The Pennsylvania State University

The Graduate School

FEATURE SCREENING AND VARIABLE SELECTION FOR ULTRAHIGH DIMENSIONAL DATA ANALYSIS

A Dissertation in

Statistics

by

Wei Zhong

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Wei Zhong was reviewed and approved* by the following:

Runze Li
Professor of Statistics and Graduate Program Chair
Dissertation Advisor and Chair of Committee

Bruce G. Lindsay
Willaman Professor of Statistics and Department Head

Dennis K.J. Lin
Distinguished Professor of Statistics and Supply Chain

Jingzhi Huang
David H.Mckinley Professor of Business and Associate Professor of Finance

*Signatures are on file in the Graduate School.

Abstract

This dissertation is concerned with feature screening and variable selection in ultrahigh dimensional data analysis, where the number of predictors, p, greatly exceeds the sample size n. That is, $p \gg n$. Ultrahigh dimensional data analysis has become increasingly important in diverse fields of scientific fields, such as genetics and finance.

In Chapter 3, we develop a sure independence screening procedure based on the distance correlation learning (DC-SIS, for short) to select important predictors for ultrahigh dimensional data. The DC-SIS can be implemented as easily as the sure independence screening procedure based on the Pearson correlation (SIS, for short) proposed by Fan and Lv (2008). However, the DC-SIS can significantly improve the SIS. Fan and Lv (2008) established the sure screening property for the SIS based on linear models. That is, with a proper threshold, it can select all important predictors with probability approaching to one as $n \to \infty$. We show that the sure screening property is valid for the DC-SIS under more general settings including linear models. Furthermore, the implementation of the DC-SIS does not require model specification (e.g., linear model or generalized linear

model) for responses or predictors. This is a very appealing property in ultrahigh dimensional data analysis. Moreover, the DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables. We establish its sure screening property for the DC-SIS, and conduct simulations to examine its finite sample performance. An iterative procedure DC-ISIS is also proposed to enhance the finite sample performance. Numerical comparison indicates that the DC-SIS performs much better than the SIS in various models. We also illustrate the performance of DC-SIS and DC-ISIS through two real data examples.

In Chapter 4, we propose a two-stage feature screening and variable selection procedure to study the estimation of the index parameter in heteroscedastic single-index models with ultrahigh dimensional covariates. In the screening stage, we propose a robust independent ranking and screening (RIRS) procedure to reduce the ultrahigh dimensionality of the covariates to a moderate scale. Aside from its computational simplicity, the RIRS procedure maintains the ranking consistency property in the terminology of Zhu, Li, Li and Zhu (2011) and the sure screening property in the terminology of Fan and Lv (2008). Therefore, in an asymptotic sense the RIRS procedure guarantees to retain all the truly active predictors. However, some inactive predictors may be selected as well. In the cleaning stage, we propose penalized linear quantile regression to refine the selection of the preceding RIRS procedure, and to simultaneously estimate the direction of the index parameter. We establish the consistency and the oracle property of the resulting penalized estimator, and demonstrate through comprehensive numerical studies that the two-stage estimation procedure is computationally expedient and presents an outstanding finite sample performance.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I am sincerely grateful to my dissertation advisor, Dr. Runze Li, for his valuable guidance, inspirational encouragement and helpful comments on my academic career. Without his supervision and support, it would be impossible for me to survive in the academic research and complete this dissertation. Second, I wish to express my propound appreciation to my committee members, Dr. Bruce Lindsay, Dr. Dennis Lin and Dr. Jingzhi Huang, for their precious time and comments to improve this dissertation.

Further, I would like to say special thanks to my collaborator, Dr. Liping Zhu, for his valuable comments and discussion during the research study. Meanwhile, I also want to thank all who helped and supported me during the completion of the dissertation.

Finally, I am highly grateful to my loving family – my grandfather, Mr. Keda Zhong, my grandmother, Ms. Yuru Zeng, my father, Mr. Binggang Zhong, my mother, Ms. Lifeng You and my fiancé, Jingyuan Liu. Without their love, support and understanding, I cannot complete the dissertation on my own.

This dissertation research was supported by National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIDA.

Dedication

This dissertation is dedicated to the memory of my loving father and hero,

Mr. Binggang Zhong.

He lives in my heart forever.

Chapter 1

Introduction

1.1 Background

Various regularization methods have been proposed for feature selection in high dimensional data analysis, which has become increasingly frequent and important in various research fields. These methods include, but are not limited to, the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001; Kim, Choi and Oh, 2008; Zou and Li, 2008), the LARS algorithm (Efron, Hastie, Johnstone and Tibshirani, 2004), the elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), the adaptive LASSO (Zou, 2006) and the Dantzig selector (Candes and Tao, 2007). All these methods allow the number of predictors to be greater than the sample size, and perform quite well for high dimensional data.

With the advent of modern technology for data collection, researchers are able to collect ultrahigh dimensional data at relatively low cost in diverse fields of scientific research. The aforementioned regularization methods may not perform well for ultrahigh dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009). These challenges call for new statistical modeling techniques for

ultrahigh dimensional data. Fan and Lv (2008) proposed the sure independence screening (SIS) and showed that the Pearson correlation ranking procedure possesses a sure screening property for linear regressions with Gaussian predictors and responses. That is, all truly important predictors can be selected with probability approaching one as the sample size diverges to ∞. Hall and Miller (2009) extended Pearson correlation learning by considering polynomial transformations of predictors. To rank the importance of each predictor, they suggested a bootstrap procedure. However, how to choose an optimal transformation remains an open issue and is often difficult. Fan, Samworth and Wu (2009) and Fan and Song (2010) proposed a more general version of independent learning which ranks the maximum marginal likelihood estimators or the maximum marginal likelihood for generalized linear models. Fan, Feng and Song (2011) considered nonparametric independence screening in sparse ultrahigh dimensional additive models. They suggested estimating the nonparametric components marginally with spline approximation, and ranking the importance of predictors using the magnitude of nonparametric components. They also demonstrated that this procedure possesses the sure screening property with vanishing false selection rate. Wang (2009) also proposed a variable screening method, called forward regression (FR), to identify the relevant predictors consistently even when $p \gg n$. Zhu, Li, Li and Zhu (2011) proposed a sure independent ranking and screening (SIRS) procedure to screen significant predictors in multi-index models. They further show that under linearity condition assumption on the predictor vector, the SIRS enjoys the ranking consistency property (i.e, the SIRS can rank the important predictors in the top asymptotically). Ji and Jin (2012) proposed the two-stage method: screening by Univariate thresholding and cleaning by Penalized least squares for Selecting variables, namely UPS. They further theoretically demonstrated that under certain

settings, the UPS can outperform the LASSO and subset selection, both of which are one-stage approaches. This motivates us to develop more effective screening procedures using two-stage approaches for ultrahigh dimensional data analysis.

1.2 Contribution

This dissertation consists of two main parts based on two research manuscripts. In Chapter 3, we propose a new feature screening procedure for ultrahigh dimensional data based on distance correlation. Szekely, Rizzo and Bakirov (2007) and Szekely and Rizzo (2009) showed that the distance correlation of two random vectors equals to zero if and only if these two random vectors are independent. Furthermore, the distance correlation of two univariate normal random variables is a strictly increasing function of the absolute value of the Pearson correlation of these two normal random variables. These two remarkable properties motivate us to use the distance correlation for feature screening in ultrahigh dimensional data. We refer to our Sure Independence Screening procedure based on the Distance Correlation as the DC-SIS. The DC-SIS can be implemented as easily as the SIS. It is equivalent to the SIS when both the response and predictor variables are normally distributed. However, the DC-SIS has appealing features that existing screening procedures including SIS do not possess. For instance, none of the aforementioned screening procedures can handle grouped predictors or multivariate responses. The proposed DC-SIS can be directly employed for screening grouped variables, and it can be directly utilized for ultrahigh dimensional data with multivariate responses. Feature screening for multivariate responses and/or grouped predictors is of great interest in pathway analyses. As in Chen, et al. (2011), pathway here means sets of proteins that are relevant to specific biological functions without regard to the state of knowledge concerning the interplay among such protein. Since proteins may work

interactively to perform various biological functions, pathway analyses complement the marginal association analyses for individual protein, and aim to detect a priori defined set of proteins that are associated with phenotypes of interest. There is a surged interest in pathway analyses in the recent literature (Ashburner, et al., 2000; Mootha, et al., 2003; Subramanian, et al., 2005; Tian, et al., 2005; Bild, et al., 2006; Efron and Tibsirani, 2007; Jones, et al., 2008). Thus, it is of importance to develop feature screening procedures for multivariate responses and/or grouped predictors.

We systematically study the theoretic properties of the DC-SIS, and prove that the DC-SIS possesses the sure screening property in the terminology of Fan and Lv (2008) under very general model settings including linear regression models, for which Fan and Lv (2008) established the sure screening property of the SIS. The sure screening property is a desirable property for feature screening in ultrahigh dimensional data. Even importantly, the DC-SIS can be used for screening features without specifying a regression model between the response and the predictors. Compared with the model-based screening procedures (Fan and Lv, 2008; Fan, Samworth and Wu, 2009; Wang, 2009; Fan and Song, 2010; Fan, Feng and Song, 2011), the DC-SIS is a model-free screening procedure. This virtue makes the proposed procedure robust to model mis-specification. This is a very appealing feature of the proposed procedure in that it may be very difficult in specifying an appropriate regression model for the response and the predictors with little information about the actual model in ultrahigh dimensional data.

We conduct Monte Carlo simulation studies to numerically compare the DC-SIS with the SIS and SIRS. Our simulation results indicate that the DC-SIS can significantly outperform the SIS and the SIRS under many model settings. We also assess the performance of the DC-SIS as a grouped variable screener, and the

simulation results show that the DC-SIS performs very well. We further examine the performance of the DC-SIS for feature screening in ultrahigh dimensional data with multivariate responses; simulation results demonstrate that screening features for multiple responses jointly may have dramatic advantage over screening features with each response separately. Fan and Lv (2008) developed an iterative SIS (ISIS) which performs much better than the SIS when some important predictors are marginally independent of the response. The development of DC-ISIS is indeed challenging and interesting because, unlike the SIS, we do not want to specify a regression model for the response and the predictors. Following Zhu, Li, Li and Zhu (2011), we further propose an iterative DC-SIS procedure (DC-ISIS). We also conduct a Monte Carlo simulation to examine the finite sample performance of DC-ISIS and demonstrate that the DC-ISIS is a dramatic improvement over the DC-SIS.

In Chapter 4, we consider heteroscedastic single-index model which assumes that both the mean and the variance functions of Y vary with the values of $\mathbf{x}$. To be precise, we assume that $E(Y \mid \mathbf{x}) = G(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathrm{var}(Y \mid \mathbf{x}) = \sigma^2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. Following the convention of the literature, we can write equivalently the heteroscedastic single-index model as follows,

$$Y = G(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \sigma(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\varepsilon. \tag{1.1}$$

For identification purpose, we assume the independent error term ε has zero mean and unit variance. Because both $G(\cdot)$ and $\sigma(\cdot)$ are unknown functions, the index parameter $\boldsymbol{\beta}$ is not identifiable. Thus, the direction of $\boldsymbol{\beta}$, rather than its true value, is of primary interest in the literature. Our goal is to identify the indices of the zero elements of $\boldsymbol{\beta}$ and to estimate the magnitudes of the nonzero elements of $\boldsymbol{\beta}$ up to a proportionality constant.

We develop a two-stage feature screening and variable selection procedure for ultrahigh dimensional heteroscedastic single-index models. In the first stage, a novel robust independent ranking and screening procedure (RIRS, for short) is proposed to reduce ultrahigh dimensionality down to a moderate scale for a general framework including heteroscedastic single-index model (1.1). Unlike the SIS (Fan and Lv, 2008) which utilizes the marginal Pearson correlation between each predictor and the response, the RIRS defines the marginal utility as the marginal Pearson correlation between each predictor and the rank of the response. Thus, the RIRS is insensitive to the extreme values and outliers in the response variable. Then, we establish both the ranking consistency property in the terminology of Zhu, Li, Li and Zhu (2011) and the sure screening property in the terminology of Fan and Lv (2008) for the proposed RIRS under mild conditions. Monte Carlo simulation studies and real data analysis demonstrate that the RIRS performs very well, especially in the presence of heteroscedasticity and outliers in the response, compared to the existing independence screening procedures, such as SIS by Fan and Lv (2008) and DC-SIS by Li, Zhong and Zhu (2012).

In the second stage, we propose to apply penalized linear quantile regression to further exclude unimportant covariates selected by the screening stage and to estimate the direction of the index parameter in the heteroscedastic single-index models. The penalized linear quantile regression inherits the merit of RIRS in the sense that it is also robust to the extreme values and outliers in the response. We remark here that, the underlying heteroscedastic single-index models are possibly nonlinear, however, the resultant estimator from penalized linear quantile regression is still consistent up to a proportionality constant, and has the oracle property in the terminology of Fan and Li (2001). The two-stage estimation procedure avoids estimating the nonlinear functions, and is very computationally

efficient in ultrahigh dimensional setting. We also demonstrate through comprehensive numerical studies that the whole procedure presents an outstanding finite sample performance.

1.3 Organization

The dissertation is organized as follows. In Chapter 2, the existing variable selection procedures for high dimensional data analysis and independence screening methods for ultrahigh dimensional data analysis are reviewed as well as their theoretic properties. The preliminary study on distance correlation is introduced in this chapter. In Chapter 3, we propose the DC-SIS and study its theoretic sure screening property. An iterative independence screening procedure (DC-ISIS) is also proposed to further enhance the finite sample performance. In Chapter 4, we propose a robust two-stage feature screening and variable selection procedure for ultrahigh dimensional heteroscedastic single-index models, and examine the theoretic properties and the finite sample performance via comprehensive numerical studies. Concluding remarks and future research are discussed in Chapter 5. Finally, the references and the curriculum vitae are provided in the end.

Chapter 2

Literature Review

Ultrahigh dimensional data analysis has become increasingly frequent and popular due to the modern technologies and methodologies of data collection in diverse scientific fields such as microarrays, genomics and finance. In the ultrahigh dimensional data, the number of predictors, say p, is usually much larger than the sample size, say n. That is, $p \gg n$. In particular, $p = O(\exp(\alpha n))$ with $\alpha > 0$. It is certainly challenging to select important predictors from ultrahigh dimensional candidates using well-established variable selection methods, such as best subset selection, the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001; Kim, Choi and Oh, 2008; Zou and Li, 2008) and among others. These existing methods may not perform well due to computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009).

Recently researchers advocated a two-stage variable selection procedure, screening and cleaning (Ji and Jin, 2012), to identify important predictors in the ultrahigh dimensional analysis. At the screening stage, marginal screening procedures are proposed to remove as many irrelevant variables as possible to reduce the dimensionality from ultrahigh p to a relatively large scale d, which can be less than n. Therefore, the marginal screening procedures can dramatically reduce the compu-

tational complexity. At the same time, the screening procedures can possess some favorable theoretic properties, such as the sure screening property (Fan and Lv, 2008) and the ranking consistency property (Zhu, Li, Li and Zhu, 2011). At the cleaning stage, the well-established variable selection methods, such as the LASSO and the SCAD, are proposed to simultaneously select significant variables from the remaining ones in the screening stage and estimate statistical effects of selected variables.

This chapter is organized as follows. Section 2.1 reviews well-established variable selection methods for linear regression models. Section 2.2 provides a brief review of the existing independence screening methods. Section 2.3 presents the definition of the distance correlation (Szekely, Rizzo and Bakirov, 2007) and its theoretic properties, based on which we will propose a new independence screening procedure in Chapter 3.

2.1 Variable Selection Approaches

Variable selection techniques play an increasingly important role in the high dimensional problems. Here, the high dimensionality means that $p = O(n^{\alpha})$ with $0 < \alpha < 1$, comparing with the ultrahigh dimensionality $p = O(\exp(\alpha n))$ with $\alpha > 0$. At the beginning stage of statistical modeling, it is typical to include as many as potential influential predictors into the model to reduce possible model bias. It is natural to assume that only a subset of predictors contribute to the response in the true model. Under this sparsity assumption, variable selection can improve both the prediction accuracy and the interpretability of the fitted model.

Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ is an $n \times p$ random design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an $n \times 1$ vector of independent and identically distributed(i.i.d.) random errors. When the dimension p is large, it is natural to assume the model is sparse. That is, only a small subset of predictors, say true predictors $\{X_j : \beta_j \neq 0, j = 1, \ldots, p\}$, contribute to the response Y.

2.1.1 Classic Variable Selection Criteria

A variable selection criterion is a statistic of a fitted model to measure the goodness of fit. In the past forty years, there are many literature covering this topic. For instance, Akaike (1973) proposed the Akaike's information criterion (AIC); Schwartz (1978) suggested the Bayesian information criterion (BIC); Craven and Wahba (1979) proposed the generalized cross validation statistic (GCV); Shao (1997) discussed the consistency and efficiency of variable selection and Miller (2002) provided a comprehensive review of the subsect selection in regression.

Residual Sum of Squares (RSS). For the linear regression model (2.1), many variable selection criteria are built on the residual sum of squares (RSS), which is defined by

$$RSS = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})^2, \tag{2.2}$$

where $\widehat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. Because $\mathbf{x}_i\widehat{\boldsymbol{\beta}}$ is the fitted value of the ith observation Y_i, RSS can measure the goodness of model fit.

R^2 and Adjusted R^2. R^2 is a commonly used statistic for model fitting , which

is defined based on RSS by,

$$R_d^2 = 1 - \frac{RSS_d}{RSS_0}, \tag{2.3}$$

where RSS_d is the residual sum of squares when an intercept and d predictors are fitted in the model, where $1 \leq d \leq p$, and RSS_0 is the RSS with only the intercept fitted. R^2 can measure how well the fitting of the d predictors is. However, R^2 increases with the number of predictors in the model. Therefore, R^2 cannot serve as a variable selection criterion. The adjusted R^2 can improve the performance of R^2 via adding a penalty on the increase of the number of predictors. The adjusted R^2 is also called Fisher's A-statistic, which is defined by

$$A_d = 1 - (1 - R_d^2)\frac{n-1}{n-d} = 1 - \frac{RSS_d/(n-d)}{RSS_0/(n-1)}. \tag{2.4}$$

Fisher's A-statistic A_d doesn't necessarily increase when a new predictor is added into the model. Therefore, it can be a variable selection criterion from an aspect of model fitting.

Prediction Sum of Squares (PRESS). Allen (1974) proposed a prediction-based variable selection criterion, the prediction sum of squares ($PRESS$). When the model includes d predictors, $PRESS_d$ is defined as

$$PRESS_d = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{id})^2, \tag{2.5}$$

where $\widehat{Y}_{id}$ is the predicted value of Y_i from the fitted model using all observations but ith one.

The idea of $PRESS$ can be generalized to the cross validation (CV). The idea of CV is that we randomly set a small number of observations (the testing set)

aside, and then use the remaining observations (the training set) to fit the model and predict the testing dataset, and summarize the performance of the prediction. For example, K-fold cross validation first partitions the data into K subsets with equal size n_K, and then we denote by $\mathbb{Y}_k$ the response in the kth subset and $\widehat{\mathbb{Y}}_k$ its predicted value based on the other $(K-1)$ subsets. Then, the K-fold CV, denoted by CV_K, is defined by

$$CV_k = \frac{1}{K}\sum_{k=1}^{K} \|\mathbb{Y}_k - \widehat{\mathbb{Y}}_k\|^2/n_K. \tag{2.6}$$

If we only set one observation aside each time, this so-called leave-one-out CV is essentially the $PRESS$. In practice, we can partition the dataset into K equivalent parts, leave one part out each time and predict this part using the remaining $K-1$ parts. Both CV and $PRESS$ can estimate the prediction errors of the fitted model and provide a good measure of how well the prediction of the proposed model is.

Craven and Wahba (1979) proposed the generalized cross validation statistic (GCV) for the linear regression model, which is defined by

$$GCV = \frac{RSS_d/n}{(1-d/n)^2}. \tag{2.7}$$

It is shown that under the mild conditions, the $PRESS$ statistic can be asymptotically approximated by

$$PRESS_d \approx \frac{n^2}{(n-d)^2} RSS_d = \frac{RSS_d}{(1-d/n)^2} = nGCV, \tag{2.8}$$

if n is much larger than d. Therefore, GCV is also a widely used variable selection criterion.

