PAWLS approach we develop here, we allow weights to be data-driven and propose to obtain $\hat{\mathbf{w}}$ and $\hat{\beta}$ simultaneously. In particular, a PAWLS method with the Lasso penalty is to solve

$$(\hat{\beta}, \hat{\mathbf{w}})(\lambda_{1n}, \lambda_{2n}) = \underset{0 < w_i \leq 1}{\arg\min} \{ \frac{1}{2n} \sum_{i=1}^{n} w_i^2 (y_i - \mathbf{x}_i' \beta)^2 + \lambda_{1n} \sum_{j=1}^{p} |\beta_j| + \lambda_{2n} \sum_{i=1}^{n} |1 - w_i| \}, \tag{4.3}$$

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 observation, 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 (4.3). 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,

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

Thus the PAWLS estimation $(\hat{\beta}, \widehat{\mathbf{w}})$ in (4.3) with $\lambda_{1n} = \lambda_{10}/(2n)$ and $\lambda_{2n} = \lambda_{20}/(2n)$ is equivalent to a corresponding posterior mode of β and $\mathbf{w}$. In the left panel of Figure 3, 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 (4.3) from a Bayesian perspective.

4.2.1 A general threshold rule and its link to sparse M-estimation

In fact, the PAWLS with Lasso in (4.3) 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 4.1. (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

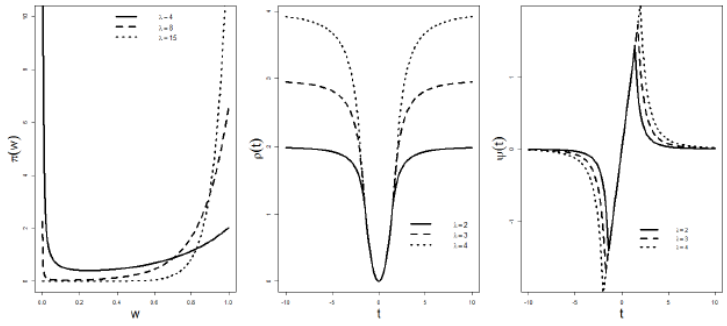
(3) (Two extremes) $\lim_{t\to 0} \Theta_\lambda(t) = 1$ and $\lim_{t\to\infty} \Theta_\lambda(t) = 0$.

The scale threshold function in Definition 4.1 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 (4.3).

Theorem 4.2. *Suppose* $\tilde{\beta} = \tilde{\beta}(0, \tilde{\mathbf{w}})$ *is a solution in* (4.2) *for* $\lambda_{1n} = 0$ *and* $\widetilde{w}_i^2 = \Theta_\lambda(y_i - \mathbf{x}_i\tilde{\beta})$, $1 \leq i \leq n$. *Here* $\Theta_\lambda(\cdot)$ *for some* $\lambda > 0$ *is a threshold function defined in Definition* 4.1. *Then* $\tilde{\beta}$ *is also an M-estimator such that* $\tilde{\beta} = \arg\min_\beta\{\sum_{i=1}^n \rho_\lambda(y_i - \mathbf{x}_i'\beta)\}$. *In particular,* $\psi_\lambda(t) = \frac{d\rho_\lambda(t)}{dt}$ *satisfies,*

$$\psi_\lambda(t) = t\Theta_\lambda(t). \tag{4.4}$$

Figure 3. 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 $\lambda = 2, 3, 4$



The proof of Theorem 4.2 is given in Appendix. Theorem 4.2 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 (4.3) indicates that $\hat{w}_i = \{n\lambda_{2n}/(y_i - \mathbf{x}_i'\hat{\beta})^2\} \wedge 1$. Thus, if we let $\lambda = n\lambda_{2n}$, then the scale shrinkage rule for (4.3) becomes

$$\Theta_\lambda(t) = \begin{cases} \lambda^2/t^4 & if\ t^2 > \lambda, \\ 1 & if\ t^2 \leq \lambda. \end{cases} \tag{4.5}$$

Thus from Theorem 4.2, the PAWLS estimation in (4.3) is linked to a corresponding sparse M-estimator with ψ function with

$$\psi_\lambda(t) = \begin{cases} \lambda^2/t^3 & if\ t^2 > \lambda, \\ t & if\ t^2 \leq \lambda, \end{cases} \tag{4.6}$$

and the corresponding ρ function,

$$\rho_\lambda(t) = \begin{cases} -\lambda^2/(2t^2) + \lambda, & if\ t^2 > \lambda, \\ t^2/2, & if\ t^2 \leq \lambda. \end{cases} \tag{4.7}$$

See the left and right panels in Figure 3 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 (4.7) gives a weakly redescending M estimation with strong robustness. Naturally, the PAWLS solution in (4.3) 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.

4.3 Implementation

4.3.1 Coordinate decent Algorithm for PAWLS

We first notice that (4.3) 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 4.2.1. However, for a given $\mathbf{w}$, the function of β is a convex optimization problem, and the vice versa. Therefore, the objective function (4.3) 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 $\beta^{(1)}$, $\mathbf{w}^{(1)}$ play important roles during alterative iterative process. We suggest to use a multiple iterative strategy as follows: (1) when updating β, we start from $\beta^{(1)} = \mathbf{0}$ 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 $\beta^{(1)} = \hat{\beta}(\widetilde{\lambda}_{1n}, \lambda_{2n})$, 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 $\beta^{(k)}$ and $\mathbf{w}^{(k)}$ are alteratively 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 $1 \leq j \leq p$, let $\lambda_{2i} = \lambda_{2n}$ for $1 \leq i \leq n$
  let $k = 1$ and obtain an initial $\boldsymbol{\beta^{(k)}}$, $\mathbf{w}^{(k)}$, and $\mathbf{r}^{(k)} = \mathbf{y} - \mathbf{X}\beta$
While not converged **do**
  [*update* β]
  $c_j = n^{-1}\mathbf{X}_j'\mathbf{w}^{(k)'}\mathbf{w}\mathbf{X}_j$, $z_j = n^{-1}\mathbf{X}_j'\mathbf{w}^{(k)'}\mathbf{w}\mathbf{r} + c_j\boldsymbol{\beta_j^{(k)}}$
  $\boldsymbol{\beta_j^{(k+1)}} = S(z_j, \lambda_{1j})^1/c_j$
  $\mathbf{r} = \mathbf{r} - \mathbf{X}_j'(\boldsymbol{\beta^{(k+1)}} - \boldsymbol{\beta^{(k)}})$
  [*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 $\|\boldsymbol{\beta^{(k+1)}} - \boldsymbol{\beta^{(k)}}\|_\infty < \epsilon$ and $\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_\infty < \epsilon$
  $k \leftarrow k + 1$
end while
deliver $\hat{\beta} = \beta^{(k)}$ and $\widehat{\mathbf{w}} = \mathbf{w}^{(k)}$

4.3.2 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

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

optimal tuning parameters λ_{1n}^{opt} and λ_{2n}^{opt} by modifying BIC as follows,

$$\mathbf{BIC}(\lambda_{1n}, \lambda_{2n}) = n \log \left\{ \sum_{i=1}^{n} \hat{w}_i^2(\lambda_{1n}, \lambda_{2n})(y_i - \mathbf{x}_i' \hat{\beta}(\lambda_{1n}, \lambda_{2n}))^2 + \frac{p}{n+p} \right\} + \widehat{s}(\lambda_{1n}, \lambda_{2n}) \log(n), \tag{4.8}$$

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 (4.8) 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 search alternatively by minimizing BIC in (4.8) 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 for 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.