Akaike's information criterion (AIC). Akaike (1973) proposed the Akaike's in-

formation criterion (AIC) via considering the Kullback-Leibler mean information. It is defined as

$$AIC = RSS_d + 2d\sigma^2, \tag{2.9}$$

which is equivalent to the famous Mallows' C_p (Mallows (1973)) in the linear regression model. Mallows' C_p of the model with d predictors is defined by

$$C_p = \frac{RSS_d}{\sigma^2} - (n - 2d). \tag{2.10}$$

Bayesian information criterion (BIC). Schwartz (1978) suggested the Bayesian information criterion (BIC), which is defined by

$$BIC = RSS_d + \log(n)d\sigma^2. \tag{2.11}$$

In practice, we choose a model with the smallest information criterion to achieve variable selection. It can be shown that the BIC is a consistent criterion. That is, when assuming there exists a true model with finite parameters, the BIC can determine the true model as the sample size approaches the infinity. However, the AIC may provide an overfitted model. On the other hand, the AIC is an asymptotically loss efficient criterion (Shao, 1997), but the BIC is not.

2.1.2 Penalized Least Squares

It is demonstrated that the best subset selection with classic variable selection criteria can perform well in practice, but it suffers from the highly expensive computational cost, especially for the high dimensional regression models. Furthermore, the subset selection approaches are lack of stability and their theoretical properties

are difficult to examine (Breiman, 1996).

To overcome these drawbacks, the penalized least squares (PLS) methods were proposed for (2.1) via minimizing the following objective function $Q(\boldsymbol{\beta})$ to obtain the estimate $\widehat{\boldsymbol{\beta}}$:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.12}$$

where $p_\lambda(\cdot)$ is the penalty function and λ is the regularity parameter to control the size of the penalty.

In the rest of this section, we will discuss some well-known penalty functions as well as how to choose a good penalty function. Moreover, we will provide the connection between the penalized least squares (2.12) and the classic best subset selection and the ridge regression.

2.1.2.1 L_q Penalties with $0 \leq q \leq 2$

- **L_0 Penalty: Best Subset Selection.**

  The best subset selection with classic variable selection criteria can be written as the form of PLS with some L_0 penalty functions. Notice that choosing a model with the minimum C_p is equivalent to minimizing the following PLS object function

  $$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2\sum_{j=1}^{p} I(|\beta_j| \neq 0). \tag{2.13}$$

  Motivated by (2.13), the best subset selection with classic variable selection criteria is equivalent to minimizing the object function (2.12) with the

following L_0 penalty function

$$p_\lambda(|\beta_j|) = \frac{\lambda^2}{2} I(|\beta_j| \neq 0) \tag{2.14}$$

with different tunning parameters λ's.

For example, AIC is asymptotically equivalent to the following PLS

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\frac{(\sigma\sqrt{2/n})^2}{2}\sum_{j=1}^{p} I(|\beta_j| \neq 0). \tag{2.15}$$

with $\lambda = \sigma\sqrt{2/n}$.

BIC is asymptotically equivalent to the following PLS

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\frac{(\sigma\sqrt{\log(n)/n})^2}{2}\sum_{j=1}^{p} I(|\beta_j| \neq 0). \tag{2.16}$$

with $\lambda = \sigma\sqrt{\log(n)/n}$.

- L_2 **Penalty: Ridge Regression.**

  The well-known ridge regression was proposed by Hoerl and Kennard (1970) to deal with collinearity problem in predictors. Although ridge regression cannot possess the variable selection feature, it is also a solution of penalized least squares (2.12) with L_2 penalty. That is, $p_\lambda(|\beta_j|) = \frac{\lambda}{2}|\beta_j|^2$. Therefore, the ridge regression estimates can be obtain via minimizing the following PLS object function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{n\lambda}{2}\sum_{j=1}^{p} |\beta_j|^2. \tag{2.17}$$

  Like the ordinary least squares, the ridge regression also has the explicit

solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + n\lambda I_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{2.18}$$

where I_p is a $p \times p$ identity matrix.

- L_q **Penalty: Bridge Regression.**

  Frank and Friedman (1993) proposed the bridge regression with L_q penalty via minimizing

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{n\lambda}{q}\sum_{j=1}^{p}|\beta_j|^q, \tag{2.19}$$

  where $0 < q < 2$. L_q penalty bridges L_0 penalty and L_2 penalty.

- L_1 **Penalty: LASSO.**

  Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO) to shrink coefficients and select significant predictors. The LASSO solution is obtained by

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s, \tag{2.20}$$

  where the tuning parameter s controls the regularization size. It is equivalent to the penalized least squares with L_1 penalty

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\lambda\sum_{j=1}^{p}|\beta_j|. \tag{2.21}$$

  Therefore, the LASSO is a special case of the bridge regression with $q = 1$.

  The following Figure 2.1 shows the difference of solutions via the LASSO and

the ridge regression. In Figure 2.1(a), the LASSO solution is the first place that the contours touch the constrained square (shaded). When the touch happens at a corner, the LASSO produces a corresponding zero coefficient. Figure 2.1(b) shows that the ridge regression rarely obtains a zero solution, because the constrained circle provides no corner for contours to hit. Therefore, the LASSO can exactly shrink some coefficients to zero and hence gives a sparse model, while the ridge regression can only shrink coefficients.

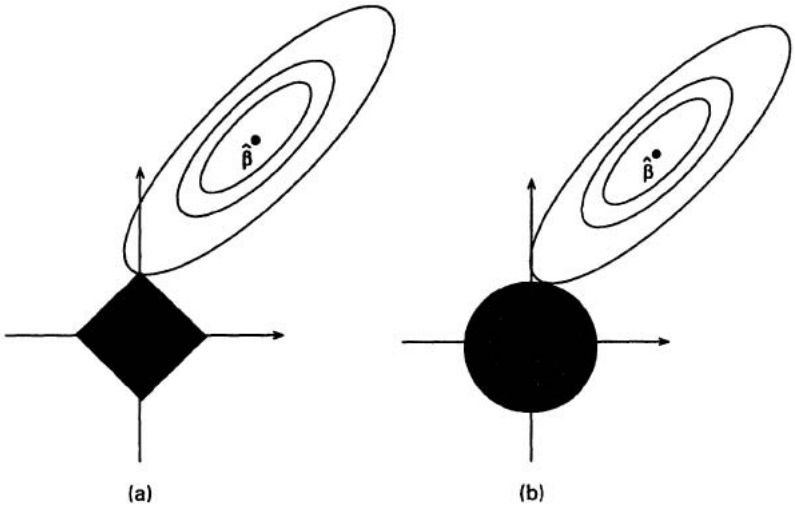

Figure 2.1. Estimation Pictures for (a) the LASSO and (b) the Ridge Regression (Tibshirani, 1996).

2.1.2.2 Seeking a Good Penalty Function

Penalized L_0 regression can conduct variable selection, but the computation is expensive and the result is unstable. Penalized L_2 regression (ridge regression) can shrink the estimated coefficients to make the result stable, but it cannot possess the variable selection feature. Penalized L_1 regression (LASSO) can provide shrinkage estimation and variable selection, but the estimators are biased even for the large true coefficients. The natural question: **What kind of penalty functions are good for variable selection and parameters estimation?**

Fan and Li (2001) advocated that good penalty functions should provide the estimators with the following three properties in the high dimensional regression problems:

(1) Unbiasedness: the penalized estimator should be nearly unbiased to reduce model bias, especially for the large true coefficients.

(2) Sparsity: the penalized estimator can automatically set small estimated coefficients to zero to achieve variable selection and reduce model complexity.

(3) Continuity: the penalized estimator is continuous in the data to avoid instability in model prediction.

To understand the above properties and discover a good penalty function, Fan and Li (2001) considered the linear regression model (2.1) with the design matrix $\mathbf{X}$ satisfying $\mathbf{X}^{\mathrm{T}}\mathbf{X} = nI_p$, where I_p is a $p \times p$ identity matrix. Then, (2.12) reduces to

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_0\|^2 + \frac{n}{2}\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\|^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \\ &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_0\|^2 + n\sum_{j=1}^{d}\left\{\frac{1}{2}(\widehat{\beta}_{0j} - \beta_j)^2 + p_\lambda(|\beta_j|)\right\}, \end{aligned} \tag{2.22}$$

where $\widehat{\boldsymbol{\beta}}_0 = \mathbf{X}^{\mathrm{T}}\mathbf{y}/n$ is the ordinary least square estimate. The first term of (2.22) is constant with respect to $\boldsymbol{\beta}$, so minimizing the object $Q(\boldsymbol{\beta})$ reduces to a componentwise regression problem. Consider the univariate minimization problem

$$\widehat{\theta}(z) = \arg\min_{\theta \in \mathbf{R}}\left\{\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)\right\}. \tag{2.23}$$

Antoniadis and Fan (2001), Fan and Li (2001) examined the conditions under which the univariate penalized estimator $\widehat{\theta}(z)$ can possess the above three properties:

(1) Unbiasedness if $p'_\lambda(|\theta|) = 0$ for large $|\theta|$,

(2) Sparsity if $\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} > 0$,

(3) Continuity if and only if $\underset{\theta\neq 0}{\operatorname{argmin}}\{|\theta| + p'_\lambda(|\theta|)\}=0$,

where $p_\lambda(\theta)$ is nondecreasing and continuously differentiable on $[0, \infty)$, and $p'_\lambda(0)$ means $p'_\lambda(0+)$ here. In general, a good penalty function $p_\lambda(\theta)$ should be singular at the origin to generate sparse estimators in variable selection, and concave when θ is large to reduce the model bias.

2.1.2.3 The SCAD Penalty

In the principle of a good penalty function, Fan and Li (2001) showed that the convex L_1 penalty (LASSO) does not satisfy the unbiasedness condition, so it increases the model bias. Furthermore, other L_q penalties can't satisfies all three conditions either. To construct a penalty function satisfying all three conditions, Fan and Li (2001) introduced the smoothly clipped absolute deviation (SCAD) penalty, which has the first derivative:

$$p'_\lambda(\theta) = \lambda\left\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\}, \tag{2.24}$$

where $a > 2$ is a constant and $\theta > 0$. The resulting PLS solution to (2.23) is given by

$$\widehat{\theta}(z) = \begin{cases} 0, & \text{if } |z| \leq \lambda \\ \operatorname{sgn}(z)(|z| - \lambda), & \text{if } \lambda < |z| \leq 2\lambda \\ \{(a-1)z - \operatorname{sgn}(z)a\lambda\}/(a-2), & \text{if } 2\lambda < |z| \leq a\lambda \\ z, & \text{if } |z| \geq a\lambda \end{cases}, \tag{2.25}$$

where $\operatorname{sgn}(\cdot)$ is the sign function.

To compare SCAD thresholding rule (2.25) with other rules, we introduce the well-known hard and soft thresholding rules here. Antoniadis (1997) and Fan (1997) introduced the hard thresholding penalty

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{2.26}$$

which leads to the hard thresholding rule

$$\widehat{\theta}(z) = zI(|z| > \lambda). \tag{2.27}$$

The LASSO (Tibshirani, 1996) yields a soft thresholding rule

$$\widehat{\theta}(z) = \text{sgn}(z)(|z| - \lambda)_+. \tag{2.28}$$

The Figure 2.2 displays the three different penalty functions.

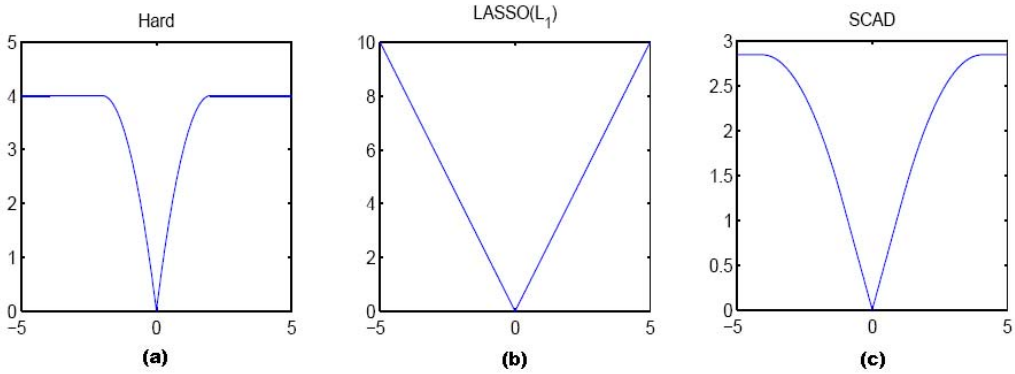

Figure 2.2. Plot of Penalty Functions for (a) the Hard Thresholding Penalty, (b) the LASSO (L_1) Penalty and (c) the SCAD Penalty with $\lambda = 2$ and $a = 3.7$.(Fan and Li, 2001).

The Figure 2.3 shows the three different thresholding rules. In Figure 2.3(a), the hard thresholding solution is sparse and unbiased for the large coefficients. However, it is not continuous and hence it is unstable. That is, a small change of

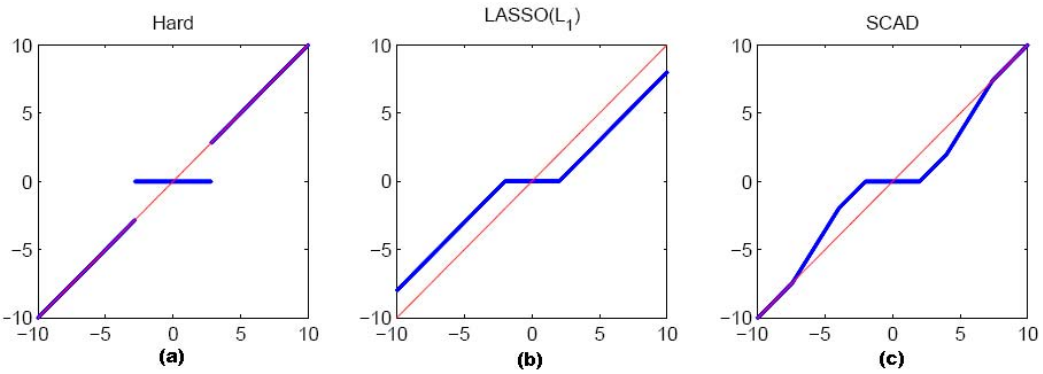

Figure 2.3. Plot of Thresholding Rules for (a) the Hard Thresholding, (b) the Soft (LASSO) Thresholding and (c) the SCAD Thresholding with $\lambda = 2$ and $a = 3.7$.(Fan and Li, 2001).

the data may result in a big change of the estimate. In Figure 2.3(b), the LASSO can provide the sparse model and the corresponding soft thresholding solution is continuous. But the resulting estimates are biased for the large coefficients. In Figure 2.3(c), the SCAD solution can satisfy all desirable properties and hence it is better than the other two rules.

To compare the performance of those thresholding estimators, we compute the corresponding risks $R(\theta) = E[\widehat{\theta}(Z) - \theta]^2$ in the fundamental model in which $Z \sim N(\theta, 1)$. Figure 2.4 shows the risk functions $R(\theta)$ for three commonly used penalty functions with $\lambda = 2$. Overall, the SCAD gives the smallest risk and performs better that the other two under this model setting.

Fan and Li (2001) has systematically studied the asymptotic oracle property of the proposed SCAD-penalized likelihood estimator. That is, the regularization estimation with the SCAD penalty works as well as if the correct submodel were known, when the regularization parameter is appropriately chosen. In other words, it can estimate the zero components to exactly zero with probability tending to 1, and the onezero components as well as when the correct submodel is known. We denote by $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ the true parameter with assuming that $\boldsymbol{\beta}_{20} = 0$. Let

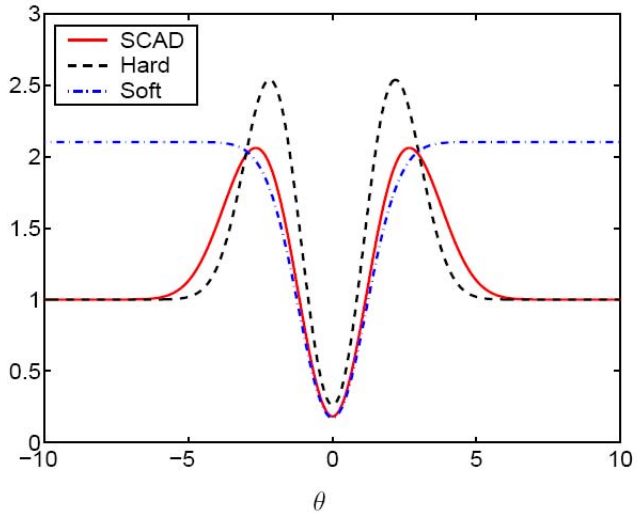

Figure 2.4. The Risk Functions for Penalized Least Squares under the Gaussian Model for the Hard Thresholding Penalty (Hard), the LASSO (Soft) and the SCAD With $\lambda = 2$ and $a = 3.7$ for the SCAD.(Fan and Li, 2001).

$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_2^{\mathrm{T}})^{\mathrm{T}}$ be the SCAD-penalized likelihood estimator. We further denote

$$\Sigma = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|), \ldots, p''_{\lambda_n}(|\beta_{s0}|)\}, \quad \text{and}$$

$$\mathbf{b} = (p'_{\lambda_n}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \ldots, p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0}))^{\mathrm{T}},$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. The following theorem presents the desirable oracle property of $\widehat{\boldsymbol{\beta}}$.

Theorem 2.1.1. (Oracle Property). *Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be independent and identically distributed, each with a density $f(\mathbf{X}_i, Y_i, \boldsymbol{\beta})$ satisfying regularity conditions (A)-(C) in Appendix of Fan and Li (2001). Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0_+} \left\{p'_{\lambda_n}(\theta)/\lambda_n\right\} > 0.$$

If $\lambda_n \to n$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the

root-n consistent $\widehat{\boldsymbol{\beta}}$ *must satisfy:*

(a) Sparsity: $\widehat{\boldsymbol{\beta}}_2 = 0$.

(b) Asymptotic Normality:

$$\sqrt{n}\left(I_1(\boldsymbol{\beta}_{10}) + \Sigma\right)\left\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \left(I_1(\boldsymbol{\beta}_{10}) + \Sigma\right)^{-1}\mathbf{b}\right\} \to N\left\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\right\},$$

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I(\boldsymbol{\beta}_{10}, \mathbf{0})$*, the Fisher information knowing* $\boldsymbol{\beta}_{20} = 0$.

Besides, there is a vast literature of penalty functions. For example, the adaptive LASSO proposed by Zou (2006), the elastic net which is a linear combination of L_1 and L_2 penalties in Zou and Hastie (2005) and minimax concave penalty (MCP) in Zhang (2010). For details, see the corresponding reference.

2.1.3 Computational Algorithms

- **LQA Algorithm.** When the convex penalty function (e.g. the L_1 penalty) is used, the object function (2.12) is convex and hence convex optimization algorithms can be applied. However, some penalty functions (e.g. the SCAD penalty) are used, and then the object is not convex any more. Fan and Li (2001) proposed a unified and effective local quadratic approximation (LQA) algorithm for optimizing nonconvex penalized object function. The idea is to use the quadratic curve to locally approximate the object function. To be specific, for a given initial value $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0})^{\mathrm{T}}$ which is not close to 0, the penalty function $p_\lambda(\cdot)$ can be locally approximated by a quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j, \quad \text{for } \beta_j \approx \beta_{j0}. \quad (2.29)$$

In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2), \quad \text{for } \beta_j \approx \beta_{j0}. \tag{2.30}$$

With the LQA, the object function (2.12) with nonconvex penalty becomes a convex function and admit a closed-form solution. The LQA algorithm set the sufficiently small coefficients to zero and hence produce a sparse model. But a drawback is that once a coefficient is shrunken to zero, it will remains zero in subsequent iterations.

- **LLA Algorithm.** Instead of using LQA, Zou and Li (2008) proposed a better approximation by using the local linear approximation (LLA):

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + p'_\lambda(|\beta_{j0}|)(|\beta_j| - |\beta_{j0}|), \quad \text{for } \beta_j \approx \beta_{j0}. \tag{2.31}$$

Figure 2.5 displays the local linear and local quadratic approximations to the SCAD penalty function. Figure 2.5 also shows that the LLA is the minimum convex majorant of the concave function on $[0, \infty)$. With the LLA, the object function (2.12) with a nonconvex penalty becomes an iteratively reweighted penalized L_1 regression. See Zou and Li (2008) for details.