4.3.3 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

$$\frac{1}{2n}\sum_{i=1}^{n} w_i^2(y_i - \mathbf{x}_i'\beta)^2 + \lambda_{1n}\sum_{j=1}^{p}|\beta_j|/|\beta_j^{(0)}| + \lambda_{2n}\sum_{i=1}^{n}|1-w_i|/|1-w_i^{(0)}|, \qquad (4.9)$$

where $w_i^{(0)}$ and $\beta_j^{(0)}$ are two initial estimates of w_i and β_j, respectively. The computation of (4.9) 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 (4.9) becomes the PAWLS in (4.3).

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 4.4, the PAWLS-Lasso estimates are reasonable choices for $\beta_j^{(0)}$ and $w_i^{(0)}$ in (4.9). This is also demonstrated in our numerical studies to be presented in the next section.

From our empirical experiences, the above procedure of works very well in all our numerical studies in the next section.

4.4 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 $0 \leq \alpha < 1$. To simplify the presentation, we omit the intercept in model (4.1) in this section. All proofs are given in Appendix.

For notation's convenience, we replace $\nu_i = 1 - w_i$ for $1 \leq i \leq n$ in some scenarios and assume all covariates to be standardized such that $\sum_{i=1}^{n} x_{ij}^2 = n,\ 1 \leq j \leq n$ in this section. We put all weights and covariates coefficients together and denote a $n + p$ dimensional unknown parameters vector $\theta = (\theta_1', \theta_2')'$, where $\theta_1 = (\beta_1, \cdots, \beta_p)'$ with true values $\theta_1^* = \beta^*$ and $\theta_2 = (\lambda_{2n}/\lambda_{1n})(\nu_1, \cdots, \nu_n)'$ with true values $\theta_2^* = (\lambda_{2n}/\lambda_{1n})\mathbf{w}^*$. Here $\mathbf{w}^* = (w_1^*, \cdots, w_n^*)'$. Let $S_{10} = \{1 \leq j \leq p : \beta_j^* \neq 0\}$ with the cardinal value $s_1 = |S_{10}|$, $S_{20} = \{1 \leq i \leq n : w_i^* < 1\}$ with the cardinal value $s_2 = |S_{20}|$, and $J_0 = \{1 \leq k \leq n + p,\ \theta_k^* \neq 0\}$ be the true active set for θ^* with the cardinal value $|J_0| s_1 + s_2 = s$. We also denote $a_n = \min\limits_{1 \leq i \leq S_{20}} w_i^*$.

We consider the fixed design such that $|x_{ij}| \leq b_n$ 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) $\dfrac{s b_n}{n^{1/2}} = o(1)$; (ii) $\dfrac{s \log(n)}{n a_n^2} = o(1)$.

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

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

$$\kappa(s, c) = \min_{\substack{\mathbf{d} \neq 0 \\ \|\mathbf{d}_{J_0^c}\|_1 \leq c\|\mathbf{d}_{J_0}\|_1 \\ |J_0| \leq s}} \frac{\|\mathbf{\Psi}^{1/2}\mathbf{d}\|_2}{\|\mathbf{d}_{J_0}\|_2} > 0, \tag{4.10}$$

where $\|\cdot\|_q$ is the ℓ_q norm, $\mathbf{d} = (\mathbf{d}_1', \mathbf{d}_2')'$ and $\mathbf{\Psi} = \frac{1}{n}\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \sigma^2\omega^{*-2} \end{pmatrix}$ with ω^* 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,

- $\mathbb{A}_1 = \{\|\epsilon'\mathbf{X}\|_\infty < n\lambda_{1n}/4\}$;
- $\mathbb{A}_2 = \{\max_{1 \leq i \leq n} \epsilon_i^2/w_i^* < n\lambda_{2n}/4\}$;
- $\mathbb{A}_3 = \{\|\epsilon'\mathbf{D}_{\tilde{\nu}}\mathbf{X}\|_\infty < n\lambda_{1n}/4\}$, 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 4.3. *On event* $\mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3$,

$$\|\hat{\theta} - \theta^*\|_1 \leq 4\|\hat{\theta}_{J_0} - \theta^*_{J_0}\|_1 \tag{4.11}$$

Lemma 4.4. *Under (A1), we have*

$$P(\mathbb{A}_1^c) \leq 2pe^{-\frac{n\lambda_1^2}{32\sigma^2}} \tag{4.12a}$$

$$P(\mathbb{A}_2^c) \leq 2n \exp\left\{-\frac{n\lambda_{2n}a_n^2}{8\sigma^2}\right\} \tag{4.12b}$$

$$P(\mathbb{A}_3^c) \leq 2\exp\left\{-M_0 \min\left\{\frac{n\lambda_{1n}^4}{256K^2\sigma^4}, \frac{n\lambda_{1n}^2}{16K\sigma^2}\right\}\right\}, \tag{4.12c}$$

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. In particular, if we choose* $\lambda_{1n} \geq \sigma(c_1)^{1/2}(\ln(p)/n)^{1/2}$ *for* $c_1 > 32$, *then*

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

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

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

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 4.3 provides an upper bound of the PAWLS estimator under three events. Lemma 4.4 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 $\hat{\beta}$.

Theorem 4.5. *Suppose A1 and RE(s,3) hold. Then with probability at least* $1 - \sum_{k=1}^{5} h_i$, *we have*

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

and

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

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 4.5 gives the oracle inequalities of joint estimators of θ. Those properties are similar to ones for the PLS estimator (with the Lasso penalty) of β only when $\mathbf{w}^*$ is given in advance. When $\mathbf{w}$ is jointly estimated with β, the non-asymptotic properties for both $\hat{\beta}$ 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 $\hat{\beta}$ 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 4.4 and Theorem 4.5.