- **LARS Algorithm.** Efron, Hastie, Johnstone and Tibshirani (2004) proposed the least angle regression (LARS) algorithm for penalized variable selection. This fast and efficient algorithm can produce the entire LASSO solution path $\{\widehat{\boldsymbol{\beta}}(\lambda), \lambda > 0\}$, which is piecewise linear in λ. See Efron, Hastie, Johnstone and Tibshirani (2004) for details.

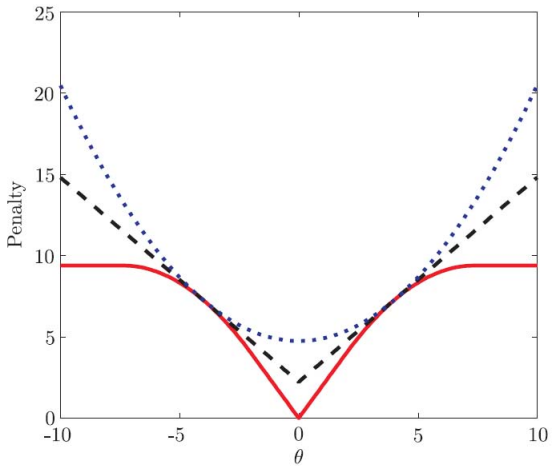

Figure 2.5. The Local Quadratic (dotted) and the Local Linear (dashed) Approximations to the SCAD Penalty Function (solid) With $\lambda = 2$ and $a = 3.7$ at a given point $|\theta| = 4$.(Zou and Li, 2008).

2.2 Independence Screening Procedures

2.2.1 Sure Independence Screening

For ultrahigh dimensional linear regression model, Fan and Lv (2008) proposed the *Sure Independence Screening* procedure via Pearson correlation learning(SIS, for short) to reduce the ultrahigh dimension down to a relative large scale.

Consider the same linear regression model as (2.1):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Under the sparsity assumption, denote the true model as $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ with the model size $s = |\mathcal{M}_*|$, where $|\mathcal{M}_*|$ represents the number of elements in the set $\mathcal{M}_*$. Then denote by $\mathbf{X}_s$ the design matrix standardized columnwisely and define $\omega = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ as follows

$$\omega = \mathbf{X}_s^{\mathrm{T}}\mathbf{y}. \tag{2.32}$$

Note that ω_j is the marginal Pearson correlation between the jth predictor X_j and the response Y scaled by the standard deviation of the response. On the other hand, ω_j can also be considered as the least square estimated coefficient for standardized X_j in the marginal regression $\mathbf{y} = X_j\beta_j + \boldsymbol{\varepsilon}$. Therefore, $|\omega_j|$ can characterize the magnitude of marginal relationship between the predictor X_j and the response Y.

The SIS ranks the importance of all predictors according to $|\omega_j|$ and removes those predictors weakly correlated with the response Y, i.e., ones with small absolute values of ω_j. To be specific, for any given $\gamma \in (0, 1)$, the SIS selects predictors with the first $[\gamma n]$ largest $|\omega_j|$ and defines the submodel

$$\widehat{\mathcal{M}}_\gamma = \{1 \le j \le p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.33}$$

where $[\gamma n]$ denotes the integer part of γn.

Sure Screening Property. Define

$$\mathbf{z} = \boldsymbol{\Sigma}^{-\mathbf{1/2}}\mathbf{x}, \text{ and } \mathbf{Z} = \mathbf{X}\boldsymbol{\Sigma}^{-\mathbf{1/2}}, \tag{2.34}$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$. Fan and Lv (2008) imposed the following five conditions/assumptions to establish the sure screening property of the SIS.

(C1) $p > n$ and $\log p = O(n^\xi)$ for some $\xi > 0$.

(C2) $\mathbf{z}$ has a spherically symmetric distribution (Chmielewski, 1981) and the random matrix $\mathbf{Z}$ is has the concentration property. That is, there exist some constants $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P\left(\lambda_{\max}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\mathrm{T}}) > c_1 \text{ and } \lambda_{\min}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\mathrm{T}}) > 1/c_1\right) \le e^{-C_1 n}, \tag{2.35}$$

holds for any $n \times \widetilde{p}$ submatrix $\widetilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \widetilde{p} \leq p$.

(C3) $\text{var}(Y) = O(1)$ and for some $\kappa \geq 0$ and $c_2, c_3 > 0$,

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geq \frac{c_2}{n^\kappa} \text{ and } \min_{i \in \mathcal{M}_*} |\text{cov}(\beta_i^{-1} Y, X_i)| \geq c_3. \tag{2.36}$$

(C4) For some $\tau \geq 0$ and $c_4 > 0$ such that

$$\lambda_{\max}(\Sigma) \leq c_4 n^\tau. \tag{2.37}$$

(C5) $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

Remark: Condition (C1) shows that the proposed SIS is suitable for the ultra-high dimensional problem. Although there is no explicit restriction on ξ, the concentration property in condition (C2) makes restriction on ξ. Condition (C3) removes the case in which a significant variable is marginally uncorrelated with the response Y, but jointly correlated with Y. Condition (C4) rules out the situation of strong collinearity among predictors. Condition (C5) provides a restriction on the error distribution.

Theorem 2.2.1. (Sure Screening Property) *Under above conditions (C1)-(C5), if $2\kappa + \tau < 1$ then there exists some $\theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$, assume the true model size $s \leq [\gamma n]$, we have for some $C > 0$,*

$$\mathbf{P}(\mathcal{M}_* \subset \widehat{\mathcal{M}}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n)), \tag{2.38}$$

therefore,

$$\mathbf{P}(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_\gamma) \to 1, \ \ as \ \ n \to \infty. \tag{2.39}$$

Theorem 2.2.1 shows that the proposed SIS can efficiently shrink the ultrahigh dimension p down to a relatively large scale $d = [\gamma n] = O(n^{1-\theta})$ for some $\theta > 0$, while all truly important predictors can be selected into the submodel $\widehat{\mathcal{M}}_\gamma$ with probability approaching one as the sample size tends to ∞.

2.2.2 Generalized Correlation Ranking

Sure independence screening via Pearson correlation learning (Fan and Lv, 2008) can perform well in the ultrahigh dimensional linear regression model. However, Pearson correlation can only capture the linear relationship between each predictor X_j and the response Y. When Pearson correlation $\rho(X_j, Y)$ is zero, it only means that the response Y is linearly uncorrelated with the predictor X_j. If the predictor X_j is nonlinearly but not linearly influential to the response Y, the SIS is most likely to miss this important predictor. In order to capture the nonlinearity in the ultrahigh dimensional problems, Hall and Miller (2009) suggested techniques based on ranking generalized empirical correlation between the response Y and each predictor X_j, which can capture both linearity and nonlinearity.

Hall and Miller (2009) defined the *generalized correlation* between two random variables X and Y as

$$\rho_g(X, Y) = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X), Y\}}{\sqrt{\text{var}\{h(X)\}\text{var}(Y)}}, \tag{2.40}$$

where $\mathcal{H}$ is a class of functions including all linear functions. For example, it is a class of polynomial functions up to a given degree. Remark that if $\mathcal{H}$ is restricted to be a class of all linear functions, $\rho_g(X, Y)$ is the absolute value of Pearson correlation $\rho(X, Y)$. Therefore, $\rho_g(X, Y)$ can be naturally considered as a generalization of the conventional Pearson correlation.

Assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent and identically

distributed observed pairs of two random variables X and Y. The generalized correlation $\rho_g(X, Y)$ between X and Y can be estimated by

$$\widehat{\rho_g}(X, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^{n} \{h(X_i) - \bar{h}\}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} \{h(X_i)^2 - \bar{h}^2\} \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}, \tag{2.41}$$

where $\bar{h} = n^{-1} \sum_{i=1}^{n} h(X_i)$ and $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

The proposed generalized correlation characterizes both linear and nonlinear relationships between two random variables. Therefore, the generalized correlation $\rho_g(X_j, Y)$ can be considered as a marginal utility to measure the influential effort of the predictor X_j on the response Y. In practice, Hall and Miller (2009) suggested to rank the predictors based on the magnitude of estimated generalized correlation $\widehat{\rho_g}(X_j, Y)$. In the result, one orders $\widehat{\rho_g}(X_j, Y)$ as $\widehat{\rho_g}(X_{\widehat{j}_1}, Y) \geq \widehat{\rho_g}(X_{\widehat{j}_2}, Y) \geq \ldots \geq \widehat{\rho_g}(X_{\widehat{j}_p}, Y)$ and takes

$$\widehat{j}_1 \succeq \widehat{j}_2 \succeq \ldots \succeq \widehat{j}_p$$

to denote the empirical ranking of the indices of all predictors. Intuitively, the higher ranking the predictor has, the more important it is on the response in term of the generalized correlation. Therefore, given a suitable cutoff, one can select predictors with higher rankings and thus reduce the ultrahigh dimensionality to a relatively low scale.

Hall and Miller (2009) suggested the bootstrap procedure to choose a cut-off. Let $S = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be the original dataset, and $S^* = \{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_n^*, Y_n^*)\}$ be a resample drawn randomly from S with replacement. Denote $r(j)$ be the ranking of the jth predictor X_j such as $\widehat{j}_{r(j)} = j$. Let $r^*(j)$ be the ranking of X_j using the bootstrapped resample S^*. Given a value α, such as 0.05, one may compute a nominal $(1 - \alpha)$-level two-sided prediction interval of the ranking, $[\widehat{r}_-(j), \widehat{r}_+(j)]$, based on the bootstrapped $r^*(j)$'s. Hall

and Miller (2009) proposed a criterion to regard the predictor X_j as influential if $\widehat{r}_+(j) < \frac{1}{2}p$. In practice, the cutoff can also be replaced by some smaller fraction of p, such as $\frac{1}{4}p$. Therefore, the proposed generalized correlation ranking reduces the ultrahigh p down to the size of the selected model $\widehat{\mathcal{M}}_k = \{j : \widehat{r}_+(j) < kp\}$, where $0 < k < 1/2$ is a constant multiplier to control the size of the selected model $\widehat{\mathcal{M}}_k$.

To present a theoretical property of the proposed generalized correlation ranking, Hall and Miller (2009) imposed the following assumptions:

(D1) $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent and identically distributed;

(D2) $\mathcal{H}$ is the class of polynomial functions up to a given degree $d \geq 1$;

(D3) $\text{var}\{h(X_{ij})\} = \text{var}(Y_i) = 1$ for all i and j;

(D4) for constants $\gamma > 0$, $c > 0$ and sufficiently large n, $p \leq cn^{\gamma}$;

(D5) $\sup_n \max_{j \leq p} E|X_{1j}|^C < \infty$, $\sup_n E|Y_1|^C < \infty$, for a constant $C > 4d(\gamma + 1)$.

Given constants $0 < c_1 < c_2 < \infty$, define $\mathcal{I}_1(c_1) = \{j : |\text{cov}(X_j, Y)| \leq c_1\sqrt{\log n/n}\}$ and $\mathcal{I}_2(c_2) = \{j : |\text{cov}(X_j, Y)| \geq c_2\sqrt{\log n/n}\}$.

Theorem 2.2.2. *Under assumptions (D1)-(D5), for sufficiently small c_1 and sufficiently large c_2, in the correlation-based ranking $\widehat{j}_1 \succeq \widehat{j}_2 \succeq \ldots \succeq \widehat{j}_p$, all the indices in $\mathcal{I}_2(c_2)$ are listed before any of the indices in $\mathcal{I}_1(c_1)$ with the probability converging to 1 as $n \to \infty$.*

Theorem 2.2.2 shows that variables with sufficiently large covariance with the response Y in the order of $\sqrt{\log n/n}$ are very likely to be ranked ahead of those with smaller covariances.

2.2.3 Sure Independence Screening for GLM

The SIS procedure (Fan and Lv, 2008) provides a possibility to handle the ultrahigh dimensional problems. However, the SIS only restricts to the ordinary linear regression model and the theoretical properties rely heavily on the joint normality assumptions on the response and predictors. This limits significantly its applicability for categorical variables, even within the context of linear models.

To this end, Fan and Song (2010) proposed a more general version of sure independence screening procedure for generalized linear models. They considered the maximum marginal likelihood estimator (the MMLE, for short) or the marginal likelihood ratio as a marginal utility to rank the importance of each predictor. The conditions under which the proposed MMLE possesses the sure screening property are also explored. Moreover, Fan and Song (2010) discussed the size of the selected model and false positive rate controlling.

First, consider the generalized linear model (GLM) with canonical link. That is, the response variable Y conditional on the predictors $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is from an exponential family, whose probability density function takes the canonical form

$$f_{Y|\mathbf{x}}(y|\mathbf{x}) = \exp\left\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(y)\right\}, \tag{2.42}$$

for some known functions $b(\cdot)$, $c(\cdot)$ and $\theta(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. Without loss of generality, assume that the dispersion parameter $\phi = 1$ and each predictor is standardized with mean 0 and variance 1. Therefore, the log-likelihood for the natural parameter θ of the GLM is

$$\ell(\theta, y) = b(\theta) - y\theta. \tag{2.43}$$

Parallel to Fan and Lv (2008), let $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the true

model with the model size $s = |\mathcal{M}_*|$. Fan and Song (2010) defined the maximum marginal likelihood estimator (MMLE) $\widehat{\boldsymbol{\beta}}_j^M$ of the jth predictor X_j as

$$\widehat{\boldsymbol{\beta}}_j^M = (\widehat{\beta}_{j,0}^M, \widehat{\beta}_j^M) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \ell(\beta_0 + \beta_1 X_{ij}, Y_i), \tag{2.44}$$

where Y_i is the ith observed response and X_{ij} is the ith observation of the jth predictor. Although the MMLE $\widehat{\beta}_j^M$ is a incorrectly estimated coefficient for jth predictor X_j in the joint model, the $\widehat{\beta}_j^M$ can preserve useful non-sparsity information of X_j in the joint model for variables screening under some mild conditions. Therefore, it is reasonable to consider the magnitude of $\widehat{\beta}_j^M$ as a marginal utility to rank the importance of X_j and select a submodel, given a prespecified threshold γ_n,

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p : |\widehat{\beta}_j^M| \ge \gamma_n\}. \tag{2.45}$$

To establish the theoretical properties of MMLE, Fan and Song (2010) denoted the population version of the marginal likelihood maximizer as

$$\boldsymbol{\beta}_j^M = (\beta_{j,0}^M, \beta_j^M) = \arg\min_{\beta_0,\beta_1} E\ell(\beta_0 + \beta_1 X_{ij}, Y_i),$$

and provided the following conditions:

(D1) The marginal Fisher information: $I_j(\boldsymbol{\beta}_j) = E\{b''(X_j^{\mathrm{T}}\boldsymbol{\beta}_j)X_j X_j^{\mathrm{T}}\}$ is finite and positive definite at $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^M$, for $j = 1, \ldots, p$. Moreover, $\|I_j(\boldsymbol{\beta}_j)\|_{\mathcal{B}} = \sup_{\boldsymbol{\beta}\in\mathcal{B},\|\mathbf{x}\|=1} \|I(\boldsymbol{\beta})^{1/2}\mathbf{x}\|$ is bounded from above, where $\mathcal{B} = \{|\beta_{j,0}^M| \le B, |\beta_j^M| \le B\}$ is a square with the width B.

(D2) $b''(\theta)$ is continuous and positive. There exists an $\varepsilon_1 > 0$ such that for some

sufficiently large positive constants K_n and all $j = 1, \ldots, p$,

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|\leq\varepsilon_1} |Eb(X_j^{\mathrm{T}}\boldsymbol{\beta})I(|X_j| > K_n)| \leq o(n^{-1}).$$

(D3) For all $\boldsymbol{\beta}_j \in \mathcal{B}$, $E(l(X_j^{\mathrm{T}}\boldsymbol{\beta}_j, Y) - l(X_j^{\mathrm{T}}\boldsymbol{\beta}_j^M, Y)) \leq V\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^M\|^2$, for some $V > 0$, bounded from below uniformly over $j = 1, \ldots, p$.

(D4) There exists some positive constants m_0, m_1, s_0, s_1 and α, such that for sufficiently large t,

$$P(|X_j| > t) \leq (m_1 - s_1)\exp\{-m_0 t^\alpha\}, \quad \text{for } j = 1, \ldots, p,$$

and that

$$E\exp(b(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + s_0) - b(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})) + E\exp(b(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - s_0) - b(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})) \leq s_1.$$

(D5) $|\mathrm{cov}(b'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), X_j)| \leq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$ and a constant $c_1 > 0$.

Theorem 2.2.3. (Sure Screening Property) *Assume above conditions (D1)-(D4) hold, then*

(i) If $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, where $k_n = b'(K_n B + B) + m_0 K_n^\alpha / s_0$, then for any $c_3 > 0$, there exists a constant $c_4 > 0$ such that

$$\begin{aligned}&\mathbf{P}\left(\max_{1\leq j\leq p} |\widehat{\beta}_j^M - \beta_j^M| \geq c_3 n^{-\kappa}\right)\\ &\leq p\left\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^\alpha)\right\}. \qquad (2.46)\end{aligned}$$

(ii) In addition, condition (D5) holds, then by taking $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \leq c_2/2$,

the following inequality holds,

$$\mathbf{P}(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \leq 1 - s\left\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha})\right\} \quad (2.47)$$

Theorem 2.2.3 shows that the MMLEs are uniformly convergent to the population values and establishes the sure screening property of the MMLE screening procedure. Fan and Song (2010) also indicated that the MMLE can handle the ultrahigh NP-dimensionality:

$$\log p = o(n^{(1-2\kappa)\alpha/(\alpha+2)}).$$

Specifically, it can deal with the NP-dimensionality as high as $\log p = o(n^{(1-2\kappa)})$ for the logistic model with bounded predictors, and $\log p = o(n^{(1-2\kappa)/4})$ for the ordinary linear model without the joint normality assumption.

Fan and Song (2010) further discussed how large the selected model $\widehat{\mathcal{M}}_{\gamma_n}$ is. Under some regularity conditions, they showed that with probability approaching one,

$$|\widehat{\mathcal{M}}_{\gamma_n}| = O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}, \quad (2.48)$$

where κ is a constant in condition (D5), which determines how large the thresholding parameter γ_n is, and $\lambda_{\max}(\Sigma)$ is the maximum eigenvalue of the covariance matrix Σ of predictors $\mathbf{x}$, which controls how correlated the predictors are. If $\lambda_{\max}(\Sigma) = O(n^{\tau})$, the size of $\widehat{\mathcal{M}}_{\gamma_n}$ has the order $O(n^{2\kappa+\tau})$, which can guide practitioners to choose the thresholding rule.

In addition, Fan and Song (2010) also proposed another feature screening method via using the marginal likelihood ratio test, which was demonstrated to

have the same spirit as the MMLE screening. For details, one can refer to Fan and Song (2010).

2.2.4 Sure Independent Ranking and Screening

Zhu, Li, Li and Zhu (2011) proposed a model-free feature screening, called *sure independent ranking and screening* (SIRS), for ultrahigh dimensional data. Compared with the SIS and other model-based sure independence screening approaches, the SIRS works for a very general model framework including many commonly used parametric and semiparametric models, including the linear model, the generalized linear model, the index model and others. Therefore, the proposed SIRS is more robust to possible model mis-specification and can be considered as model-free.

Let Ψ_y be the support of the response Y and denote the conditional distribution function of Y given $\mathbf{x}$ as $F(y|\mathbf{x}) = P(Y \leq y|\mathbf{x})$. Define the indices sets of active predictors and inactive predictors, respectively, by

$$\begin{aligned} \mathcal{A} &= \{k : F(y \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(y \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } y \in \Psi_y\}. \end{aligned}$$

X_k for $k \in \mathcal{A}$ is called an active predictor, whereas X_k for $k \in \mathcal{I}$ is called an inactive predictor. Without of generality, assume that the first p_1 predictors are active and the rest $p-p_1$ predictors are inactive. In other words, $\mathcal{A} = \{1, 2, \ldots, p_1\}$ and $\mathcal{I} = \{p_1 + 1, \ldots, p\}$.

Considering a general model framework, we assume that $F(y \mid \mathbf{x})$ depends on $\mathbf{x}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$ for some $p_1 \times K$ constant matrix $\boldsymbol{\beta}$. That is,

$$F(y \mid \mathbf{x}) = F_0(y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}), \tag{2.49}$$

where $F_0(\cdot \mid \cdot)$ is an unknown function.