Corollary 4.6. *Suppose* $p = O\left(\exp(n^\alpha)\right)$ *for* $1/2 < \alpha < 1$ *and all assumptions in Theorem* 4.5 *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*

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

and

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

4.5 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 $p \gg n$.

4.5.1 Simulation studies

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

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

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 (4.1). However, we will demonstrate that the advantage of the PAWLS are still obvious compared with other methods from simulation studies.

Example 4.7. (Low-dimensional case) We choose $n = 50$, $p = 8$, and set $\beta^* = (3, 2, 1.5, 0, 0, 0, 0, 0)'$. 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 estimation mentioned in Section 1 including the SROS, MMNNG, and sparse LTS (sLTS). 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 $\{1 \leq j \leq p : \hat{\beta}_j \neq 0\} = \{1 \leq j \leq p : \beta_j^* \neq 0\}$; 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 $\#\{1 \leq j \leq p : \hat{\beta}_j \neq 0\}$) is also reported. All those results are summarized in Table 2. 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), $\|\hat{\beta} - \beta\|^2$ out of all repetitions. Those results of MSE (after removing 10% of largest ones) from Case A, C and D are plotted in Figure 4. 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 3. 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.

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

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

Figure 4. Boxplot of MSE in Example 1. The first row: Example 1 (Case A, C and D from the left to right); The second row: Example 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.

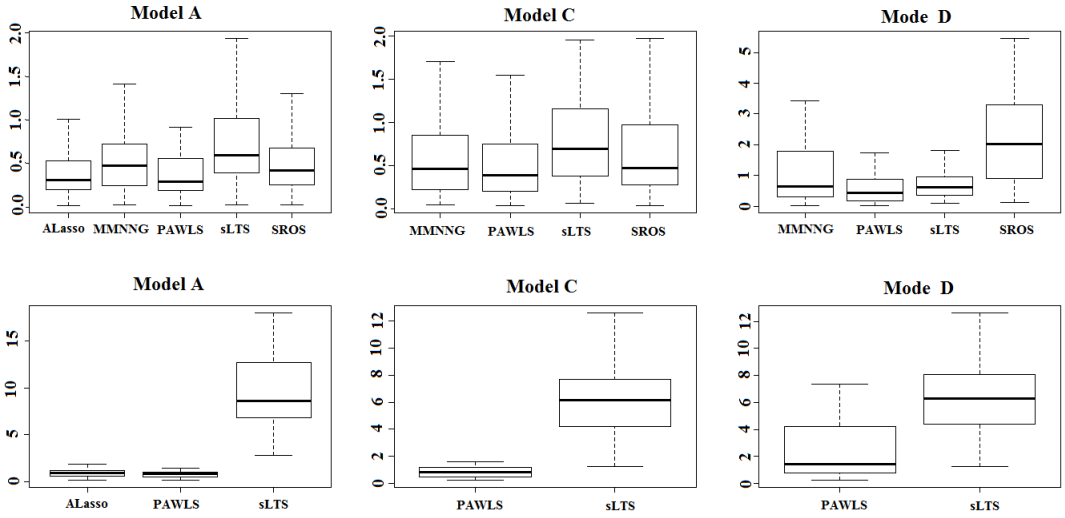


Table 3. Outlier Detection Evaluation in Example 1 and 2

		sLTS			PAWLS		
	Model	M (%)	S (%)	JD(%)	M (%)	S (%)	JD(%)
Example 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 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

Example 4.8. (high-dimensional case) Similar to Example 1, except that $n = 100$, $p = 500$ and $\beta = (\mathbf{2}'_{10}, \mathbf{0}'_{p-10})'$, where $\mathbf{c}_k$ is a k-dimensional vector consists of all c.

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 4. Besides OFR, CFR and AN reported in Example 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 3. Some of MSE results are reported in those Boxplots in Figure 4. It is observed that the advantages of the PAWLS is even more obvious is 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.

4.5.2 Real data applications

The two datasets introduced in 1.3 will be studied in this section: Air pollution data ($p < n$) and NTC-60 data ($p > n$).

4.5.2.1 Air pollution

As mentioned earlier, observation 21 had to be removed since it contains two missing values, resulting in $n = 59$ and $p = 14$ in our study. We also 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.

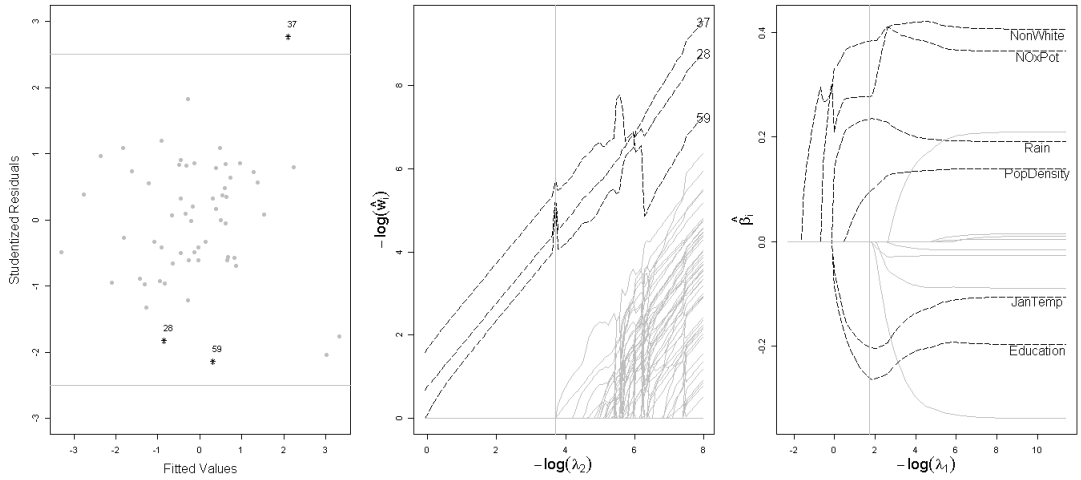
[GV15] analyzed the data with a QQ-plot and reveals the possible contamination of the data set. The PAWLS estimates of β are compared with output from four other methods in Table 5. The PAWLS selects 7 variables from 14 of them. Among them, Rain, PopDensity, NonWhite, and SO2Pot are positively correlated 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 5, where three suspected outliers detected by the PAWLS are highlighted by "*". See the studentized residual plot in the left panel Figure 5. 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 5. 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.

Figure 5. 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}).



4.5.2.2 NCI-60 cancer cell panel

As to the NCI-60 dataset, 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

Table 4. Variable Selection Results for Example 2 ($\beta' = (\mathbf{2}'_{10}, \mathbf{0}'_{p-10})$)

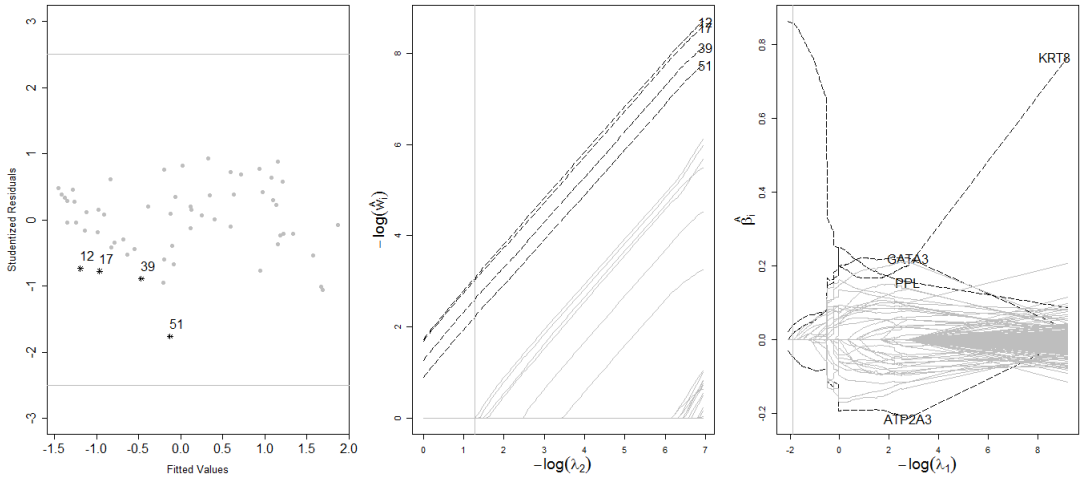
Method	CFR (%)	CFR+2 (%)	OFR (%)	AN (%)	CFR (%)	CFR+2 (%)	OFR (%)	AN
	Case A				**Case B**			
ALasso	55	90	45	11.0	48	74	45	13.1
sLTS	0	0	74	32.6	0	0	91	28.3
PAWLS	92	100	8	10.1	96	98	2	10.0
	Case C				**Case D**			
ALasso	0	0	5	40.2	0	0	3	39.0
sLTS	0	0	93	32.3	0	0	92	31.9
PAWLS	84	97	13	10.0	44	71	43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

Table 5. 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

to analyze this data, where the former selects 27 genes including KRT8 and GATA3, and the latter selects only KRT8.

Figure 6. 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 6. 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 V

DISCUSSION AND FUTURE WORK

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 data-adaptively joint estimation of weight and coefficients vectors and its link to M-estimation justify both of the strong robustness and estimation efficiency of this PAWLS approach under 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 a prior information or 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 of 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 extend the study of regularized M-estimation in high-dimensional settings. For example, [NYWR09] established consistency and convergence rates for regularized M-estimators under high-dimensional setting 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 square 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 have several other relevant research questions not fully addressed in this thesis. For example, the robustness of regression can be also 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 of 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.

There are many extensions of this work to other types of penalized weighted approach. For example, one can extend it to penalized weighted generalized linear model, or penalized weighted ridge regression.

REFERENCES

[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.

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

[AR+13] Anthony Atkinson, Marco Riani, et al., *Exploring multivariate data with the forward search*, Springer Science & Business Media, 2013.

[ARC10] Anthony C Atkinson, Marco Riani, and Andrea Cerioli, *The forward search: Theory and data analysis*, Journal of the Korean Statistical Society **39** (2010), no. 2, 117–134.

[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.

[AY01] Charu C Aggarwal and Philip S Yu, *Outlier detection for high dimensional data*, ACM Sigmod Record, vol. 30, ACM, 2001, pp. 37–46.

[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.

[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.

[CC08] Jiahua Chen and Zehua Chen, *Extended bayesian information criteria for model selection with large model spaces*, Biometrika **95** (2008), no. 3, 759–771.

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

[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.

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

[DSP66] Norman Richard Draper, Harry Smith, and Elizabeth Pownell, *Applied regression analysis*, vol. 3, Wiley New York, 1966.

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

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[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).

[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.

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

[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.

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

[HBK84] Douglas M Hawkins, Dan Bradu, and Gordon V Kass, *Location of several outliers in multiple-regression data using elemental sets*, Technometrics **26** (1984), no. 3, 197–208.

[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.

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

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

[Hös94] Ola Hössjer, *Rank-based estimates in the linear model with high breakdown point*, Journal of the American Statistical Association **89** (1994), no. 425, 149–158.

[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.

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

[KVAZ07] Jafar A Khan, Stefan Van Aelst, and Ruben H Zamar, *Robust linear model selection based on least angle regression*, Journal of the American Statistical Association **102** (2007), no. 480, 1289–1299.

[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.

[LZ86] H Linhart and W Zucchini, *Wiley series in probability and mathematical statistics. model selection*, 1986.

[MB06] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The annals of statistics (2006), 1436–1462.

[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.

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

[Rey12] William Rey, *Introduction to robust and quasi-robust statistical methods*, Springer Science & Business Media, 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.

[RVD06] Peter J Rousseeuw and Katrien Van Driessen, *Computing lts regression for large data sets*, Data mining and knowledge discovery **12** (2006), no. 1, 29–45.

[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.

[Sie82] Andrew F Siegel, *Robust regression using repeated medians*, Biometrika **69** (1982), no. 1, 242–244.

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

[SRC92] Douglas G Simpson, David Ruppert, and Raymond J Carroll, *On one-step gm estimates and stability of inferences in linear regression*, Journal of the American Statistical Association **87** (1992), no. 418, 439–450.

[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.

[SY15] Ezequiel Smucler and Víctor J Yohai, *Robust and sparse estimators for linear regression models*, arXiv preprint arXiv:1508.01967 (2015).

[Tib96] Robert Tibshirani, *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.

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

[Wei05] Sanford Weisberg, *Applied linear regression*, vol. 528, John Wiley & Sons, 2005.

[Wik16a] Wikipedia, *Heteroscedasticity — wikipedia, the free encyclopedia*, 2016, [Online; accessed 10-March-2016].

[Wik16b] ———, *Outlier — wikipedia, the free encyclopedia*, 2016, [Online; accessed 10-March-2016].

[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.

[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.

[WM15] Cen Wu and Shuangge Ma, *A selective review of robust variable selection with applications in bioinformatics*, Briefings in bioinformatics **16** (2015), no. 5, 873–883.

[XH09] Huiliang Xie and Jian Huang, *Scad-penalized regression in high-dimensional partially linear models*, The Annals of Statistics (2009), 673–696.

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

[Xu07] Shizhong Xu, *An empirical bayes method for estimating epistatic effects of quantitative trait loci*, Biometrics **63** (2007), no. 2, 513–521.