Without of generality, assume that $E(X_k) = 0$ and $\text{var}(X_k) = 1$ for $k = 1, \ldots, p$. Define

$$\Omega(y) = E\{\mathbf{x}F(y|\mathbf{x})\} = E\{\mathbf{x}E[\mathbf{1}(Y \leq y)|\mathbf{x}]\} = \text{cov}\{\mathbf{x}, \mathbf{1}(Y \leq y)\}.$$

Then define a new marginal utility ω_k at the population level by

$$\omega_k = E\{\Omega_k^2(Y)\}, \quad k = 1, \ldots, p, \tag{2.50}$$

where $\Omega_k(y)$ is the kth element of $\Omega(y)$. Intuitively, if X_k and Y are independent, then X_k and $\mathbf{1}(Y \leq y)$ for any $y \in \Psi_y$ are independent resulting so that $\omega_k = 0$. On the other hand, if X_k and Y are correlated, then X_k and $\mathbf{1}(Y \leq y)$ for some $y \in \Psi_y$ are correlated and thus $\omega_k > 0$.

For a random sample $\{(X_{i1}, \ldots, X_{ip}, Y_i), i = 1, \ldots, n\}$ from $\{\mathbf{x}, Y\}$, the sample moment estimator of ω_k is derived by

$$\widehat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n} \widehat{\Omega}_k^2(Y_j) = \frac{1}{n}\sum_{j=1}^{n} \left\{ \frac{1}{n}\sum_{i=1}^{n} X_{ik}\mathbf{1}(Y_i \leq Y_j) \right\}^2, \quad k = 1, \ldots, p, \tag{2.51}$$

Zhu, Li, Li and Zhu (2011) suggested to employ the sample estimate of ω_k to rank all the candidate predictors, and select the top ones as the estimate of the active predictors.

Ranking consistency property. To demonstrate the utility of the proposed SIRS, Zhu, Li, Li and Zhu (2011) established the consistency in ranking of the SIRS based on the following conditions:

(F1) The following inequality condition holds uniformly for p:

$$\frac{K^2\lambda_{\max}\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}}{\lambda_{\min}^2\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\}} < \frac{\min_{k\in\mathcal{A}}\omega_k}{\lambda_{\max}\{\Omega_{\mathcal{A}}\}}, \tag{2.52}$$

where $\Omega_{\mathcal{A}} = E\{\Omega_{\mathcal{A}}(Y)\Omega_{\mathcal{A}}^{\mathrm{T}}(Y)\}$, $\Omega_{\mathcal{A}}(y) = \{\Omega_1(y), \ldots, \Omega_{p_1}(y)\}^{\mathrm{T}}$, and $\lambda_{\max}\{M\}$ and $\lambda_{\min}\{M\}$ represent the maximum and minimum eigenvalues of a matrix M, respectively. Note that $\lambda_{\max}\{M\}$ and $\lambda_{\min}\{M\}$ may depend on the dimension of M. Throughout this dissertation, "$a < b$ holds uniformly for p" means that "$\limsup_{p\to\infty}\{a(p) - b(p)\} < 0$".

(F2) The linearity condition:

$$E\{\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}\} = \mathrm{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}\{\mathrm{cov}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})\}^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}$$

(F3) The moment condition: there exists a constant $t_0 > 0$ such that

$$\max_{1\le k\le p} E\{\exp(tX_k)\} < \infty, \quad \text{for } 0 < t \le t_0.$$

Remark: Condition (F1) provides an assumption on the correlations among the predictors. Condition (F2) holds if $\mathbf{x}$ is normal or follows an elliptically symmetric distribution. Furthermore, when the number of predictors p diverges while the dimension K is fixed, the linearity condition (F2) can hold asymptotically. Condition (F3) assumes that all moments of the predictors are uniformly bounded. Therefore, condition (F3) holds for the normal distribution and others with bounded support.

Theorem 2.2.4. *Under Conditions (F1)-(F3), the following inequality holds uni-*

formly for p:

$$\max_{k\in\mathcal{I}}\omega_k < \min_{k\in\mathcal{A}}\omega_k. \tag{2.53}$$

Theorem 2.2.4 shows that ω_k of an inactive predictor is always smaller than ω_k of an active predictor, which will provide a theoretical separation of predictor ranking.

Theorem 2.2.5. (Consistency in Ranking) *In addition to Conditions (F1)-(F3), assume that* $p = o\{\exp(an)\}$ *for* $a > 0$. *Then, for any* $\varepsilon > 0$, *there exists a sufficiently small constant* $s_\varepsilon > 0$ *such that*

$$\mathbf{P}\left(\sup_{k=1,\ldots,p}|\widehat{\omega}_k - \omega_k| > \varepsilon\right) \le 2p\exp\{n\log(1-\varepsilon s_\varepsilon/2)/3\}. \tag{2.54}$$

In addition, if let $\delta = \min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k$, *then there exists a sufficiently small constant* $s_\delta > 0$ *such that*

$$\mathbf{P}\left(\max_{k\in\mathcal{I}}\widehat{\omega}_k < \min_{k\in\mathcal{A}}\widehat{\omega}_k\right) \le 2p\exp\{n\log(1-\delta s_\delta/4)/3\}. \tag{2.55}$$

Theorem 2.2.5 demonstrates the ranking consistency property of the proposed SIRS. That is, the SIRS screening method using $\widehat{\omega}_k$ always ranks an active predictor ahead of an inactive one with the probability tending to one. Further, consistency in ranking provides a clear separation between the active and inactive predictors. Thus, with an appropriate cutoff, the SIRS can be consistent in selection in the ultrahigh dimensional problems.

For the independence screening, Fan and Lv (2008) suggested a hard thresholding rule to choose the top variables in the order of $O(n/\log n)$. Besides, Zhu, Li, Li and Zhu (2011) recommended a soft thresholding rule based on adding arti-

ficial auxiliary variables to the data. First, randomly generate q auxiliary variables $\{Z_1, \ldots, Z_q\}$ which are independent of both $\mathbf{x}$ and $\mathbf{Y}$. Then, consider the $(p+q)$ dimensional vector $(X_1, \ldots, X_p, Z_1, \ldots, Z_q)$ as the predictors and apply the independence screening method to pick top variables. In details, denote ω_k as the marginal utility for kth predictor for $k = 1, \ldots, p+q$. Because $\{Z_1, \ldots, Z_q\}$ are truly inactive, $\max_{l=1,\ldots,q} \omega_{p+l} < \min_{k \in \mathcal{A}} \omega_k$ holds by Theorem 2.2.4 and under some mild conditions, $\max_{l=1,\ldots,q} \widehat{\omega}_{p+l} < \min_{k \in \mathcal{A}} \widehat{\omega}_k$ holds with probability tending to one by Theorem 2.2.5. Then select the predictor subset

$$\widehat{\mathcal{M}}_s = \{k : \widehat{\omega}_k > \max_{l=1,\ldots,q} \widehat{\omega}_{p+l}\} \tag{2.56}$$

Zhu, Li, Li and Zhu (2011) also gave an upper bound on the probability of selecting any inactive predictors by this soft thresholding rule.

Theorem 2.2.6. *Assume the exchangeability condition. That is,* $\{X_j, j \in \mathcal{I}\}$ *and* $\{Z_j, j = 1, \ldots, q\}$ *are exchangeable in the sense that both any* X_j *for* $j \in \mathcal{I}$ *and* Z_j *are equally likely to be selected by the soft thresholding procedure. Then, for* $r \in \mathbb{N}_+$,

$$P\left(|\widehat{\mathcal{M}}_s \cap \mathcal{I}| \geq r\right) \leq \left(1 - \frac{r}{p+q}\right)^q. \tag{2.57}$$

Zhu, Li, Li and Zhu (2011) suggested to choose $q = p$ empirically and used numerical studies to show that the soft thresholding rule with this choice can work quite well. For details about the SIRS, one can refer to Zhu, Li, Li and Zhu (2011).

2.2.5 Extensions of Independence Screening

2.2.5.1 Iterative Version of Independence Screening

Fan and Lv (2008) has shown that the SIS can perform very well when the conditions are satisfied. However, when these restrictive conditions fail, the SIS procedure may be problematic. For example, when a variable is jointly correlated, but marginally uncorrelated with the response, the SIS is unlikely to select this important variable, resulting in high false negative rate. On the other hand, when a variable is jointly uncorrelated but highly marginally correlated with the response, the SIS is likely to select this unimportant variable, resulting in high false positive rate. To overcome this problem, Fan and Lv (2008) provided an important methodological extension of the SIS, called the *Iterative Sure Independence Screening* (the ISIS, for short).

The steps of the ISIS procedure are provided as follows:

Step.1 Apply the SIS to the full dataset and select an indices set $\widehat{\mathcal{A}}_1$ of size $d = [n/\log n]$. Then implement the variable selection approaches, such as penalized least square with SCAD penalty, on the indices set $\widehat{\mathcal{A}}_1$ to select a submodel $\widehat{\mathcal{M}}_1$. Let $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_1$.

Step.2 Compute the residuals from regressing the response Y over $\{X_j : j \in \widehat{\mathcal{M}}\}$. Then treat these residuals as the new responses and apply the same procedure in Step 1 to the remaining variables with indices $\{1, \ldots, p\}\backslash\widehat{\mathcal{M}}$ to obtain another submodel $\widehat{\mathcal{M}}_2$. Let $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{M}}_2$.

Step.3 Iterate the process until $|\widehat{\mathcal{M}}| \leq d'$, where d' is the prescribed number and $d' \leq n$. The indices set $\widehat{\mathcal{M}}$ is the final selected submodel by the ISIS.

The ISIS procedure has been empirically proved by Fan and Lv (2008) that it can perform better than the ordinary SIS. Besides this version of iterative inde-

pendence screening, Fan, Samworth and Wu (2009) extended the idea to a model general version using the marginal likelihood to rank the importance of variables; Fan, Feng and Song (2011) provided iterative nonparametric independence screening for the sparse ultrahigh dimensional additive models; Zhu, Li, Li and Zhu (2011) also created an iterative version of the model-free independence screening with iteratively transforming the space of predictors.

2.2.5.2 Reduction of False Positive Rate

The independence screening procedures are commonly used for feature selection, but they are usually conservative and result in many false positive variables. Fan, Samworth and Wu (2009) proposed a simple resampling technique to reduce the false positive rate.

Let $\mathcal{A}$ be the set of active indices. Partition the samples randomly into two parts with the same sample size, and then apply one independence screening, such as the SIS and the ISIS, to two halves. Denote $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ as the selected submodel based on the first half and the second half of the samples, respectively. Under some conditions, both $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ possess the sure screening property. That is, both $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ can contain all active indices (i.e. $\mathcal{A}$) with the probability tending to one, i.e.

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}_1) \to 1, \quad P(\mathcal{A} \subseteq \widehat{\mathcal{A}}_2) \to 1, \quad \text{as } n \to \infty.$$

Then define $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_1 \cap \widehat{\mathcal{A}}_2$ as a new estimate of the active set $\mathcal{A}$. Therefore, the estimate $\widehat{\mathcal{A}}$ also satisfies the sure screening property:

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \to 1, \quad \text{as } n \to \infty.$$

Intuitively, the probability that one unimportant variable has to be selected twice

into both $\widehat{\mathcal{A}}_1$ and $\widehat{\mathcal{A}}_2$ is very small, so $\widehat{\mathcal{A}}$ can be expected to contain much fewer unimportant variables which may be falsely selected into $\widehat{\mathcal{A}}_1$ or $\widehat{\mathcal{A}}_2$. In the result, this simple resampling approach reduces the false positive rate efficiently.

Fan, Samworth and Wu (2009) constructed a theoretical upper bound on the probability of selecting any unimportant variable into the model based on the following exchangeability condition.

(G1) The model satisfies the exchangeability condition at level $r \in \mathbb{N}_+$ if the set of random vectors

$$\{(Y, X_{\mathcal{A}}, X_{j_1}, \ldots, X_{j_r}) : j_1, \ldots, j_r \text{ are distinct elements of } \mathcal{A}^c\}$$

is exchangeable.

This condition guarantees that each unimportant variable is equally likely to be selected by the independence screening procedure.

Theorem 2.2.7. (Upper Bound of False Positive) *Under the exchangeability condition (G1),*

$$\mathbf{P}(|\widehat{\mathcal{A}} \cap \mathcal{A}^c| \geq r) \leq \frac{\binom{d}{r}^2}{\binom{p-|\mathcal{A}|}{r}} \leq \frac{1}{r!}\left(\frac{d^2}{p-|\mathcal{A}|}\right)^r, \tag{2.58}$$

where d is the prespecified size of the selected set $\widehat{\mathcal{A}}_1$ or $\widehat{\mathcal{A}}_2$, and $d^2 \leq p - |\mathcal{A}|$ is required for the second inequality.

Theorem 2.2.7 shows that the probability of selecting at least r unimportant variables can be very small when p is large, d is small and $|\mathcal{A}|$, the number of

important variables, is small. This result seems a little bit unusual. However, we realize that the probability of missing important variables is expected to increase with p and decrease with d.

2.3 Distance Correlation

2.3.1 Definition of Distance Correlation

Szekely, Rizzo and Bakirov (2007) advocated using the distance correlation for measuring dependence between two random vectors. To be precise, let $\phi_{\mathbf{u}}(\mathbf{t})$ and $\phi_{\mathbf{v}}(\mathbf{s})$ be the respective characteristic functions of the random vectors $\mathbf{u}$ and $\mathbf{v}$, and $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})$ be the joint characteristic function of $\mathbf{u}$ and $\mathbf{v}$.

Definition 2.3.1. (Distance Covariance) *Szekely, Rizzo and Bakirov (2007) defined the* ***distance covariance*** *between* $\mathbf{u}$ *and* $\mathbf{v}$ *with finite first moments to be the nonnegative number* $\mathrm{dcov}(\mathbf{u},\mathbf{v})$ *given by*

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = \int_{R^{d_u+d_v}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\|^2 \, w(\mathbf{t},\mathbf{s})\, d\mathbf{t}\, d\mathbf{s}, \tag{2.59}$$

where d_u *and* d_v *are the dimensions of* $\mathbf{u}$ *and* $\mathbf{v}$*, respectively, and*

$$w(\mathbf{t},\mathbf{s}) = \left\{ c_{d_u} c_{d_v} \|\mathbf{t}\|_{d_u}^{1+d_u} \|\mathbf{s}\|_{d_v}^{1+d_v} \right\}^{-1}$$

with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$.

We let $\|\mathbf{a}\|_d$ stand for the Euclidean norm of $\mathbf{a} \in \mathbb{R}^d$, and $\|\phi\|^2 = \phi\bar{\phi}$ for a complex-valued function ϕ with $\bar{\phi}$ being the conjugate of ϕ. The integral at 0 and ∞ in (2.59) is meant in the sense: $\lim_{\eta\to 0} \int_{R^{d_u+d_v}\setminus\{\eta B + \eta^{-1}B^c\}}$, where B is the unit ball in $R^{d_u+d_v}$ with center at 0 and B^c stands for the complement of B.

Definition 2.3.2. (DISTANCE CORRELATION) *The distance correlation (DC) between* $\mathbf{u}$ *and* $\mathbf{v}$ *with finite first moments is defined as*

$$\mathrm{dcorr}(\mathbf{u},\mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u},\mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u},\mathbf{u})\mathrm{dcov}(\mathbf{v},\mathbf{v})}}, \tag{2.60}$$

if $\mathrm{dcov}(\mathbf{u},\mathbf{u})\mathrm{dcov}(\mathbf{v},\mathbf{v}) > 0$. *Otherwise,* $\mathrm{dcorr}(\mathbf{u},\mathbf{v}) = 0$.

2.3.2 Estimate of Distance Correlation

Szekely, Rizzo and Bakirov (2007, Remark 3) stated that

$$\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) = S_1 + S_2 - 2S_3,$$

where S_j, $j = 1$, 2 and 3, are defined below:

$$\begin{aligned} S_1 &= E\left\{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}\right\}, \\ S_2 &= E\left\{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}\right\} E\left\{\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}\right\}, \\ S_3 &= E\left\{E\left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{d_u}|\ \mathbf{u}\right) E\left(\|\mathbf{v}-\widetilde{\mathbf{v}}\|_{d_v}|\ \mathbf{v}\right)\right\}. \end{aligned} \tag{2.61}$$

where $(\widetilde{\mathbf{u}},\widetilde{\mathbf{v}})$ is an independent copy of $(\mathbf{u},\mathbf{v})$.

Suppose that $\{(\mathbf{u}_i,\mathbf{v}_i), i = 1,\cdots,n\}$ is a random sample from the population $(\mathbf{u},\mathbf{v})$. Szekely, Rizzo and Bakirov (2007) proposed to estimate S_1, S_2 and S_3 through the usual moment estimation. To be precise,

$$\begin{aligned} \widehat{S}_1 &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_{d_u}\|\mathbf{v}_i-\mathbf{v}_j\|_{d_v}, \\ \widehat{S}_2 &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_{d_u} n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{v}_i-\mathbf{v}_j\|_{d_v}, \text{ and} \end{aligned}$$

$$\widehat{S}_3 \;=\; n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|\mathbf{u}_i-\mathbf{u}_l\|_{d_u}\|\mathbf{v}_j-\mathbf{v}_l\|_{d_v}.$$

Thus, a natural estimator of $\mathrm{dcov}^2(\mathbf{u},\mathbf{v})$ is given by

$$\widehat{\mathrm{dcov}}^2(\mathbf{u},\mathbf{v}) \;=\; \widehat{S}_1+\widehat{S}_2-2\widehat{S}_3.$$

Similarly, we can define the sample distance covariances $\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})$ and $\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})$. Accordingly, the sample distance correlation between $\mathbf{u}$ and $\mathbf{v}$ can be defined by

$$\widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v})=\frac{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{v})}{\sqrt{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{u})\widehat{\mathrm{dcov}}(\mathbf{v},\mathbf{v})}}.$$

2.3.3 Properties of Distance Correlation

Szekely, Rizzo and Bakirov (2007) systematically studied the theoretic properties of the distance correlation. Note that the definition of the dcorr in (2.60) suggests an analogy with the corresponding Pearson's product-moment correlation. Analogous properties of the dcorr are established in the following theorem.

Theorem 2.3.3. (Properties of dcorr)

(a) If $E(|\mathbf{u}|_{d_u}+|\mathbf{v}|_{d_v})<\infty$, *then* $0\le \mathrm{dcorr}(\mathbf{u},\mathbf{v})\le 1$, *and* $\mathrm{dcorr}(\mathbf{u},\mathbf{v})=0$ *if and only if* $\mathbf{u}$ *and* $\mathbf{v}$ *are independent.*

(b) $0\le \widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v})\le 1$.

(c) If $\widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v})=1$, *then there exist a vector* $\mathbf{a}$, *a nonzero real number* $\mathbf{b}$ *and an orthogonal matrix* $\mathbf{C}$ *such that* $\mathbf{Y}=\mathbf{a}+\mathbf{bXC}$.

The property (a) of the DC in Theorem 2.3.3 motivates us to utilize it in a feature screening procedure, which will be detailed in the next chapter. Note

that two univariate random variables U and V are independent if and only if U and $T(V)$, a strictly monotone transformation of V, are independent. This implies that a DC-based feature screening procedure can be more effective than the marginal Pearson correlation learning in the presence of nonlinear relationship between U and V. Furthermore, Chapter 3 will demonstrate that a DC-based screening procedure is a model-free procedure, in which one does not need to specify a model structure between the predictors and the response.

Szekely, Rizzo and Bakirov (2007) also presented the relationship between the distance correlation and the Pearson correlation coefficient between two univariate normal random variables in the following theorem.

Theorem 2.3.4. (Results for the Bivariate Normal Distribution) *If two univariate random variables U and V follow standard normal distributions, let ρ be the classic Pearson correlation between U and V, then*

(a) $\mathrm{dcorr}(U,V) \leq |\rho|$,

(b) $\mathrm{dcorr}^2(U,V) = \frac{\rho \arcsin(\rho) + \sqrt{1-\rho^2} - \rho \arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1+\pi/3-\sqrt{3}}$,

(c) $\inf_{\rho \neq 0} \frac{\mathrm{dcorr}(U,V)}{|\rho|} = \lim_{\rho \to 0} \frac{\mathrm{dcorr}(U,V)}{|\rho|} = \frac{1}{2(1+\pi/3-\sqrt{3})^{1/2}} \approx 0.89066.$

The plot of $\mathrm{dcorr}^2(U,V)$ versus ρ^2 in the following Figure 2.6 shows the relationship between $\mathrm{dcorr}(U,V)$ and ρ derived in Theorem 2.3.4.

The property (b) of the DC in Theorem 2.3.4 shows that the distance correlation $\mathrm{dcorr}(U,V)$ is strictly increasing in $|\rho|$. This property implies that the DC-based feature screening procedure is equivalent to the marginal Pearson correlation learning for linear regression with normally distributed predictors and random error. In such a situation, Fan and Lv (2008) showed that the Pearson correlation learning has the sure screening property.

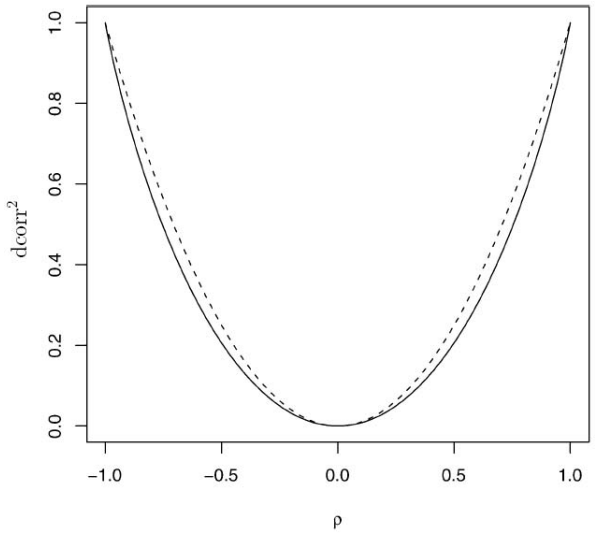

Figure 2.6. Square of Distance Correlation dcorr^2 (solid line) and Square of Pearson Correlation ρ^2 (dashed line) between two univariate normal variables.(Szekely, Rizzo and Bakirov, 2007).

Chapter 3

Feature Screening via Distance Correlation Learning

3.1 Introduction

In this chapter, we propose a new feature screening procedure for ultrahigh dimensional data based on distance correlation (DC-SIS, for short) and an iterative DC-SIS procedure (DC-ISIS). We systematically study the theoretic properties of the DC-SIS, and prove that the DC-SIS possesses the sure screening property in the terminology of Fan and Lv (2008). Monte Carlo simulation studies and real data analysis are conducted to examine the finite sample performance of both DC-SIS and DC-ISIS, and demonstrate their outstanding finite sample performance.

The rest of the chapter is organized as follows. In Section 3.2, we develop the DC-SIS for the ultrahigh dimensional data. Then, the sure screening property is established for the DC-SIS in Section 3.3. In Section 3.4, we examine the finite sample performance of the DC-SIS via Monte Carlo simulations as well as a real data example. In Section 3.5, we propose the DC-ISIS to further enhance the finite sample performance of the DC-SIS. All technical proofs are given in Section 3.6.

3.2 A New Independence Screening Procedure

In this section we propose an independence screening procedure built upon the DC. Let $\mathbf{y} = (Y_1, \cdots, Y_q)^{\mathrm{T}}$ be the response vector with support Ψ_y, and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be the predictor vector. We regard q as a fixed number in this context. In an ultrahigh-dimensional setting the dimensionality p greatly exceeds the sample size n. It is thus natural to assume that only a small number of predictors are relevant to $\mathbf{y}$. Denote by $F(\mathbf{y} \mid \mathbf{x})$ the conditional distribution function of $\mathbf{y}$ given $\mathbf{x}$. Without specifying a regression model, we define the index set of the active and inactive predictors by

$$\begin{aligned} \mathcal{D} &= \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } \mathbf{y} \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } \mathbf{y} \in \Psi_y\}. \end{aligned} \tag{3.1}$$

We further write $\mathbf{x}_{\mathcal{D}} = \{X_k : k \in \mathcal{D}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$, and refer to $\mathbf{x}_{\mathcal{D}}$ as an *active* predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ as an *inactive* predictor vector. The index subset $\mathcal{D}$ of all active predictors or, equivalently, the index subset $\mathcal{I}$ of all inactive predictors, is the objective of our primary interest. Definition (3.1) implies that $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{\mathcal{I}} \mid \mathbf{x}_{\mathcal{D}}$, where $\perp\!\!\!\perp$ denotes statistical independence. That is, given $\mathbf{x}_{\mathcal{D}}$, the remaining predictors $\mathbf{x}_{\mathcal{I}}$ are independent of $\mathbf{y}$. Thus the inactive predictors $\mathbf{x}_{\mathcal{I}}$ are redundant when the active predictors $\mathbf{x}_{\mathcal{D}}$ are known.

For ease of presentation, we write

$$\omega_k = \mathrm{dcorr}^2(X_k, \mathbf{y}), \quad \text{and} \quad \widehat{\omega}_k = \widehat{\mathrm{dcorr}}^2(X_k, \mathbf{y}), \text{ for } k = 1, \cdots, p.$$

based on a random sample $\{\mathbf{x}_i, \mathbf{y}_i\}$, $i = 1, \ldots, n$. We consider using ω_k as a marginal utility to rank the importance of X_k at the population level. We utilize the DC because it allows for arbitrary regression relationship of $\mathbf{y}$ onto $\mathbf{x}$,

regardless of whether it is linear or nonlinear. The DC also permits univariate and multivariate response, regardless of whether it is continuous, discrete or categorical. In addition, it allows for groupwise predictors. Thus, this DC based screening procedure is completely model-free. We select a set of important predictors with large $\hat{\omega}_k$. That is, we define

$$\widehat{\mathcal{D}}^{\star} = \left\{k : \widehat{\omega}_k \geq cn^{-\kappa}, \text{ for } 1 \leq k \leq p\right\},$$

where c and κ are pre-specified threshold values which will be defined in condition (C3.2) in the subsequent section.

3.3 Theoretical Properties

3.3.1 Preliminary Lemmas

The following three lemmas will be used in the proof of Theorems 3.3.4 sure screening property. The Lemma 3.3.1 below provides a useful decomposition of the distance correlation (Szekely, Rizzo and Bakirov, 2007) between the kth predictor X_k and the q-dimemsional response $\mathbf{y}$ at the population level.

Lemma 3.3.1. *Let $\phi_{X_k}(t)$ and $\phi_{\mathbf{y}}(\mathbf{s})$ be the respective characteristic functions of X_k and $\mathbf{y}$, and $\phi_{X_k,\mathbf{y}}(t,\mathbf{s})$ be the joint characteristic function of X_k and $\mathbf{y}$. Let $w(t,\mathbf{s}) = \left\{c_1 c_q |t|_1^2 |\mathbf{s}|_q^{q+1}\right\}^{-1}$ with $c_q = \pi^{(q+1)/2}/\Gamma\left\{(q+1)/2\right\}$ and $c_1 = \pi$. Then*

$$\begin{aligned} \mathrm{dcov}^2(X_k, \mathbf{y}) &= \int_{\mathbb{R}^{q+1}} \left|\phi_{X_k,\mathbf{y}}(t,\mathbf{s}) - \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\right|^2 w(t,\mathbf{s}) dt d\mathbf{s} \\ &= S_{k1} + S_{k2} - 2S_{k3}, \end{aligned}$$

where

$$S_{k1} = E\|X_k - \tilde{X}_k\|_1\|\mathbf{y} - \tilde{\mathbf{y}}\|_q,$$
$$S_{k2} = E\|X_k - \tilde{X}_k\|_1 E\|\mathbf{y} - \tilde{\mathbf{y}}\|_q, \quad \textit{and}$$
$$S_{k3} = E\{E(\|X_k - \tilde{X}_k\|_1|X_k)E(\|\mathbf{y} - \tilde{\mathbf{y}}\|_q|\mathbf{y})\},$$

and $\{\tilde{X}_k, \tilde{\mathbf{y}}\}$ *is an independent copy of* $\{X_k, \mathbf{y}\}$*, respectively.*

Szekely, Rizzo and Bakirov (2007) mentioned the result in Lemma 3.3.1 in their Remark 3. However, they did not provide a rigorous proof. A detailed technical proof is provided in Section 3.6.

Note that the sample counterparts of S_{ki}'s can be estimated by the method of moment as follows,

$$\widehat{S}_{k1} = \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_{ik} - X_{jk}\|_1\|\mathbf{y}_i - \mathbf{y}_j\|_q,$$
$$\widehat{S}_{k2} = \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_{ik} - X_{jk}\|_1\frac{1}{n^2}\sum_{i,j=1}^{n}\|\mathbf{y}_i - \mathbf{y}_j\|_q, \quad \text{and}$$
$$\widehat{S}_{k3} = \frac{1}{n^3}\sum_{i,j,l=1}^{n}\|X_{ik} - X_{lk}\|_1\|\mathbf{y}_j - \mathbf{y}_l\|_q.$$

Then, by definitions of distance covariance and sample distance covariance, we have that

$$\widehat{\mathrm{dcov}}^2(X_k, \mathbf{y}) = \widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}.$$

These following two lemmas provide us two exponential inequalities, and are extracted from Lemma 5.6.1.A and Theorem 5.6.1.A of Serfling (1980).

Lemma 3.3.2. *Let* $\mu = E(Y)$*. If* $\Pr(a \le Y \le b) = 1$*, then*

$$E\left[\exp\left\{s(Y-\mu)\right\}\right] \le \exp\left\{s^2(b-a)^2/8\right\}, \textit{ for any } s > 0.$$

Lemma 3.3.3. *Let $h(Y_1, \cdots, Y_m)$ be a kernel of the U-statistic U_n, and the parameter $\theta = E\{h(Y_1, \cdots, Y_m)\}$. If $a \leq h(Y_1, \cdots, Y_m) \leq b$, then, for any $t > 0$ and $n \geq m$,*

$$Pr(U_n - \theta \geq t) \leq \exp\left\{-2[n/m]t^2/(b-a)^2\right\},$$

where $[n/m]$ denotes the integer part of n/m.

Due to the symmetry of U-statistic, Lemma 3.3.3 entails that

$$\Pr(|U_n - \theta| \geq t) \leq 2\exp\left\{-2[n/m]t^2/(b-a)^2\right\}.$$

3.3.2 Sure Screening Property

Next we study the theoretical properties of the proposed independence screening procedure built upon the DC. The following conditions are imposed to facilitate the technical proofs, although they may not be the weakest ones.

(C3.1) Both $\mathbf{x}$ and $\mathbf{y}$ satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0 < s \leq 2s_0$,

$$\sup_p \max_{1\leq k\leq p} E\left\{\exp(s\|X_k\|_1^2)\right\} < \infty, \text{ and } E\{\exp(s\|\mathbf{y}\|_q^2)\} < \infty.$$

(C3.2) The minimum distance correlation of active predictors satisfies

$$\min_{k\in\mathcal{D}} \omega_k \geq 2cn^{-\kappa}, \text{ for some constants } c > 0 \text{ and } 0 \leq \kappa < 1/2.$$

Condition (C3.1) follows immediately when $\mathbf{x}$ and $\mathbf{y}$ are bounded uniformly, or when they have multivariate normal distribution. The normality assumption has

been widely used in the area of ultrahigh dimensional data analysis to facilitate the technical derivations. See, for example, Fan and Lv (2008) and Wang (2009).

Next we explore condition (C3.2). When $\mathbf{x}$ and $\mathbf{y}$ have multivariate normal distribution, Theorem 2.3.4 (b) gives an explicit relationship between the DC and the squared Pearson correlation. For simplicity, we write $\text{dcorr}(X_k, \mathbf{y}) = T_0\left(|\rho(X_k, \mathbf{y})|\right)$ where $T_0(\cdot)$ is strictly increasing given in Theorem 2.3.4 (b). In this situation, condition (C3.2) requires essentially that $\min_{k\in\mathcal{D}}|\rho(X_k, \mathbf{y})| \geq T_{inv}(2cn^{-\kappa})$, where $T_{inv}(\cdot)$ is the inverse function of $T_0(\cdot)$. This is parallel to condition 3 of Fan and Lv (2008) where it is assumed that $\min_{k\in\mathcal{D}}|\rho(X_k, \mathbf{y})| \geq 2cn^{-\kappa}$. This intuitive illustration implies that condition (C3.2) requires that the marginal DC of active predictors cannot be too small, which is similar to condition 3 of Fan and Lv (2008). We remark here that, although we illustrate the intuition by assuming that $\mathbf{x}$ and $\mathbf{y}$ are multivariate normal, we do not require this assumption explicitly in our context.

The following Theorem 3.3.4 establishes the sure screening property for the DC-SIS procedure, which is a desired property for ultrahigh dimensional statistical learning. The technical proof of Theorem 3.3.4 is provided in Section 3.6.

Theorem 3.3.4. (Sure Screening Property) *Under condition (C3.1), for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that*

$$Pr\left(\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right) \leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right). \quad (3.2)$$

Under conditions (C3.1) and (C3.2), we have that

$$Pr\left(\mathcal{D} \subseteq \widehat{\mathcal{D}}^{\star}\right) \geq 1 - O\left(s_n\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + n\exp\left(-c_2 n^{\gamma}\right)\right]\right), \quad (3.3)$$

where s_n is the cardinality of $\mathcal{D}$.

The sure screening property holds for the DC-SIS under milder conditions than those for the SIS (Fan and Lv, 2008) in that we do not require the regression function of $\mathbf{y}$ onto $\mathbf{x}$ to be linear. Thus, the DC-SIS provides a unified alternative to existing model-based sure screening procedures. Compared with the SIRS (Zhu, Li, Li and Zhu, 2011), the DC-SIS can effectively handle grouped predictors and multivariate responses.

To balance the two terms in the right hand side of (3.2), we choose the optimal order $\gamma = (1-2\kappa)/3$, then the first part of Theorem 3.3.4 becomes

$$\Pr\Big(\max_{1\le k\le p}|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\Big)\le O\left(p\left[\exp\left\{-c_1n^{(1-2\kappa)/3}\right\}\right]\right),$$

for some constant $c_1 > 0$, indicating that we can handle the NP-dimensionality of order $\log p = o\left(n^{(1-2\kappa)/3}\right)$. If we further assume that X_k and $\mathbf{y}$ are bounded uniformly in p, then we can obtain without much difficulty that

$$\Pr\left(\max_{1\le k\le p}|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\right)\le O\left(p\left[\exp\left\{-c_1n^{1-2\kappa}\right\}\right]\right).$$

In this case, we can handle the NP-dimensionality $\log p = o\left(n^{1-2\kappa}\right)$.

3.4 Numerical Studies

In this section we assess the performance of the DC-SIS by Monte Carlo simulation. Our simulation studies were conducted using R code. We further illustrate the proposed screening procedure with an empirical analysis of a real data example.

In Examples 1, 2 and 3, we generate $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, and the error term ε from standard normal distribution $\mathcal{N}(0,1)$. We consider two covariance matrices

to assess the performance of the DC-SIS and to compare with existing methods: (i) $\sigma_{ij} = 0.8^{|i-j|}$ and (ii) $\sigma_{ij} = 0.5^{|i-j|}$. We fix the sample size n to be 200 and vary the dimension p from 2,000 to 5,000. We repeat each experiment 500 times, and evaluate the performance through the following three criteria.

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 500 replications.

2. $\mathcal{P}_s$: the proportion that an individual active predictor is selected for a given model size d in the 500 replications.

3. $\mathcal{P}_a$: the proportion that all active predictors are selected for a given model size d in the 500 replications.

The $\mathcal{S}$ is expected to be close to the number of truly active predictors. The sure screening property ensures that $\mathcal{P}_s$ and $\mathcal{P}_a$ are both close to one when the estimated model size d is sufficiently large. We choose d to be $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$ throughout our simulations to empirically examine the effect of the cutoff, where $[a]$ denotes the integer part of a.

Example 1. This example is designed to compare the finite sample performance of the DC-SIS with the SIS (Fan and Lv, 2008) and SIRS (Zhu, Li, Li and Zhu, 2011). In this example, we generate the response from the following four models:

$$\begin{aligned}
\textbf{(1.a):} \qquad Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + c_4\beta_4 X_{22} + \varepsilon, \\
\textbf{(1.b):} \qquad Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) + c_4\beta_3 X_{22} + \varepsilon, \\
\textbf{(1.c):} \qquad Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) X_{22} + \varepsilon, \\
\textbf{(1.d):} \qquad Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + \exp(c_4|X_{22}|)\varepsilon,
\end{aligned}$$

where $\mathbf{1}(X_{12} < 0)$ is an indicator function. The regression functions $E(Y \mid \mathbf{x})$ in

Table 3.1. The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 1.

$\mathcal{S}$	SIS					SIRS					DC-SIS				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
case 1: $p = 2000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$															
(1.a)	4.0	4.0	5.0	7.0	21.2	4.0	4.0	5.0	7.0	45.1	4.0	4.0	4.0	6.0	18.0
(1.b)	68.0	578.5	1180.5	1634.5	1938.0	232.9	871.5	1386.0	1725.2	1942.4	5.0	9.0	24.5	73.0	345.1
(1.c)	395.9	1037.2	1438.0	1745.0	1945.1	238.5	805.0	1320.0	1697.0	1946.0	6.0	10.0	22.0	59.0	324.1
(1.d)	130.5	611.2	1166.0	1637.0	1936.5	42.0	304.2	797.0	1432.2	1846.1	4.0	5.0	9.0	41.0	336.2
case 2: $p = 2000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$															
(1.a)	5.0	9.0	16.0	97.0	729.4	5.0	9.0	18.0	112.8	957.1	4.0	7.0	11.0	31.2	507.2
(1.b)	26.0	283.2	852.0	1541.2	1919.0	103.9	603.0	1174.0	1699.2	1968.0	5.0	8.0	11.0	17.0	98.0
(1.c)	224.5	775.2	1249.5	1670.0	1951.1	118.6	573.2	1201.5	1685.2	1955.0	7.0	10.0	15.0	38.0	198.3
(1.d)	79.0	583.8	1107.5	1626.2	1930.0	50.9	300.5	728.0	1368.2	1900.1	4.0	7.0	17.0	73.2	653.1
case 3: $p = 5000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$															
(1.a)	4.0	4.0	5.0	6.0	59.0	4.0	4.0	5.0	7.0	88.4	4.0	4.0	4.0	6.0	34.1
(1.b)	165.1	1112.5	2729.0	3997.2	4851.5	560.8	1913.0	3249.0	4329.0	4869.1	5.0	11.8	45.0	168.8	956.7
(1.c)	1183.7	2712.0	3604.5	4380.2	4885.0	440.4	1949.0	3205.5	4242.8	4883.1	7.0	17.0	53.0	179.5	732.0
(1.d)	259.9	1338.5	2808.5	3990.8	4764.9	118.7	823.2	1833.5	3314.5	4706.1	4.0	5.0	15.0	77.2	848.2
case 4: $p = 5000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$															
(1.a)	5.0	10.0	26.5	251.5	2522.7	5.0	10.0	28.0	324.8	3246.4	5.0	8.0	14.0	69.0	1455.1
(1.b)	40.7	639.8	2072.0	3803.8	4801.7	215.7	1677.8	3010.0	4352.2	4934.1	5.0	8.0	11.0	21.0	162.0
(1.c)	479.2	1884.8	3347.5	4298.5	4875.2	297.7	1359.2	2738.5	4072.5	4877.6	8.0	12.0	22.0	83.0	657.9
(1.d)	307.0	1544.0	2832.5	4026.2	4785.2	148.2	672.0	1874.0	3330.0	4665.2	4.0	7.0	21.0	165.2	1330.0

models **(1.a)**-**(1.d)** are all nonlinear in X_{12}. In addition, models (**1.b**) and (**1.c**) contain an interaction term X_1X_2, and model (**1.d**) is heteroscedastic. Following Fan and Lv (2008), we choose $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, 2, 3$ and 4, where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim \mathcal{N}(0, 1)$. We set $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$ in this example to challenge the feature screening procedures under consideration. For each independence screening procedure, we compute the associated marginal utility between each predictor X_k and the response Y, that is, we regard $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ as the predictor vector in this example.