[YZ88] Victor J Yohai and Ruben H Zamar, *High breakdown-point estimates of regression by means of the minimization of an efficient scale*, Journal of the American statistical association **83** (1988), no. 402, 406–413.

[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.

[Zha] Z Zhang, *Parameter estimation techniques: A tutorial with application to conic fitting (1996)*, URL http://www-sop. inria. fr/robotvis/personnel/zzhang/Publis/Tutorial-Estim/Main. html.

[Zha07] Cun Hui Zhang, *Penalized linear unbiased selection*, Department of Statistics and Bioinformatics, Rutgers University (2007), 2007–003.

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

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

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 217–242.

APPENDIX A

PROOF

A.0.3 Proof in Section 4.2.1

Proof of Theorem 4.2

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

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

and

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

Thus

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\tilde{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{W}}^2\mathbf{y} = \tilde{\mathbf{W}}^{-1}(\mathbf{I} - \mathbf{H}_{\tilde{\mathbf{W}}\mathbf{X}})\tilde{\mathbf{W}}\mathbf{y}, \qquad (1.1)$$

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 (4.4). Then from (1.1),

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

□

A.0.4 Proof in Section 4.4

To prove those lemmas and Theorem 4.5 in Section 4.4, we need to reformulate the model as follows. In particular, we define $r_{i,\beta} = y_i - \mathbf{x}_i'\beta$ and a $n \times n$ matrix $\mathbf{R}_\beta = diag\{r_{1,\beta}, \cdots, r_{n,\beta}\}$. Let $\mathbf{r}_{i,\beta}$ be the ith column vector of $\mathbf{R}_\beta$. Recall the notation $\nu_i = 1 - w_i$ and $\theta = (\theta_1', \theta_2')'$, where $\theta_1 = (\beta_1, \cdots, \beta_p)'$ and $\theta_2 = (\lambda_{2n}/\lambda_{1n})(\nu_1, \cdots, \nu_n)'$. Define $\mathbf{z}_{i,\beta}' = \left(\mathbf{x}_i', (\lambda_{1n}/\lambda_{2n})\mathbf{r}_{i,\beta}'\right)$ and $\mathbf{Z}_\beta = \begin{pmatrix} \mathbf{z}_{1,\beta}' \\ \cdots \\ \mathbf{z}_{n,\beta}' \end{pmatrix} = \begin{pmatrix} \mathbf{X} & (\lambda_{1n}/\lambda_{2n})\mathbf{R}_\beta \end{pmatrix}$. Then model (4.1) with true parameter values becomes

$$y_i = \mathbf{r}_{i,\beta}'\nu^* + \mathbf{x}_i'\beta^* + \epsilon_i = \mathbf{z}_{i,\beta^*}'\theta^* + \epsilon_i. \tag{1.2}$$

Recall that the penalized likelihood of PAWLS in (4.3),

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

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

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

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

Proof of Lemma 4.3

Using the definition,

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

Then

$$\begin{aligned}\frac{1}{2n}\|\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*\|^2 &\leq \frac{1}{n}\epsilon'(\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*) + \lambda_{1n}[\|\theta^*\|_1 - \|\hat{\theta}\|_1] \\ &\leq \frac{1}{n}|\epsilon'\mathbf{Z}_{\beta^*}(\hat{\theta} - \theta^*)| + \frac{1}{n}|\epsilon'(\mathbf{Z}_{\beta^*} - \mathbf{Z}_{\hat{\beta}})\hat{\theta}| + \lambda_{1n}[\|\theta^*\|_1 - \|\hat{\theta}\|_1]\end{aligned} \tag{1.3}$$

Notice that

$$\begin{aligned}\mathbf{Z}_{\beta^*}(\hat{\theta} - \theta^*) &= \mathbf{X}(\hat{\theta}_1 - \theta_1^*) + (\lambda_{1n}/\lambda_{2n})\mathbf{R}_{\beta^*}(\hat{\theta}_2 - \theta_2^*) \\ &= \mathbf{X}(\hat{\theta}_1 - \theta_1^*) + (\lambda_{1n}/\lambda_{2n})\omega^{*-1}\mathbf{D}_\epsilon(\hat{\theta}_2 - \theta_2^*),\end{aligned}$$

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 $\mathbf{D}_\nu$. Then on event $\mathbb{A}_1 \cap \mathbb{A}_2$, we have

$$\begin{aligned}\frac{1}{n}|\epsilon'\mathbf{Z}_\beta^*(\hat{\theta} - \theta^*)| &\leq \frac{1}{n}\|\epsilon'\mathbf{X}\|_\infty\|\hat{\theta}_1 - \theta_1^*\|_1 + \frac{\lambda_{1n}}{n\lambda_{2n}}\max_{1\leq i\leq n}\frac{\epsilon_i^2}{w_i^*}\|\hat{\theta}_2 - \theta_2^*\|_1 \\ &\leq \frac{\lambda_{1n}}{4}\|\hat{\theta}_1 - \theta_1^*\|_1 + \frac{\lambda_{1n}}{4}\|\hat{\theta}_2 - \theta_2^*\|_1 \\ &\leq \frac{\lambda_{1n}}{4}\|\hat{\theta} - \theta^*\|_1.\end{aligned} \tag{1.4}$$

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

$$(\mathbf{Z}_{\beta^*} - \mathbf{Z}_{\hat{\beta}})\hat{\theta} = (\lambda_{1n}/\lambda_{2n}) diag(\mathbf{x}_1'(\hat{\beta} - \beta^*), \cdots, \mathbf{x}_n'(\hat{\beta} - \beta^*))\hat{\theta}_2 = \mathbf{D}_{\tilde{\nu}}\mathbf{X}(\hat{\beta} - \beta^*).$$

Then

$$\begin{aligned} \frac{1}{n}|\epsilon'(\mathbf{Z}_{\beta^*} - \mathbf{Z}_{\hat{\beta}})\hat{\theta}| &= \frac{1}{n}|\epsilon'\mathbf{D}_{\tilde{\nu}}\mathbf{X}(\hat{\beta} - \beta^*)| \\ &\leq \|\epsilon'\mathbf{D}_{\tilde{\nu}}\mathbf{X}\|_\infty \|\hat{\beta} - \beta^*\|_1 \\ &\leq (\lambda_{1n}/4)\|\hat{\beta} - \beta^*\|_1, \end{aligned} \tag{1.5}$$

where the last "$\leq$" holds on events $\mathbb{A}_3$.