Tables 4.1 and 3.2 depict the simulation results for $\mathcal{S}$, $\mathcal{P}_s$ and $\mathcal{P}_a$. The performances of the DC-SIS, SIS and SIRS are quite similar in model **(1.a)**, indicating that the the SIS has a robust performance if the working linear model does not deviate far from the underlying true model. The DC-SIS outperforms the SIS and SIRS significantly in models **(1.b)**, **(1.c)** and **(1.d)**. Both the SIS and SIRS have little chance to identify the important predictors X_1 and X_2 in models **(1.b)** and **(1.c)**, and X_{22} in model **(1.d)**. The main reason is that both the SIS and SIRS

fail to identify important predictors which are symmetrically relevant to the response variable. To be precise, the marginal utilities of both the SIS and SIRS will be exactly zero if X_k satisfies $E(X_k \mid Y) = E(X_k)$. In models **(1.b)** and **(1.c)**, because X_1 and X_2 are highly correlated, both exhibit symmetric patterns with Y (plots are not shown here). In model **(1.d)** the symmetry of X_{22} is obvious in that $E(X_{22} \mid Y) = E(X_{22})$. In contrast, the DC-SIS does not suffer from the symmetry issue. It performs quite well throughout all scenarios. This demonstrates a specific advantage of the distance correlation over the Pearson correlation upon which the SIS and SIRS are built.

This can be interpreted to mean that the regularity conditions for the SIS or the SIRS may not hold for the current model settings. It is challenging to impose a model structure with little prior information between the response and the predictors under the ultrahigh dimensional setting. Thus, the DC-SIS procedure may be more desirable than the SIS, because the screening property of the SIS was established based on the linear regression model in Fan and Lv (2008). Although Zhu, Li, Li and Zhu (2011) claims that the SIRS is model-free, it fails to select some predictors either, such as X_{22} in the model **(1.d)**. The potential reason is that the SIRS cannot handle the interaction information in models **(1.b), (1.c)** or the signal contained in the conditional variance in the model **(1.d)**.

Example 2. We illustrate that the DC-SIS can be directly used for screening grouped predictors. In many regression problems, some predictors can be naturally grouped. The most common example which contains group variables is the multi-factor ANOVA problem, in which each factor may have several levels and can be expressed through a group of dummy variables. The goal of ANOVA is to select important main effects and interactions for accurate predictions, which amounts to the selection of groups of dummy variables. To demonstrate the practicability

Table 3.2. The empirical probabilities of each active predictor (denoted by $\mathcal{P}_s$) and all active predictors (denoted by $\mathcal{P}_a$) are chosen for a given model size d_i, where $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

		SIS					SIRS					DC-SIS				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
model	size	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
case 1: $p = 2000$ and $\sigma_{ij} = 0.5^{\lvert i-j\rvert}$																
	d_1	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.95	1.00	0.94	1.00	1.00	0.97	1.00	0.96
(1.a)	d_2	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.98	1.00	0.98
	d_3	1.00	1.00	0.98	1.00	0.98	1.00	1.00	0.97	1.00	0.97	1.00	1.00	0.99	1.00	0.98
	d_1	0.08	0.07	0.97	1.00	0.03	0.02	0.03	0.98	1.00	0.00	0.72	0.70	0.99	1.00	0.58
(1.b)	d_2	0.12	0.13	0.98	1.00	0.06	0.05	0.05	0.99	1.00	0.01	0.85	0.84	1.00	1.00	0.76
	d_3	0.15	0.17	0.99	1.00	0.07	0.06	0.06	0.99	1.00	0.01	0.89	0.88	1.00	1.00	0.82
	d_1	0.12	0.13	0.01	0.99	0.00	0.04	0.03	0.51	1.00	0.01	0.93	0.93	0.77	1.00	0.65
(1.c)	d_2	0.17	0.18	0.03	0.99	0.00	0.07	0.05	0.67	1.00	0.01	0.97	0.96	0.84	1.00	0.79
	d_3	0.21	0.21	0.05	0.99	0.00	0.09	0.08	0.75	1.00	0.02	0.98	0.97	0.89	1.00	0.84
	d_1	0.42	0.22	0.14	0.42	0.02	1.00	0.98	0.87	0.05	0.04	1.00	0.91	0.81	0.99	0.73
(1.d)	d_2	0.48	0.29	0.22	0.50	0.03	1.00	0.99	0.91	0.10	0.09	1.00	0.94	0.87	1.00	0.82
	d_3	0.56	0.32	0.26	0.54	0.04	1.00	0.99	0.93	0.12	0.11	1.00	0.96	0.92	1.00	0.88
case 2: $p = 2000$ and $\sigma_{ij} = 0.8^{\lvert i-j\rvert}$																
	d_1	1.00	1.00	0.63	1.00	0.63	1.00	1.00	0.62	1.00	0.62	1.00	1.00	0.78	1.00	0.77
(1.a)	d_2	1.00	1.00	0.71	1.00	0.72	1.00	1.00	0.70	1.00	0.69	1.00	1.00	0.84	1.00	0.84
	d_3	1.00	1.00	0.77	1.00	0.78	1.00	1.00	0.75	1.00	0.75	1.00	1.00	0.86	1.00	0.86
	d_1	0.12	0.13	0.81	1.00	0.06	0.04	0.04	0.88	1.00	0.02	0.97	0.98	0.92	1.00	0.88
(1.b)	d_2	0.19	0.19	0.86	1.00	0.12	0.07	0.07	0.91	1.00	0.03	0.99	0.99	0.95	1.00	0.94
	d_3	0.22	0.23	0.88	1.00	0.15	0.09	0.11	0.93	1.00	0.06	1.00	0.99	0.96	1.00	0.96
	d_1	0.17	0.16	0.03	0.99	0.00	0.04	0.04	0.53	1.00	0.02	1.00	1.00	0.75	1.00	0.75
(1.c)	d_2	0.22	0.22	0.06	1.00	0.01	0.08	0.08	0.71	1.00	0.03	1.00	1.00	0.85	1.00	0.86
	d_3	0.27	0.27	0.10	1.00	0.03	0.10	0.10	0.81	1.00	0.05	1.00	1.00	0.90	1.00	0.90
	d_1	0.44	0.38	0.11	0.45	0.03	1.00	1.00	0.73	0.05	0.04	0.99	0.98	0.68	1.00	0.67
(1.d)	d_2	0.51	0.46	0.18	0.53	0.05	1.00	1.00	0.81	0.09	0.08	1.00	0.98	0.76	1.00	0.75
	d_3	0.55	0.49	0.22	0.57	0.06	1.00	1.00	0.84	0.14	0.11	1.00	0.99	0.80	1.00	0.80
case 3: $p = 5000$ and $\sigma_{ij} = 0.5^{\lvert i-j\rvert}$																
	d_1	1.00	1.00	0.94	1.00	0.94	1.00	0.99	0.92	1.00	0.92	1.00	0.99	0.96	1.00	0.95
(1.a)	d_2	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.97	1.00	0.97
	d_3	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.96	1.00	0.96	1.00	1.00	0.98	1.00	0.98
	d_1	0.06	0.06	0.94	1.00	0.02	0.02	0.02	0.96	1.00	0.00	0.59	0.60	0.98	1.00	0.46
(1.b)	d_2	0.09	0.09	0.96	1.00	0.03	0.03	0.03	0.97	1.00	0.01	0.72	0.72	0.99	1.00	0.61
	d_3	0.12	0.10	0.97	1.00	0.04	0.05	0.04	0.98	1.00	0.01	0.79	0.78	0.99	1.00	0.68
	d_1	0.06	0.06	0.01	0.99	0.00	0.03	0.02	0.30	1.00	0.00	0.86	0.87	0.61	1.00	0.41
(1.c)	d_2	0.10	0.10	0.02	1.00	0.00	0.04	0.03	0.45	1.00	0.00	0.92	0.93	0.69	1.00	0.57
	d_3	0.12	0.12	0.02	1.00	0.00	0.05	0.05	0.53	1.00	0.00	0.94	0.95	0.73	1.00	0.64
	d_1	0.39	0.21	0.11	0.40	0.01	1.00	0.97	0.82	0.02	0.02	0.99	0.87	0.74	0.99	0.65
(1.d)	d_2	0.44	0.24	0.14	0.45	0.01	1.00	0.98	0.88	0.04	0.03	0.99	0.90	0.81	0.99	0.75
	d_3	0.48	0.28	0.17	0.47	0.02	1.00	0.99	0.90	0.06	0.05	0.99	0.92	0.85	1.00	0.79
case 4: $p = 5000$ and $\sigma_{ij} = 0.8^{\lvert i-j\rvert}$																
	d_1	1.00	1.00	0.55	1.00	0.55	1.00	1.00	0.55	1.00	0.55	1.00	1.00	0.70	1.00	0.69
(1.a)	d_2	1.00	1.00	0.61	1.00	0.62	1.00	1.00	0.61	1.00	0.61	1.00	1.00	0.76	1.00	0.76
	d_3	1.00	1.00	0.67	1.00	0.67	1.00	1.00	0.64	1.00	0.64	1.00	1.00	0.80	1.00	0.80
	d_1	0.10	0.09	0.74	1.00	0.05	0.02	0.02	0.83	1.00	0.00	0.94	0.94	0.90	1.00	0.82
(1.b)	d_2	0.12	0.13	0.81	1.00	0.07	0.03	0.04	0.87	1.00	0.01	0.97	0.97	0.93	1.00	0.89
	d_3	0.15	0.16	0.84	1.00	0.10	0.05	0.06	0.90	1.00	0.02	0.98	0.98	0.95	1.00	0.92
	d_1	0.10	0.10	0.02	0.98	0.00	0.02	0.03	0.34	1.00	0.00	1.00	1.00	0.64	1.00	0.63
(1.c)	d_2	0.13	0.14	0.04	0.99	0.01	0.04	0.04	0.50	1.00	0.01	1.00	1.00	0.74	1.00	0.74
	d_3	0.16	0.18	0.05	0.99	0.01	0.05	0.05	0.61	1.00	0.02	1.00	1.00	0.79	1.00	0.79
	d_1	0.42	0.32	0.09	0.40	0.01	1.00	1.00	0.66	0.02	0.01	0.99	0.97	0.63	0.98	0.59
(1.d)	d_2	0.48	0.39	0.12	0.44	0.02	1.00	1.00	0.74	0.04	0.03	0.99	0.97	0.70	1.00	0.68
	d_3	0.51	0.42	0.15	0.46	0.02	1.00	1.00	0.78	0.05	0.04	0.99	0.98	0.73	1.00	0.71

of the DC-SIS, we adopt the following model:

$$\begin{aligned} Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3\{\mathbf{1}(X_{12} < q_1) + 1.5 \times \mathbf{1}(q_1 \le X_{12} < q_2) \\ &\quad +2 \times \mathbf{1}(q_2 \le X_{12} < q_3)\} + c_4\beta_4 X_{22} + \varepsilon, \end{aligned}$$

where q_1, q_2 and q_3 are the 25%, 50% and 75% quantiles of X_{12}, respectively. The variables X with the coefficients c_i's and β_i's are the same as those in Example 1. We write

$$\widetilde{\mathbf{x}}_{12} = \{\mathbf{1}(X_{12} < q_1),\ \mathbf{1}(q_1 \le X_{12} < q_2),\ \mathbf{1}(q_2 \le X_{12} < q_3))\}^{\mathrm{T}}.$$

These three correlated variables naturally become a group. The predictor vector in this example becomes $\mathbf{x} = (X_1, \ldots, X_{11}, \widetilde{\mathbf{x}}_{12}, X_{13}, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^{p+2}$. We remark here that the marginal utility of the grouped variable $\widetilde{\mathbf{x}}_{12}$ is defined by

$$\widehat{\omega}_{12} = \widehat{\mathrm{dcorr}}^2(\widetilde{\mathbf{x}}_{12}, Y).$$

The 5%, 25%, 50%, 75% and 95% percentiles of the minimum model size $\mathcal{S}$ are summarized in Table 3.3. These percentiles indicate that with very high probability, the minimum model size $\mathcal{S}$ to ensure the inclusion of all active predictors is small. Note that $[n/\log(n)] = 37$. Thus, almost all $\mathcal{P}_s$s and $\mathcal{P}_a$s equal 100%. All active predictors including the grouped variable $\widetilde{\mathbf{x}}_{12}$ can almost perfectly be selected into the resulting model across all three different model sizes. Hence, the DC-SIS is efficient to select the grouped predictors.

Example 3. In this example, we investigate the performance of the DC-SIS with multivariate responses. The SIS proposed in Fan and Lv (2008) cannot be directly applied for such settings. In contrast, the DC-SIS is ready for screening the active

Table 3.3. The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 2.

$\mathcal{S}$	$p = 2000$					$p = 5000$				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	4.0	4.0	4.0	5.0	12.0	4.0	4.0	4.0	6.0	16.1
$\sigma_{ij} = 0.8^{\|i-j\|}$	4.0	5.0	7.0	9.0	15.2	4.0	5.0	7.0	9.0	21.0

Table 3.4. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 2. The user-specified model size $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

		$p = 2000$					$p = 5000$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
	size	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
$\sigma_{ij} = 0.5^{\|i-j\|}$	d_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\sigma_{ij} = 0.8^{\|i-j\|}$	d_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

predictors by the nature of DC. In this example, we generate $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$ from normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}} = (\sigma_{\mathbf{x},ij})_{2\times 2}$, where $\sigma_{\mathbf{x},11} = \sigma_{\mathbf{x},22} = 1$ and $\sigma_{\mathbf{x},12} = \sigma_{\mathbf{x},21} = \sigma(\mathbf{x})$.

We consider two scenarios for the correlation function $\sigma(\mathbf{x})$:

(3.a): $\sigma(\mathbf{x}) = \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$, where $\boldsymbol{\beta}_1 = (0.8, 0.6, 0, \ldots, 0)^{\mathrm{T}}$.

(3.b): $\sigma(\mathbf{x}) = \{\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) - 1\} / \{\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) + 1\}$, where $\boldsymbol{\beta}_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, \ldots, 0)^{\mathrm{T}}$ with U_i's being independent and identically distributed according to uniform distribution Uniform$[0, 1]$.

The simulation results are reported in Tables 3.5 and 3.6. Once again, we can see from Table 3.5 that the minimum model size $\mathcal{S}$ to include all active predictors is much smaller than the sample size and close to the number of the truly active predictors. Table 3.6 indicates that the proportions that the active predictors are selected into the model are close to one, which supports the assertion that the DC-SIS processes the sure screening property. It implies that the DC-SIS

can identify the active predictors contained in correlations between multivariate responses. This may be potentially useful in gene co-expression analysis.

Table 3.5. The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 3.

	$\mathcal{S}$	$p = 2000$					$p = 5000$				
	Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	**(3.a)**	4.0	9.0	18.0	39.3	112.3	6.0	22.0	48.0	95.3	296.4
	(3.b)	6.0	19.0	43.0	92.0	253.1	14.0	45.0	92.5	198.8	571.6
$\sigma_{ij} = 0.8^{\|i-j\|}$	**(3.a)**	2.0	3.0	6.0	12.0	40.0	2.0	6.0	14.0	32.0	98.0
	(3.b)	4.0	4.0	4.0	6.0	10.0	4.0	4.0	5.0	8.0	18.1

Table 3.6. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 3. The user-specified model size $d_1 = [n/\log n]$, $d_2 = 2[n/\log n]$ and $d_3 = 3[n/\log n]$.

		$p = 2000$								$p = 5000$							
		(3.a)			**(3.b)**					**(3.a)**			**(3.b)**				
		$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
	size	X_1	X_2	ALL	X_1	X_2	X_3	X_4	ALL	X_1	X_2	ALL	X_1	X_2	X_3	X_4	ALL
$\sigma_{ij} = 0.5^{\|i-j\|}$	d_1	0.95	0.76	0.74	0.71	0.98	0.98	0.72	0.47	0.79	0.49	0.42	0.48	0.91	0.90	0.53	0.20
	d_2	0.98	0.90	0.90	0.85	0.99	0.99	0.85	0.71	0.93	0.70	0.67	0.67	0.97	0.97	0.71	0.45
	d_3	1.00	0.95	0.95	0.91	0.99	1.00	0.90	0.81	0.97	0.81	0.80	0.75	0.98	0.99	0.78	0.55
$\sigma_{ij} = 0.8^{\|i-j\|}$	d_1	0.98	0.95	0.94	1.00	1.00	1.00	1.00	1.00	0.92	0.84	0.81	1.00	1.00	1.00	0.99	0.99
	d_2	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.93	1.00	1.00	1.00	1.00	1.00
	d_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.96	1.00	1.00	1.00	1.00	1.00

Example 4. The Cardiomyopathy microarray dataset was once analyzed by Segal, Dahlquist and Conklin (2003) and Hall and Miller (2009). The goal is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice. The response Y is the Ro1 expression level, and the predictors X_k's are other gene expression levels. Compared with the sample size $n = 30$ in this dataset, the dimension $p = 6319$ is very large.

The DC-SIS procedure ranks two genes, labeled Msa.2134.0 and Msa.2877.0, at the top. The scatter plots of Y versus these two gene expression levels with cubic spline fit curves in Figure 3.1 indicate clearly the existence of nonlinear patterns. Yet, our finding is different from Hall and Miller (2009) in that they

ranked Msa.2877.0 and Msa.1166.0 at the top with their proposed generalized correlation ranking. A natural question arises: which screening procedure performs better in terms of ranking? To compare the performance of these two procedures, we fit an additive model as follows:

$$Y = \ell_{k1}(X_{k1}) + \ell_{k2}(X_{k2}) + \varepsilon_k, \text{ for } k = 1, 2.$$

The DC-SIS, corresponding to $k = 1$, regards Msa.2134.0 and Msa.2877.0 as the two predictors, while the generalized correlation ranking proposed by Hall and Miller (2009), corresponding to $k = 2$, regards Msa.2877.0 and Msa.1166.0 as predictors in the above model. We fit the unknown link functions ℓ_{ki} using the R `mgcv` package. The DC-SIS method clearly achieves better performance with the adjusted R^2 of 96.8% and the deviance explained of 98.3%, in contrast to the adjusted R^2 of 84.5% and the deviance explained of 86.6% for the generalized correlation ranking method. We remark here that deviance explained means the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance. Because both the adjusted R^2 values and the explained deviance are very large, it seems unnecessary to extract any additional genes.

3.5 The Iterative Screening Procedure

The DC-SIS may fail to identify some active predictors which are marginally independent of the response. In this section, we develop an iterative DC-SIS procedure to fix this issue. We first describe this phenomenon via an illustrative example considered by Fan and Lv (2008).

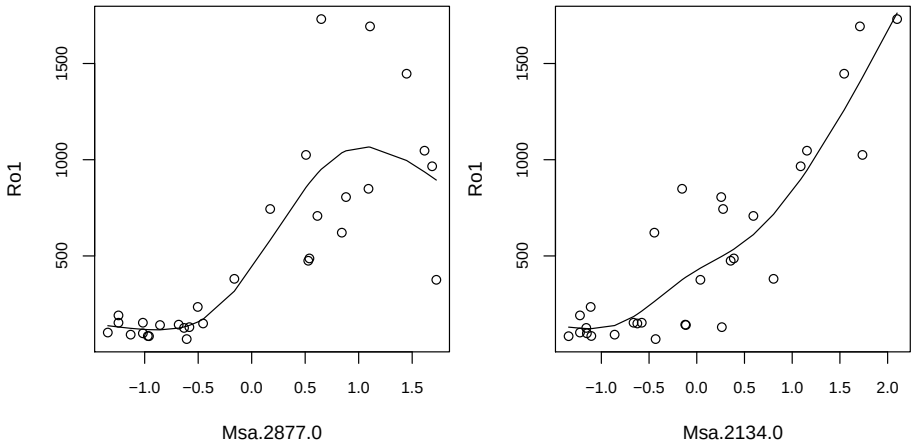

Figure 3.1. The scatter plot of Y versus two genes expression levels identified by the DC-SIS.

We consider the following model

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon. \tag{3.4}$$

Each predictor is generated from a normal distribution with zero mean and unit variance. All X_k's except X_4 are equally correlated with the Pearson correlation coefficient ρ, while X_4 has the Pearson correlation $\sqrt{\rho}$ with all other $p-1$ variables. In this example, X_4 is marginally independent of but jointly relevant to the response variable Y. Both SIS and DC-SIS can only pick out X_4 by chance, although X_4 is clearly a variable of interest.

For linear models, one may calculate the least squares fit with selected variables, and further calculate the residuals. Thus, Fan and Lv (2008) proposed an iterative SIS (ISIS) to detect significant predictors that are marginally independent of the response by regarding the residual as a new response. It is more challenging for the DC-SIS to handle such an issue because we do not want to impose a regression model, and thus residuals are not available. Below we introduce an iterative

sure independent screening via distance correlation (DC-ISIS) procedure to handle the issue. Zhu, Li, Li and Zhu (2011) proposed a similar iterative procedure for their SIRS. Below we apply their strategy for the DC-SIS. The DC-ISIS procedure consists of four steps as follows:

Step 1. In the initial stage, we apply the DC-SIS procedure for $\mathbf{y}$ and $\mathbf{x}$. Suppose we select p_1 predictors, which are denoted by $\mathcal{D}_1 = \left\{X_1^{(1)}, \ldots, X_{p_1}^{(1)}\right\}$, where $p_1 < d$, where d is user-specified model size. Fan and Lv (2008) suggested choosing $d = O(n/\log n)$. In this paper, we simply set $d = 2[n/\log n]$.

Step 2. Create new predictor variables $\mathbf{x}_{new}$ by regressing the screened-out variables on variables in $\mathcal{D}_1$. Specifically, denote by $\mathbf{X}_1$ the corresponding design matrix of variables in $\mathcal{D}_1$, and $\mathbf{X}_1^c$ the corresponding design matrix of variables in $\mathcal{D}_1^c$, the complement of $\mathcal{D}_1$. Define $\mathbf{X}_{new} = \left\{\mathbf{I}_n - \mathbf{X}_1 \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1} \mathbf{X}_1^{\mathrm{T}}\right\} \mathbf{X}_1^c$, the corresponding design matrix of new predictor variables $\mathbf{x}_{new}$. Then, apply the DC-SIS procedure for $\mathbf{y}$ and $\mathbf{x}_{new}$. Suppose we select p_2 predictors $\mathcal{D}_2 = \{X_1^{(2)}, \ldots, X_{p_2}^{(2)}\}$.

Step 3. Let $L = 2$ and $\mathcal{D}_S = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_L$. Repeat **Step 2** with replacement $\mathcal{D}_1$ and $\mathcal{D}_c^c$ with $\mathcal{D}_S$ and $\mathcal{D}_S^c$, respectively. Denote the selected p_{L+1} predictors by $\mathcal{D}_{L+1} = \{X_1^{(L+1)}, \ldots, X_{p_{L+1}}^{(L+1)}\}$.

Step 4. Let $L = 3, 4, \cdots, k$, repeat **Step 3** and update the selected predictors set with $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_k$ until $p_1 + p_2 + \cdots + p_k \geq d$.

How to decide the sizes p_i's can be challenging, and it usually depends upon model complexity. For the purpose of simplicity, we set all p_i's to be 5 in our simulation studies presented in next section. It will show that the proposed DC-ISIS can efficiently identify the active predictors that marginally independent of the responses.

Example 5. In this example, we compare the empirical performance of DC-ISIS with the DC-SIS, SIS and ISIS in the same linear model (3.4) designed by Fan and Lv (2008). We consider two different values of the correlation coefficient: (i) $\rho = 0.5$ and (ii) $\rho = 0.8$. To make the simulation more challenging, we vary the dimension p from 1000 to 5000 for the fixed sample size $n = 200$. Table 3.7 depicts the simulation results of $\mathcal{P}_s$ and $\mathcal{P}_a$ for each independence screening procedure. As illustrated before, X_4 is marginally independent of but jointly important to the response Y, so the SIS and DC-SIS can hardly select X_4. However, the proposed DC-ISIS is able to select X_4 perfectly in our model setting, and the ISIS also performs well to select X_4. We remark that the DC-ISIS doesn't implement any regression information, while the ISIS uses the true information of linear regression. In this sense, the DC-ISIS is model free and more flexible to various regression models, which can also be seen in the following examples.

Table 3.7. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 5 with the user-specified model size $d = 2[n/\log n]$.

		$\rho = 0.5$					$\rho = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
p	Method	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
1000	SIS	1.00	1.00	1.00	0.00	0.00	0.93	0.94	0.93	0.00	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
	DC-SIS	1.00	1.00	1.00	0.01	0.01	0.91	0.92	0.93	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	SIS	1.00	1.00	1.00	0.00	0.00	0.88	0.90	0.89	0.00	0.00
	ISIS	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.95	0.95
	DC-SIS	1.00	1.00	1.00	0.00	0.00	0.88	0.88	0.89	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Example 6. The proposed DC-ISIS can be directly implemented for the categorical response. This simulation is designed to access the finite-sample performance of the DC-ISIS for the regression models with the categorical response. We first

generate the response variable Y as same as Example 5, and then transform Y to a new categorical variable in the following two ways: (a) a binary response defined by $Y^* = \mathbf{1}(Y > 0)$, where $\mathbf{1}(\cdot)$ is the indicator function; (b) a multi-level categorical response defined by

$$Y^* = \begin{cases} 1, & \text{if } Y < -3; \\ 2, & \text{if } -3 \le Y < 0; \\ 3, & \text{if } 0 \le Y < 3; \\ 4, & \text{if } Y \ge 3. \end{cases}$$

In each simulated model, (X_1, X_2, X_3, X_4) are relevant to the new response Y^*, but X_4 is marginally independent of Y^*. We only report the simulation results for the sample size $n = 200$ and the dimension $p = 1000$ in the paper. The results for $p = 5000$ are available upon request. Table 3.8 summarizes the simulation results of $\mathcal{P}_s$ and $\mathcal{P}_a$ for both types of the new responses. It demonstrates that the DC-ISIS can improve the performance of the DC-SIS dramatically, in the sense that the DC-ISIS can select all relevant predictors with sufficiently high probability, especially X_4 which is always missed by the DC-SIS.

Table 3.8. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 6 with the user-specified model size $d = 2[n/\log n]$.

		$\rho = 0.5$					$\rho = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Y^*	Method	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
(a)	DC-SIS	1.00	1.00	1.00	0.00	0.00	0.89	0.89	0.88	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98
(b)	DC-SIS	1.00	1.00	1.00	0.01	0.01	0.90	0.90	0.90	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Example 7. Thanks to the appealing property of DC, both the DC-SIS and DC-ISIS can be directly used for screening grouped predictors. In many regres-

sion problems, several predictors can be represented by a group. For example, in the additive regression model with nonparametric components, each component may be expressed as a linear combination of several basis functions of the original variable, which can be naturally considered as a group. In the high/ultrahigh dimensional space, selecting the important original variables is equivalent to selecting the important groups of basis functions. Another common example is the multi-factor analysis-of-variance (ANOVA) problem (Yuan and Lin, 2008).

In this example, we generate p predictors $(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with entries $\sigma_{jj} = 1$ for $j = 1, 2, \ldots, p$, and $\sigma_{ij} = \rho_1$ for $i \neq j$. Then, we generate a new predictor X_4' from the standard normal distribution. X_4' is introduced to have the much higher correlation ρ_3 with X_4, and the lower correlation ρ_2 with other $p-1$ variables. Here, we consider two highly correlated variables X_4 and X_4' as one group, denoted by $\widetilde{X}_4 = \{X_4, X_4'\}$. Then, the predictors vector becomes $(X_1, X_2, X_3, \widetilde{X}_4, \cdots, X_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$.

We design the following model to generate the response variable Y:

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\left[\left(\frac{\rho_2\rho_3 - \rho_1}{\rho_3^2 + 1}\right)X_4 + \left(\frac{\rho_1\rho_3 - \rho_2}{\rho_3^2 + 1}\right)X_4'\right] + \varepsilon, \quad (3.5)$$

where ε is an independent random error from standard normal distribution $\mathcal{N}(0, 1)$. It can be shown that both X_4 and X_4' are marginally independent of but jointly important to the response Y, so is the grouped predictor $\widetilde{X}_4$. We remark here that the marginal utility of the grouped predictor $\widetilde{X}_4$ is defined by $\widehat{\omega}_4 = \widehat{\mathrm{dcorr}}^2(\widetilde{X}_4, Y)$. In this example, we consider two dimensions varying from 1000 to 5000 for the fixed sample size 200. We set $\rho_1 = 0.5$, $\rho_2 = 0.3$ and $\rho_3 = 0.9$, then the regression

model (3.5) becomes

$$Y = 5X_1 + 5X_2 + 5X_3 - 15(1.21X_4 - 0.79X_4') + \varepsilon.$$

The simulation results of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Table 3.9 demonstrate that the DC-ISIS is efficient to select all active predictors including the grouped variable $\widetilde{X}_4$, although $\widetilde{X}_4$ is marginally independent of the response Y.

Table 3.9. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 7 with the user-specified model size $d = 2[n/\log n]$.

		$\mathcal{P}_s$				$\mathcal{P}_a$
p	Method	X_1	X_2	X_3	X_4	ALL
1000	DC-SIS	1.00	1.00	1.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00
5000	DC-SIS	1.00	1.00	1.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	0.99	0.99

Example 8. The DC-SIS and DC-ISIS are available to screen the active predictors for the multivariate responses by the nature of DC. In this example, we investigate their empirical performance in the model with a bivariate response. We generate $(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = \rho^{|i-j|}$, for $i, j = 1, 2, \ldots, p$. We consider two covariance matrices with (i) $\rho = 0.5$ and (ii) $\rho = 0.8$, respectively. Here, the sample size n is fixed to be 200 and the dimension p varies from 2000 to 5000. For ease of interpretation, a bivariate response $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$ is generated by the following bivariate normal distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 5X_1 + 5X_2 - 5(\rho+\rho^2)X_3 \\ 5X_{11} + 5X_{12} - 5(\rho+\rho^2)X_{13} \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\}. \tag{3.6}$$

In consequence, X_3 and X_{13} are marginally independent of but jointly relevant

to the bivariate response variable $\mathbf{y}$. Table 3.10 indicates that the proportions that the active predictors are selected into the model by the DC-ISIS equal to one under the model setting, but the DC-SIS can only pick the variables X_3 and X_{13} by random chance. Thus, the DC-ISIS can efficiently select active predictors for multiple responses. This may be potentially useful in genetical pathway analyses in the ultrahigh dimensional space.

Table 3.10. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 8 with the user-specified model size $d = 2[n/\log n]$.

		$\rho = 0.5$							$\rho = 0.8$						
		$\mathcal{P}_s$						$\mathcal{P}_a$	$\mathcal{P}_s$						$\mathcal{P}_a$
p	Method	X_1	X_2	X_3	X_{11}	X_{12}	X_{13}	ALL	X_1	X_2	X_3	X_{11}	X_{12}	X_{13}	ALL
1000	DC-SIS	1.00	1.00	0.07	1.00	1.00	0.06	0.00	1.00	1.00	0.25	1.00	1.00	0.14	0.04
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	DC-SIS	1.00	1.00	0.02	1.00	1.00	0.01	0.00	1.00	1.00	0.09	1.00	1.00	0.01	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Example 9. In this example, we apply the DC-ISIS procedure for a rat eye expression dataset to conduct an empirical analysis and comparison with existing methods. This dataset was once used by Scheetz, et al. (2006) and Huang, Ma and Zhang (2008). In this dataset, 120 twelve-week-old male rats were selected for tissue harvesting from the eyes and for microarray analysis. The microarrays used to analyze the RNA from the eyes of these rats contain $31,042$ different probe sets. Following Huang, Ma and Zhang (2008), we excluded the probes that were not expressed sufficiently or that lacked sufficient variation, leaving 18,976 probes which satisfy these two criteria. The response variable TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang, et al., 2006), is one of the selected 18,976 probes. We then selected 3,000 probes with the largest variances from the remaining $18,975$ probes. The goal of our analysis is to identify the genes that are most relevant to the expression level of TRIM32 from the 3,000 candidate genes.

Huang, Ma and Zhang (2008) used the SIS to shrink the dimension p from

3,000 down to 200, then selected 19 genes with the adaptive LASSO (aLASSO) from these 200 candidate genes in the linear regression model. We refer to their procedure as HMZ(2008) below. To analyze this dataset, we first applied the DC-SIS and DC-ISIS procedures to shrink the dimension from $3,000$ down to $2[n/\log(n)] = 50$. Then, we conducted some exploratory data analysis for the selected 50 probes and the response, and found that each selected probe was linearly correlated to the response in some sense. Thus, we followed Huang, Ma and Zhang (2008) to apply linear regression model to the selected genes. For the purpose of comparison, we also implemented the aLASSO to select those most relevant genes from the selected 50 genes. The DC-ISIS+aLASSO eventually chose 12 genes, and the DC-SIS+aLASSO chose only 6 genes. The resulting dimensions are summarized in Table 3.11.

Table 3.11. Results of Example 9: rat eye expression dataset.

	#_of_Var	R^2	MSPE(RSD)
HMZ(2008)	19	73.4%	0.371(0.161)
DC-SIS+aLASSO	6	62.6%	0.411(0.232)
DC-ISIS+aLASSO	12	81.1%	0.232(0.106)

We fitted the data set by a linear model and compared their performance in terms of the adjusted R^2. The results are displayed in Table 3.11, in which the column labeled '#_of_Var' stands for the number of selected variables and the column of R^2 for the adjusted R^2. It can be clearly seen that the DC-ISIS+aLASSO performs the best with the largest R^2 value 81.1%, indicating that the iterative procedure identifies some features missed by the DC-SIS+aLASSO. Although HMZ(2008) selected 19 genes from 200 candidates, it does not perform as well as DC-ISIS+aLASSO, partly because SIS may miss some important features when the predictors are not normally distributed.

Next, we randomly partitioned the data into a training data, consisting of

100 observations, and a test data consisting of the remaining 20 observations. We fitted linear models respectively for HMZ(2008), DC-SIS+aLASSO and DC-ISIS+aLASSO with the training data, then calculated the prediction error in the test dataset. To be precise, we estimated the parameters from the training dataset, calculated the squared distance between the observed and predicted response values in the test dataset. The column labeled 'MSPE(RSD)' in Table 3.11 displays the median of the mean squared prediction errors (MSPE) with associated robust estimate of the standard deviation (RSD = IQR/1.34) in the parenthesis over 1000 repetitions. Once again, DC-ISIS+aLASSO produced the smallest MSPE, confirming the better performance of DC-ISIS+aLASSO.

3.6 Theoretical Proofs

3.6.1 Proof of Lemma 3.3.1

Following Székely, Rizzo and Bakirov (2007), we remark here that all involved integrals at 0 and ∞ are meaningful in the sense that $\lim_{\varepsilon \to 0} \int_{\mathbb{R}^d / \{\varepsilon B + \varepsilon^{-1} B^c\}}$, where B is the unit ball centered at 0 in $\mathbb{R}^d$. We first note that

$$|\phi_{X_k,\mathbf{y}}(t,\mathbf{s})|^2 = \phi_{X_k,\mathbf{y}}(t,\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s}).$$

After simple algebraic calculation, it follows that

$$\begin{aligned}
& \int_{\mathbb{R}^{q+1}} |\phi_{X_k,\mathbf{y}}(t,\mathbf{s}) - \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})|^2 \, w(t,\mathbf{s}) dt d\mathbf{s} \\
= & \int_{\mathbb{R}^{q+1}} \{\phi_{X_k,\mathbf{y}}(t,\mathbf{s}) - \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\} \{\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s}) - \phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})\} \, w(t,\mathbf{s}) dt d\mathbf{s} \\
= & \int_{\mathbb{R}^{q+1}} \phi_{X_k,\mathbf{y}}(t,\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s}) w(t,\mathbf{s}) dt d\mathbf{s}
\end{aligned}$$

$$
\begin{aligned}
- & \int_{\mathbb{R}^{q+1}} \phi_{X_k,\mathbf{y}}(t,\mathbf{s})\phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
- & \int_{\mathbb{R}^{q+1}} \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
+ & \int_{\mathbb{R}^{q+1}} \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s}.
\end{aligned}
$$

In the sequel we will deal with the above four quantities separately. Denote $F(X_k)$, $F(\mathbf{y})$ and $F(X_k,\mathbf{y})$ be the distribution functions of X_k, $\mathbf{y}$ and $(X_k,\mathbf{y})$, respectively. We first deal with the first term. Using the fact that $\exp(itX_k) = \cos(tX_k) + i\sin(tX_k)$ and the Fubini theorem, we obtain

$$
\begin{aligned}
& \int_{\mathbb{R}^{q+1}} \phi_{X_k,\mathbf{y}}(t,\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
= & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} \exp\left\{it(X_k-\widetilde{X}_k)+i\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}) \\
= & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} \cos\left\{t(X_k-\widetilde{X}_k)+\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}) \\
+ & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} i\sin\left\{t(X_k-\widetilde{X}_k)+\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}) \\
= & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} \cos\left\{t(X_k-\widetilde{X}_k)\right\}\cos\left\{\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}),
\end{aligned}
$$

where $(\widetilde{X}_k,\widetilde{\mathbf{y}})$ is an independent copy of $(X_k,\mathbf{y})$, and $S(X_k,\mathbf{y})$ and $S(\widetilde{X}_k,\widetilde{\mathbf{y}})$ are the supports of $(X_k,\mathbf{y})$ and $(\widetilde{X}_k,\widetilde{\mathbf{y}})$, respectively. Since $\sin(\cdot)w(\cdot)$ is an odd function, we can easily obtain that

$$
\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} \sin\left\{t(X_k-\widetilde{X}_k)+\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}) = 0,
$$

and

$$
\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}} \sin\left\{t(X_k-\widetilde{X}_k)\right\}\sin\left\{\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}}) = 0.
$$

Let $\mathbb{R}^d_\varepsilon = \mathbb{R}^d/\{\varepsilon B + \varepsilon^{-1}B^c\}$, where B is the unit ball centered at 0 in $\mathbb{R}^d$. Recall that the integrals at 0 and ∞ are meant in the following sense that $\lim_{\varepsilon\to 0}\int_{\mathbb{R}^d_\epsilon}$. Using Lemma 1 of Székely, Rizzo and Bakirov (2007) and the facts that

$$\begin{aligned}\cos u\cos v &= 1-(1-\cos u)-(1-\cos v)+(1-\cos u)(1-\cos v), \text{ and}\\ w(t,\mathbf{s}) &= \left\{c_1|t|_1^2\right\}^{-1}\left\{c_q|\mathbf{s}|_q^{q+1}\right\}^{-1} = w(t)w(\mathbf{s}),\end{aligned}$$

we have

$$\begin{aligned}&\int_{\mathbb{R}^{q+1}_\varepsilon}|\phi_{X_k,\mathbf{y}}(t,\mathbf{s})|^2\,w(t,\mathbf{s})dtd\mathbf{s}\\ =\;&\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}\int_{\mathbb{R}^{q+1}_\varepsilon}\Big[1-\Big\{1-\cos\Big(t(X_k-\widetilde{X}_k)\Big)\Big\}-\{1-\cos\left(\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right)\}\\ +\;&\Big\{1-\cos\Big(t(X_k-\widetilde{X}_k)\Big)\Big\}\Big\{1-\cos\left(\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right)\Big\}\Big]\,w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}})\\ =\;&\int_{\mathbb{R}^{q+1}_\varepsilon}w(t,\mathbf{s})dtd\mathbf{s}-\left\{\int_{\mathbb{R}^q_\varepsilon}w(\mathbf{s})d\mathbf{s}\right\}\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}|X_k-\widetilde{X}_k|_p dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}})\\ &-\left\{\int_{\mathbb{R}_\varepsilon}w(t)dt\right\}\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}|\mathbf{y}-\widetilde{\mathbf{y}}|_q dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}})\\ &+\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_k,\widetilde{\mathbf{y}})}|X_k-\widetilde{X}_k|_1|\mathbf{y}-\widetilde{\mathbf{y}}|_q dF(X_k,\mathbf{y})dF(\widetilde{X}_k,\widetilde{\mathbf{y}})\\ =\;&\int_{\mathbb{R}^{q+1}_\varepsilon}w(t,\mathbf{s})dtd\mathbf{s}-\left\{\int_{\mathbb{R}^q_\varepsilon}w(\mathbf{s})d\mathbf{s}\right\}E\{|X_k-\widetilde{X}_k|_1\}\\ &-\left\{\int_{\mathbb{R}_\varepsilon}w(t)dt\right\}E\{|\mathbf{y}-\widetilde{\mathbf{y}}|_q\}+E\left\{|X_k-\widetilde{X}_k|_1|\mathbf{y}-\widetilde{\mathbf{y}}|_q\right\}\\ =:\;&A_1-A_2-A_3+S_1^\varepsilon.\end{aligned}$$