From (1.4-1.5), we obtain

$$\begin{aligned} \frac{1}{2n}\|Z_{\hat{\beta}}\hat{\theta} - Z_{\beta^*}\theta^*\|^2 &\leq \frac{\lambda_{1n}}{4}\|\hat{\theta} - \theta^*\|_1 + \frac{\lambda_{1n}}{4}\|\hat{\beta} - \beta^*\|_1 + \lambda_{1n}[\|\theta^*\|_1 - \|\hat{\theta}\|_1] \\ &= \frac{\lambda_{1n}}{2}\|\hat{\beta} - \beta^*\|_1 + \lambda_{1n}[\|\beta^*\|_1 - \|\hat{\beta}\|_1] \\ &\quad + \frac{\lambda_{2n}}{4}\|\tilde{\nu} - \nu^*\|_1 + \lambda_{2n}[\|\nu^*\|_1 - \|\tilde{\nu}\|_1]. \end{aligned} \tag{1.6}$$

Adding $\frac{\lambda_{1n}}{2}\|\hat{\beta} - \beta^*\|_1 + \frac{\lambda_{2n}}{2}\|\tilde{\nu} - \nu^*\|_1$ on two sides,

$$\begin{aligned} &\frac{1}{2n}\|Z_{\hat{\beta}}\hat{\theta} - Z_{\beta^*}\theta^*\|^2 + \frac{\lambda_{1n}}{2}\|\hat{\beta} - \beta^*\|_1 + \frac{\lambda_{2n}}{2}\|\tilde{\nu} - \nu^*\|_1 \\ &\leq \lambda_{1n}(\|\hat{\beta} - \beta^*\|_1 + [\|\beta^*\|_1 - \|\hat{\beta}\|_1]) \\ &\quad + \lambda_{2n}(\|\tilde{\nu} - \nu^*\|_1 + [\|\nu^*\|_1 - \|\tilde{\nu}\|_1]) \\ &\leq 2\lambda_{1n}\|\hat{\beta}_{J_{10}} - \beta^*_{J_{10}}\|_1 + 2\lambda_{2n}\|\tilde{\nu}_{J_{20}} - \nu^*_{J_{20}}\|_1. \end{aligned} \tag{1.7}$$

The last "$\leq$" holds since $\|\hat{\beta}_{J_{10}^c} - \beta^*_{J_{10}^c}\|_1 + \|\beta^*_{J_{10}^c}\|_1 - \|\hat{\beta}_{J_{10}^c}\|_1 = 0$ and $\|\tilde{\nu}_{J_{20}^c} - \nu^*_{J_{20}^c}\|_1 + \|\nu^*_{J_{20}^c}\|_1 - \|\tilde{\nu}_{J_{20}^c}\|_1 = 0$. Thus we have

$$\|\hat{\beta} - \beta^*\|_1 + (\lambda_{2n}/\lambda_{1n})\|\tilde{\nu} - \nu^*\|_1 \leq 4\|\hat{\beta}_{J_{10}} - \beta^*_{J_{10}}\|_1 + 4(\lambda_{2n}/\lambda_{1n})\|\tilde{\nu}_{J_{20}} - \nu^*_{J_{20}}\|_1.$$

Thus (4.11) holds.

$$\|\hat{\theta} - \theta^*\|_1 \leq 4\|\hat{\theta}_{J_0} - \theta^*_{J_0}\|_1$$

and

$$\|\hat{\theta}_{J_0^c} - \theta^*_{J_0^c}\|_1 \leq 3\|\hat{\theta}_{J_0} - \theta^*_{J_0}\|_1.$$

□

Proof of Lemma 4.4

$$\begin{aligned}
P(\mathbb{A}_1^c) &= P\left(\|\mathbf{X}'\epsilon\|_\infty > n\lambda_{1n}/4\right) \\
&= P\left(\max_{1\leq j\leq p} |\sum_{i=1}^n x_{ij}\epsilon_i| > n\lambda_{1n}/4\right) \\
&= P\left(\max_{1\leq j\leq p} |\tau_j| > \sqrt{n}\lambda_{1n}/(4\sigma)\right) \\
&\leq pP\left(|\tau_j| > \sqrt{n}\lambda_{1n}/(4\sigma)\right) \\
&\leq 2pe^{-\frac{n\lambda_1^2}{32\sigma^2}}.
\end{aligned}$$

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

$$\begin{aligned} P(\mathbb{A}_3^c) &\leq P\left(\left(\sum_{1\leq i\leq n}\epsilon_i^2\right)^{1/2}\left(\max_{1\leq j\leq p}\sum_{1\leq i\leq n}\tilde{\nu}_i^2 x_{ij}\right)^{1/2} > n\lambda_{1n}/4\right) \\ &\leq P\left(\left(\sum_{1\leq i\leq n}\epsilon_i^2\right)^{1/2} n^{1/2} > n\lambda_{1n}/4\right) \\ &\leq P\left(\sum_{1\leq i\leq n}\frac{1}{n}\frac{\epsilon_i^2}{\sigma^2} > \frac{\lambda_{1n}^2}{16\sigma^2}\right) \\ &\leq 2\exp\left\{-M_0\min\left\{\frac{n\lambda_{1n}^4}{256K^2\sigma^4}, \frac{n\lambda_{1n}^2}{16K\sigma^2}\right\}\right\}, \end{aligned} \tag{1.8}$$

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 (1.8).

□

Proof of Theorem 4.5

Define $\hat{\mathbf{\Sigma}}^* = \frac{1}{n}\mathbf{Z}'_{\beta^*}\mathbf{Z}_{\beta^*}$ and $\mathbf{\Sigma} = E[\hat{\mathbf{\Sigma}}^*]$. The "$\hat{}$" on $\hat{\mathbf{\Sigma}}^*$ is used to address its stochastic property, not the estimating behavior. From the definition, we have

$$n\hat{\mathbf{\Sigma}}^* = \sum_{i=1} \mathbf{z}_{i,\beta^*}\mathbf{z}'_{i,\beta^*} = \begin{pmatrix} \sum_{1\le i\le n} \mathbf{x}_i\mathbf{x}'_i & (\lambda_{1n}/\lambda_{2n})\sum_{1\le i\le n} \mathbf{x}_i\mathbf{r}'_{i,\beta^*} \\ (\lambda_{1n}/\lambda_{2n})\sum_{1\le i\le n} \mathbf{r}_{i,\beta^*}\mathbf{x}'_i & (\lambda_{1n}/\lambda_{2n})^2\sum_{1\le i\le n} \mathbf{r}_{i,\beta^*}\mathbf{r}'_{i,\beta^*} \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \frac{1}{n}\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0}_{p\times n} \\ \mathbf{0}_{n\times p} & \sigma^2(\lambda_{1n}/\lambda_{2n})^2\omega^{*-2} \end{pmatrix}$$

since $E\left[\sum_{i=1}^n \mathbf{r}_{i,\beta^*}\mathbf{r}'_{i,\beta^*}\right] = \sigma^2\omega^{*-2} = diag\{\sigma^2/w_1^{*2}, \cdots, \sigma^2/w_n^{*2}\}$. 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

$$|(\mathbf{d}'\hat{\mathbf{\Sigma}}^*\mathbf{d}) - (\mathbf{d}'\mathbf{\Sigma}\mathbf{d})| \le \delta_n(\|\mathbf{d}\|_1)^2 \le 16\delta_n(\|\mathbf{d}_{J_0}\|_1)^2 \le 16s\delta_n(\|\mathbf{d}_{J_0}\|)^2. \tag{1.9}$$

The last "$\leq$" is from the Cauchy-Schwartz inequality. From the condition $RE(s,3)$ in (4.10) and (1.9), we have

$$\begin{aligned}\kappa(s,3)\|\mathbf{d}_{J_0}\| &\leq (\mathbf{d}'\boldsymbol{\Sigma}\mathbf{d})^{1/2} \\ &\leq (\mathbf{d}'\hat{\boldsymbol{\Sigma}}^*\mathbf{d})^{1/2} + (|\mathbf{d}'(\hat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma})\mathbf{d}|)^{1/2} \\ &\leq (1/\sqrt{n})\|\mathbf{Z}_{\beta^*}\mathbf{d}\| + 4\sqrt{s\delta_n}\|\mathbf{d}_{J_0}\|.\end{aligned}$$

Plugging in $\mathbf{d} = \hat{\theta} - \theta^*$, we obtain

$$\begin{aligned}&\kappa(s,3)\|\hat{\theta}_{J_0} - \theta^*_{J_0}\| \\ &\leq (1/\sqrt{n})\|\mathbf{Z}_{\beta^*}(\hat{\theta} - \theta^*)\| + 4\sqrt{s\delta}\|\hat{\theta}_{J_0} - \theta^*_{J_0}\| \\ &\leq (1/\sqrt{n})\|(\mathbf{Z}_{\beta^*} - \mathbf{Z}_{\hat{\beta}})\hat{\theta}\| + (1/\sqrt{n})\|\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*\| + 4\sqrt{s\delta_n}\|\hat{\theta}_{J_0} - \theta^*_{J_0}\|.\end{aligned} \tag{1.10}$$

We will check $(1/\sqrt{n})\|(\mathbf{Z}_{\beta^*} - \mathbf{Z}_{\hat{\beta}})\hat{\theta}\|$ and $(1/\sqrt{n})\|\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*\|$ separately.

First, from Lemma 4.3, we know

$$\begin{aligned}(1/2n)\|\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*\|^2 &\leq (\lambda_{1n}/2)\|\hat{\theta} - \theta^*\|_1 + \lambda_{1n}[\|\theta^*\|_1 - \|\hat{\theta}\|_1] \\ &\leq \lambda_{1n}\|\hat{\theta}_{J_0} - \theta^*_{J_0}\|_1 \\ &\leq \lambda_{1n}\sqrt{s}\|\theta^*_{J_0} - \hat{\theta}_{J_0}\|.\end{aligned}$$

Then

$$(1/\sqrt{n})\|\mathbf{Z}_{\hat{\beta}}\hat{\theta} - \mathbf{Z}_{\beta^*}\theta^*\| \leq (2\lambda_{1n})^{1/2}s^{1/4}\|\theta^*_{J_0} - \hat{\theta}_{J_0}\|^{1/2}. \tag{1.11}$$

On the other hand,

$$\begin{aligned}(1/n)\|(\mathbf{Z}_{\beta^*}-\mathbf{Z}_{\hat{\beta}})\hat{\theta}\|^2 &= (1/n)\sum_{i=1}^n\left[\widehat{\nu}_i\mathbf{x}_i'(\beta^*-\hat{\beta})\right]^2\\ &\le (1/n)\sum_{i=1}^n\left[\widehat{\nu}_i^2\max_{1\le j\le p}x_{ij}^2(\|\beta^*-\hat{\beta}\|_1)^2\right]\\ &\le (\widehat{s}_{2n}/n)b_n^2(\|\beta^*-\hat{\beta}\|_1)^2\\ &\le (\widehat{s}_{2n}/n)b_n^2 4(\|\theta^*_{J_0}-\hat{\theta}_{J_0}\|_1)^2\\ &\le (\widehat{s}_{2n}/n)b_n^2 4s(\|\theta^*_{J_0}-\hat{\theta}_{J_0}\|)^2,\end{aligned}$$

where $\widehat{s}_{2n}=\sum_{i=1}^n\widehat{\nu}_i$. Then

$$(1/\sqrt{n})\|(\mathbf{Z}_{\beta^*}-\mathbf{Z}_{\hat{\beta}})\hat{\theta}\|\le 2s^{1/2}(\widehat{s}_{2n}/n)^{1/2}b_n\|\theta^*_{J_0}-\hat{\theta}_{J_0}\|. \tag{1.12}$$

In fact, as what we will verify in Lemma 1.1 and 1.2, 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 (1.10-1.12), we have

$$\begin{aligned}\kappa(s,3)\|\hat{\theta}_{J_0}-\theta^*_{J_0}\| &\le 2s^{1/2}(\widehat{s}_{2n}/n)^{1/2}b_n\|\theta^*_{J_0}-\hat{\theta}_{J_0}\|\\ &\quad+(2\lambda_{1n})^{1/2}s^{1/4}\|\theta^*_{J_0}-\hat{\theta}_{J_0}\|^{1/2}+4(s\delta_n)^{1/2}\|\hat{\theta}_{J_0}-\theta^*_{J_0}\|.\end{aligned}$$

Then

$$\|\hat{\theta}_{J_0} - \theta^*_{J_0}\| \leq \frac{2\lambda_{1n} s^{1/2}}{[\kappa(s,3) - (2s^{1/2}(\widehat{s}_{2n}/n)^{1/2} b_n + 4(s\delta_n)^{1/2})]^2} \leq \frac{8\lambda_{1n} s^{1/2}}{\kappa^2(s,3)}.$$

Thus

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

□

Lemma 1.1. *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*

$$P\left(s\delta_n > \zeta\right) \leq \frac{3\sigma}{\sqrt{\zeta}} \frac{\lambda_{1n}\sqrt{s}}{\sqrt{n}\lambda_{2n} a_n} \sqrt{1 + \log(2n)} + \frac{3\sigma}{\sqrt{2\zeta}} \frac{s\lambda_{1n} b_n}{n\lambda_{2n} a_n} \sqrt{1 + \log(2n)} \to 0. \quad (1.13)$$

Proof of Lemma 1.1

Notice that $E[\mathbf{R}_{\beta^*}] = \mathbf{0}$ and $E[\mathbf{R}^2_{\beta^*}] = \sigma^2\omega^{*-2}$. Then

$$\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma} = (1/n) \begin{pmatrix} \mathbf{0}_{p\times p} & (\lambda_{1n}/\lambda_{2n})\mathbf{X}'\mathbf{R}_{\beta^*} \\ (\lambda_{1n}/\lambda_{2n})\mathbf{X}\mathbf{R}'_{\beta^*} & (\lambda_{1n}/\lambda_{2n})^2(\mathbf{R}^2_{\beta^*} - \sigma^2\omega^{*-2}) \end{pmatrix}.$$

Then $s\|\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\|_\infty = \max\{(1/n)(\lambda_{1n}/\lambda_{2n})^2 s\|\mathbf{R}^2_{\beta^*} - \sigma^2\omega^{*-2}\|_\infty, (1/n)(\lambda_{1n}/\lambda_{2n}) s\|\mathbf{X}'\mathbf{R}_{\beta^*}\|_\infty\}$. We will check $\frac{s\lambda^2_{1n}}{n\lambda^2_{2n}}\|\mathbf{R}^2_{\beta^*} - \sigma^2\omega^{*-2}\|_\infty \to 0$ and $(1/n)(s\lambda_{1n}/\lambda_{2n})\|\mathbf{X}'\mathbf{R}_{\beta^*}\|_\infty \to 0$ with

probability separately. For any $\zeta > 0$,

$$
\begin{aligned}
& P\left((1/n)(\lambda_{1n}/\lambda_{2n})^2 s\|\mathbf{R}^2_{\beta^*} - \sigma^2\omega^{*-2}\|_\infty > \zeta\right) \\
& \leq P\left(\max_{1\leq i\leq n}|\epsilon_i^2/\sigma^2 - 1| > (n\zeta\lambda_{2n}^2 a_n^2)/(s\lambda_{1n}^2\sigma^2)\right) \\
& \leq P\left(\max_{1\leq i\leq n}\epsilon_i^2/\sigma^2 > (n\zeta\lambda_{2n}^2 a_n^2)/(s\lambda_{1n}^2\sigma^2) - 1\right) \\
& \leq P\left(\max_{1\leq i\leq n}\epsilon_i^2/\sigma^2 > (n\zeta\lambda_{2n}^2 a_n^2)/(4s\lambda_{1n}^2\sigma^2)\right) \\
& \leq P\left(\max_{1\leq i\leq n}|\epsilon_i/\sigma| > (\sqrt{\zeta}/(2\sigma))(\sqrt{n}\lambda_{2n}a_n/(\sqrt{s}\lambda_{1n}))\right) \\
& \leq (2\sigma)/(\sqrt{\zeta})(\lambda_{1n}\sqrt{s}/(\sqrt{n}\lambda_{2n}a_n))E\left[\max_{1\leq i\leq n}|\epsilon_i/\sigma|\right] \\
& \leq (3\sigma)/(\sqrt{\zeta})(\lambda_{1n}\sqrt{s}/(\sqrt{n}\lambda_{2n}a_n))\sqrt{1+\log(2n)}.
\end{aligned}
\tag{1.14}
$$

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 $s\lambda_{1n}^2/(n\lambda_{2n}^2)\|\mathbf{R}^2_{\beta^*} - \sigma^2\omega^{*-2}\|_\infty \to 0$. Similarly for $\forall\zeta > 0$,

$$
\begin{aligned}
& P\left((1/n)(s\lambda_{1n}/\lambda_{2n})\|\mathbf{X}'\mathbf{R}_{\beta^*}\|_\infty > \zeta\right) \\
& \leq P\left(\max_{1\leq i\leq n}|x_{ij}r_{i,\beta^*}| > \zeta n\lambda_{2n}/(s\lambda_{1n})\right) \\
& \leq P\left(\max_{1\leq i\leq n}|\epsilon_i| > \zeta n\lambda_{2n}a_n/(s\lambda_{1n}b_n)\right) \\
& \leq (s\lambda_{1n}b_n)/(\zeta n\lambda_{2n}a_n)E\left[\max_{1\leq i\leq n}|\epsilon_i|\right] \\
& \leq (3\sigma s\lambda_{1n}b_n\sqrt{1+\log(2n)})/(2\zeta n\lambda_{2n}a_n).
\end{aligned}
\tag{1.15}
$$

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 4.5 are obtained by replacing ζ by $(\kappa(s,3)/16)^2$ in (1.14) and (1.15). $\square$

Lemma 1.2. *Suppose (A1), (A2-i) and (A3) hold. Then under* $\lambda_{1n}/\lambda_{2n} \leq O(1)$,

$$P\left(b_n s\widehat{s}_{2n}/n)^{1/2} > \kappa(s,3)/8\right) \to 0. \tag{1.16}$$

Proof of Lemma 1.2

From (1.7),

$$\begin{aligned} \frac{\lambda_{2n}}{2}\|\widehat{\nu} - \nu^*\|_1 &\leq \lambda_{1n}\|\beta^*\|_1 + \lambda_{2n}\|\widehat{\nu}_{J_{20}} - \nu^*_{J_{20}}\|_1 \\ &\leq s_1\|\beta^*\|_\infty\lambda_{1n} + \lambda_{2n}\|\widehat{\nu}_{J_{20}} - \nu^*_{J_{20}}\|_1 \\ &\leq Ms_1 + 2s_2\lambda_{2n}, \end{aligned} \tag{1.17}$$

where $s_1 = |J_{10}|$ and $s_2 = |J_{20}|$. The last "$\leq$" is from (A3). Thus,

$$\begin{aligned} \textstyle\sum_{i=1}^n \widehat{\nu}_i &\leq \|\nu^*\|_1 + \|\widehat{\nu} - \nu^*\|_1 \\ &\leq s_2 + \|\widehat{\nu} - \nu^*\|_1 \\ &\leq 5s_2 + 2Ms_1(\lambda_{1n}/\lambda_{2n}). \end{aligned}$$

If we $\lambda_{1n}/\lambda_{2n} \leq O(1)$, under (A2-i), we have

$$\sqrt{s\|\widehat{\nu}\|_1 b_n^2/n} \leq O\left((sb_n^2/n)^{1/2}(5s_2 + 2Ms_1)^{1/2}\right) \leq O\left(sb_n/n^{1/2}\right) \to 0.$$

□

Proof of Corollary 4.6

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 4.5, we get

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

and

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

Thus using the Cauchy-Schwarz inequality again,

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

□