The notations A_i's and S_1^ε are denoted in an obvious way. Next we turn to the second quantity. Assume $(X_k,\mathbf{y})$, $\widetilde{X}_{k1}$ and $\widetilde{\mathbf{y}}_2$ are mutually independent, and $\widetilde{X}_{k1}$ and $\widetilde{\mathbf{y}}_2$, are respective copies of X_k and $\mathbf{y}$. Using Fubini Theorem and similar

arguments for handling the first term, we obtain that

$$
\begin{aligned}
& \int_{\mathbb{R}_\varepsilon^{q+1}} \phi_{X_k,\mathbf{y}}(t,\mathbf{s})\phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
= & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)}\int_{\mathbb{R}_\varepsilon^{q+1}} \\
& \exp\left\{it(X_k-\widetilde{X}_{k1})+i\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}}_2)\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_{k1})dF(\widetilde{\mathbf{y}}_2) \\
= & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)}\int_{\mathbb{R}_\varepsilon^{q+1}} \\
& \cos\left\{t(X_k-\widetilde{X}_{k1})\right\}\cos\left\{\mathbf{s}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}}_2)\right\} w(t,\mathbf{s})dtd\mathbf{s}dF(X_k,\mathbf{y})dF(\widetilde{X}_{k1})dF(\widetilde{\mathbf{y}}_2) \\
= & \int_{\mathbb{R}_\varepsilon^{q+1}} w(t,\mathbf{s})dtd\mathbf{s} \\
- & \left\{\int_{\mathbb{R}_\varepsilon^{q}} w(\mathbf{s})d\mathbf{s}\right\}\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)} |X_k-\widetilde{X}_{k1}|_1 dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{X}_{k1})dF(X_k,\mathbf{y}) \\
- & \left\{\int_{\mathbb{R}_\varepsilon} w(t)dt\right\}\int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)} |\mathbf{y}-\widetilde{\mathbf{y}}_2|_q dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{X}_{k1})dF(X_k,\mathbf{y}) \\
+ & \int_{S(X_k,\mathbf{y})}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)} |X_k-\widetilde{X}_{k1}|_1|\mathbf{y}-\widetilde{\mathbf{y}}_2|_q dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{X}_{k1})dF(X_k,\mathbf{y}) \\
= & \int_{\mathbb{R}_\varepsilon^{q+1}} w(t,\mathbf{s})dtd\mathbf{s} - \left\{\int_{\mathbb{R}_\varepsilon^{q}} w(\mathbf{s})d\mathbf{s}\right\} E\{|X_k-\widetilde{X}_k|_1\} \\
- & \left\{\int_{\mathbb{R}_\varepsilon} w(t)dt\right\} E\{|\mathbf{y}-\widetilde{\mathbf{y}}|_q\} + E\left\{E\left(|X_k-\widetilde{X}_k|_1 \mid X_k\right) E\left(|\mathbf{y}-\widetilde{\mathbf{y}}|_q \mid \mathbf{y}\right)\right\} \\
= & A_1 - A_2 - A_3 + S_3^\varepsilon.
\end{aligned}
$$

Since

$$
\begin{aligned}
& \int_{\mathbb{R}_\varepsilon^{q+1}} \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
= & \int_{\mathbb{R}_\varepsilon^{q+1}} \phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})\phi_{X_k,\mathbf{y}}(t,\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s}.
\end{aligned}
$$

The third term is identical to the second term. Consequently,

$$
\int_{\mathbb{R}_\varepsilon^{q+1}} \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\phi_{X_k,\mathbf{y}}(-t,-\mathbf{s})w(t,\mathbf{s})dtd(\mathbf{s}) = A_1 - A_2 - A_3 + S_3^\varepsilon.
$$

Finally we move to the last quantity. Let X_{k1}, $\widetilde{X}_{k1}$, $\mathbf{y}_2$ and $\widetilde{\mathbf{y}}_2$ be mutually independent random variables. In addition, the first two have identical distribution, and the last term have identical distribution as well. Following similar arguments, we can obtain that

$$
\begin{aligned}
& \int_{\mathbb{R}_\varepsilon^{q+1}} \phi_{X_k}(t)\phi_{\mathbf{y}}(\mathbf{s})\phi_{X_k}(-t)\phi_{\mathbf{y}}(-\mathbf{s})w(t,\mathbf{s})dtd\mathbf{s} \\
= & \int_{S(X_{k1})}\int_{S(\mathbf{y}_2)}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)}\int_{\mathbb{R}_\varepsilon^{q+1}} \exp\left\{it(X_{k1}-\widetilde{X}_{k1}) + i\mathbf{s}^{\mathrm{T}}(\mathbf{y}_2-\widetilde{\mathbf{y}}_2)\right\} \\
& w(t,\mathbf{s})dtd\mathbf{s}dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{X}_{k1})dF(\mathbf{y}_2)dF(X_{k1}) \\
= & \int_{S(X_{k1})}\int_{S(\mathbf{y}_2)}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)}\int_{\mathbb{R}_\varepsilon^{q+1}} \cos\left\{t(X_{k1}-\widetilde{X}_{k1})\right\}\cos\left\{\mathbf{s}^{\mathrm{T}}(\mathbf{y}_2-\widetilde{\mathbf{y}}_2)\right\} \\
& w(t,\mathbf{s})dtd\mathbf{s}dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{X}_{k1})dF(\mathbf{y}_2)dF(X_{k1}) \\
= & \int_{\mathbb{R}_\varepsilon^{q+1}} w(t,\mathbf{s})dtd\mathbf{s} \\
- & \left\{\int_{\mathbb{R}_\varepsilon^{q}} w(\mathbf{s})d\mathbf{s}\right\}\int_{S(X_{k1})}\int_{S(\mathbf{y}_2)}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)} |X_{k1}-\widetilde{X}_{k1}|_1 dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{\mathbf{x}}_1)dF(\mathbf{y}_2)dF(\mathbf{x}_1) \\
- & \left\{\int_{\mathbb{R}_\varepsilon} w(t)dt\right\}\int_{S(X_{k1})}\int_{S(\mathbf{y}_2)}\int_{S(\widetilde{X}_{k1})}\int_{S(\widetilde{\mathbf{y}}_2)} |\mathbf{y}_2-\widetilde{\mathbf{y}}_2|_q dF(\widetilde{\mathbf{y}}_2)dF(\widetilde{\mathbf{x}}_1)dF(\mathbf{y}_2)dF(\mathbf{x}_1) \\
+ & \int_{S(X_{k1})}\int_{S(\widetilde{X}_{k1})} |X_{k1}-\widetilde{X}_{k1}|_p dF(\widetilde{X}_{k1})dF(X_{k1}) \int_{S(\mathbf{y}_2)}\int_{S(\widetilde{\mathbf{y}}_2)} |\mathbf{y}_2-\widetilde{\mathbf{y}}_2|_q dF(\widetilde{\mathbf{y}}_2)dF(\mathbf{y}_2) \\
= & \int_{\mathbb{R}_\varepsilon^{q+1}} w(t,\mathbf{s})dtd\mathbf{s} - \left\{\int_{\mathbb{R}_\varepsilon^{q}} w(\mathbf{s})d\mathbf{s}\right\} E\left\{|X_k-\widetilde{X}_k|_1\right\} \\
- & \left\{\int_{\mathbb{R}_\varepsilon} w(t)dt\right\} E\Big\{|\mathbf{y}-\widetilde{\mathbf{y}}|_q\Big\} + E\left\{|X_k-\widetilde{X}_k|_1\right\} E\Big\{|\mathbf{y}-\widetilde{\mathbf{y}}|_q\Big\} \\
= & A_1 - A_2 - A_3 + S_2^\varepsilon.
\end{aligned}
$$

Hence, we can conclude that

$$
\begin{aligned}
& \lim_{\varepsilon\to 0}\{(A_1-A_2-A_3+S_1^\varepsilon)-(A_1-A_2-A_3+S_3^\varepsilon) \\
& -(A_1-A_2-A_3+S_3^\varepsilon)+(A_1-A_2-A_3+S_2^\varepsilon)\} \\
= & \lim_{\varepsilon\to 0}(S_1^\varepsilon+S_2^\varepsilon-2S_3^\varepsilon) = S_1+S_2-2S_3.
\end{aligned}
$$

This completes the proof of Lemma 3.3.1. □

3.6.2 Proof of Thoerem 3.3.4

We aim to show the uniform consistency of the denominator and the numerator of $\widehat{\omega}_k$ under regularity conditions respectively. Because the denominator of $\widehat{\omega}_k$ has a similar form as the numerator, we deal with its numerator only below. Throughout proof, the notations C and c are generic constants which may take different values at each appearance.

We first deal with $\widehat{S}_{k1}$. Define $\widehat{S}^*_{k1} = \{n(n-1)\}^{-1}\sum_{i\neq j}\|X_{ik} - X_{jk}\|_1\|\mathbf{y}_i - \mathbf{y}_j\|_q$, which is a usual U-statistic. We shall establish the uniform consistency of $\widehat{S}^*_{k1}$ by using the theory of U-statistics (Serfling, 1980, Section 5). By using the Cauchy-Schwartz inequality,

$$\begin{aligned} S_{k1} &= E\left(\|X_{ik} - X_{jk}\|_1\|\mathbf{y}_i - \mathbf{y}_j\|_q\right) \leq \left\{E\left(\|X_{ik} - X_{jk}\|_1^2\right)E\left(\|\mathbf{y}_i - \mathbf{y}_j\|_q^2\right)\right\}^{1/2} \\ &\leq 4\left\{E(X_k^2)E\|\mathbf{y}\|_q^2\right\}^{1/2}. \end{aligned}$$

This together with condition (C3.1) implies that S_{k1} is uniformly bounded in p, that is, $\sup_p \max_{1\leq k\leq p} S_{k1} < \infty$. For any given $\varepsilon > 0$, take n large enough such that $S_{k1}/n < \varepsilon$. Then it can be easily shown that

$$\begin{aligned} \Pr\left(\left|\widehat{S}_{k1} - S_{k1}\right| \geq 2\varepsilon\right) &= \Pr\left\{\left|\widehat{S}^*_{k1}(n-1)/n - S_{k1}(n-1)/n - S_{k1}/n\right| \geq 2\varepsilon\right\} \\ &\leq \Pr\left\{\left|\widehat{S}^*_{k1} - S_{k1}\right|(n-1)/n \geq 2\varepsilon - S_{k1}/n\right\} \\ &\leq \Pr\left(\left|\widehat{S}^*_{k1} - S_{k1}\right| \geq \varepsilon\right). \end{aligned} \tag{3.7}$$

To establish the uniform consistency of $\widehat{S}_{k1}$, it thus suffices to show the uniform consistency of $\widehat{S}^*_{k1}$. Let $h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) = \|X_{ik} - X_{jk}\|_1\|\mathbf{y}_i - \mathbf{y}_j\|_q$ be the kernel

of the U-statistic $\widehat{S}^*_{k1}$. We decompose the kernel function h_1 into two parts: $h_1 = h_1\mathbf{1}(h_1 > M) + h_1\mathbf{1}(h_1 \leq M)$ where M will be specified later. The U-statistic can now be written as follows,

$$\begin{aligned}
\widehat{S}^*_{k1} &= \{n(n-1)\}^{-1} \sum_{i \neq j} h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) \leq M\right\} \\
&+ \{n(n-1)\}^{-1} \sum_{i \neq j} h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) > M\right\} \\
&= \widehat{S}^*_{k1,1} + \widehat{S}^*_{k1,2}.
\end{aligned}$$

Accordingly, we decompose S_{k1} into two parts:

$$\begin{aligned}
S_{k1} &= E\left[h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) \leq M\right\}\right] \\
&+ E\left[h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j)\mathbf{1}\left\{h_1(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j) > M\right\}\right] \\
&= S_{k1,1} + S_{k1,2}.
\end{aligned}$$

Clearly, $\widehat{S}^*_{k1,1}$ and $\widehat{S}^*_{k1,2}$ are the respectively unbiased estimates of $S_{k1,1}$ and $S_{k1,2}$.

We deal with the consistency of $\widehat{S}^*_{k1,1}$ first. With the Markov's inequality, for any $t > 0$, we can obtain that

$$\Pr(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \varepsilon) \leq \exp\left(-t\varepsilon\right) \exp(-tS_{k1,1}) E\{\exp(t\widehat{S}^*_{k1,1})\}.$$

Serfling (1980, Section 5.1.6, Pages 180-181) showed that any U-statistic can be represented as an average of averages of independent and identically distributed (i.i.d) random variables; that is, $\widehat{S}^*_{k1,1} = (n!)^{-1} \sum\limits_{n!} \Omega_1(X_{1k}, \mathbf{y}_1; \cdots; X_{nk}, \mathbf{y}_n)$, where $\sum\limits_{n!}$ denotes the summation over all possible permutations of $(1, \ldots, n)$, and each $\Omega_1(X_{1k}, \mathbf{y}_1; \cdots; X_{nk}, \mathbf{y}_n)$ is an average of $m = [n/2]$ i.i.d random variables that is, $\Omega_1 = m^{-1} \sum\limits_r h_1^{(r)} \mathbf{1}\{h_1^{(r)} \leq M\}$. Since the exponential function is convex, it follows

from Jensen's inequality that, for $0 < t \leq 2s_0$,

$$\begin{aligned} E\{\exp(t\widehat{S}^*_{k1,1})\} &= E\big[\exp\big\{t(n!)^{-1}\sum_{n!}\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)\big\}\big] \\ &\leq (n!)^{-1}\sum_{n!}E\left[\exp\left\{t\Omega_1(X_{1k},\mathbf{y}_1;\cdots;X_{nk},\mathbf{y}_n)\right\}\right] \\ &= E^m\big\{\exp\big(m^{-1}th_1^{(r)}\mathbf{1}\{h_1^{(r)}\leq M\}\big)\big\}, \end{aligned}$$

which together with Lemma 3.3.2 entails immediately that

$$\begin{aligned} \Pr(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \varepsilon) &\leq \exp(-t\varepsilon)\,E^m\big\{\exp\big(m^{-1}t\big[h_1^{(r)}\mathbf{1}\{h_1^{(r)}\leq M\} - S_{k1,1}\big]\big)\big\} \\ &\leq \exp\big\{-t\varepsilon + M^2t^2/(8m)\big\}. \end{aligned}$$

By choosing $t = 4\varepsilon m/M^2$, we have $\Pr(\widehat{S}^*_{k1,1} - S_{k1,1} \geq \varepsilon) \leq \exp(-2\varepsilon^2 m/M^2)$. Therefore, by the symmetry of U-statistic, we can obtain easily that

$$\Pr\big(\big|\widehat{S}^*_{k1,1} - S_{k1,1}\big| \geq \varepsilon\big) \leq 2\exp\big(-2\varepsilon^2 m/M^2\big). \tag{3.8}$$

Next we show the consistency of $\widehat{S}^*_{k1,2}$. With Cauchy-Schwartz and Markov's inequality,

$$\begin{aligned} S^2_{k1,2} &\leq E\left\{h_1^2(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\right\}\Pr\left\{h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j) > M\right\} \\ &\leq E\left\{h_1^2(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\right\}E\left[\exp\left\{s'h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\right\}\right]/\exp(s'M), \end{aligned}$$

for any $s' > 0$. Using the fact $(a^2+b^2)/2 \geq (a+b)^2/4 \geq |ab|$, we have

$$\begin{aligned} & h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j) = \left\{(X_{ik}-X_{jk})^2(\mathbf{y}_i-\mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i-\mathbf{y}_j)\right\}^{1/2} \\ \leq\ & 2\left\{\left(X_{ik}^2+X_{jk}^2\right)\left(\|\mathbf{y}_i\|_q^2+\|\mathbf{y}_j\|_q^2\right)\right\}^{1/2} \leq \left\{\left(X_{ik}^2+X_{jk}^2+\|\mathbf{y}_i\|_q^2+\|\mathbf{y}_j\|_q^2\right)^2\right\}^{1/2} \\ =\ & X_{ik}^2+X_{jk}^2+\|\mathbf{y}_i\|_q^2+\|\mathbf{y}_j\|_q^2, \end{aligned}$$

which yields that

$$\begin{aligned} E\left[\exp\left\{s'h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j)\right\}\right] &\leq E\left[\exp\left\{s'\left(X_{ik}^2+X_{jk}^2+\|\mathbf{y}_i\|_q^2+\|\mathbf{y}_j\|_q^2\right)\right\}\right] \\ &\leq E\left\{\exp\left(2s'X_{ik}^2\right)\right\}E\left\{\exp\left(2s'\|\mathbf{y}_i\|_q^2\right)\right\}. \end{aligned}$$

The last inequality follows from the Cauchy-Schwartz inequality. If we choose $M = cn^{\gamma}$ for $0 < \gamma < 1/2 - \kappa$, then $S_{k1,2} \leq \varepsilon/2$ when n is sufficiently large. Consequently,

$$\Pr\left(\left|\widehat{S}_{k1,2}^{*} - S_{k1,2}\right| > \varepsilon\right) \leq \Pr\left(\left|\widehat{S}_{k1,2}^{*}\right| > \varepsilon/2\right). \tag{3.9}$$

It remains to bound the probability $\Pr\left(\left|\widehat{S}_{k1,2}^{*}\right| > \varepsilon/2\right)$. We observe that the events satisfy

$$\left\{\left|\widehat{S}_{k1,2}^{*}\right| > \varepsilon/2\right\} \subseteq \left\{X_{ik}^2 + \|\mathbf{y}_i\|_q^2 > M/2, \text{ for some } 1 \leq i \leq p\right\}. \tag{3.10}$$

To see this, we assume that $X_{ik}^2 + \|\mathbf{y}_i\|_q^2 \leq M/2$ for all $1 \leq i \leq p$. This assumption will lead to a contradiction. To be precise, under this assumption, $h_1(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j) \leq X_{ik}^2 + X_{jk}^2 + \|\mathbf{y}_i\|_q^2 + \|\mathbf{y}_j\|_q^2 \leq M$. Consequently, $\left|\widehat{S}_{k1,2}^{*}\right| = 0$, which is a contrary to the event $\left|\widehat{S}_{k1,2}^{*}\right| > \varepsilon/2$. This verifies the relation (3.10) is true.

By invoking condition (C3.1), there must exist a constant C such that

$$\begin{aligned} \Pr(\|X_k\|_1^2 + \|\mathbf{y}\|_q^2 \geq M/2) &\leq \Pr(\|X_k\|_1 \geq \sqrt{M}/2) + \Pr(\|\mathbf{y}\|_q \geq \sqrt{M}/2) \\ &\leq 2C\exp(-sM/4). \end{aligned}$$

The last inequality follows from Markov's inequality for $s > 0$. Consequently,

$$\begin{aligned} \max_{1\le k\le p} \Pr\big(\big|\widehat{S}^*_{k1,2}\big| > \varepsilon/2\big) &\le n \max_{1\le k\le p} \Pr(\|X_k\|_1^2 + \|\mathbf{y}\|_q^2 \ge M/2) \\ &\le 2nC\exp(-sM/4). \end{aligned} \tag{3.11}$$

Recall that $M = cn^\gamma$. Combining the results (3.8), (3.9) and (3.11), we have

$$\Pr\big(\big|\widehat{S}_{k1} - S_{k1}\big| \ge 4\varepsilon\big) \le 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^\gamma/4\right). \tag{3.12}$$

In the sequel we turn to $\widehat{S}_{k2}$. We write $\widehat{S}_{k2} = \widehat{S}_{k2,1}\widehat{S}_{k2,2}$, where $\widehat{S}_{k2,1} = n^{-2}\sum_{i\ne j}\|X_{ik} - X_{jk}\|_1$, and $\widehat{S}_{k2,2} = n^{-2}\sum_{i\ne j}\|\mathbf{y}_i - \mathbf{y}_j\|_q$. Similarly, we write $S_{k2} = S_{k2,1}S_{k2,2}$, where $S_{k2,1} = E\{\|X_{ik} - X_{jk}\|_1\}$ and $S_{k2,2} = E\{\|\mathbf{y}_i - \mathbf{y}_j\|_q\}$. Following arguments for proving (3.12) we can show that

$$\begin{aligned} &\Pr\big(\big|\widehat{S}_{k2,1} - S_{k2,1}\big| \ge 4\varepsilon\big) \le 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^{2\gamma}/4\right), \text{ and} \\ &\Pr\big(\big|\widehat{S}_{k2,2} - S_{k2,2}\big| \ge 4\varepsilon\big) \le 2\exp\left(-\varepsilon^2 n^{1-2\gamma}\right) + 2nC\exp\left(-sn^{2\gamma}/4\right). \end{aligned} \tag{3.13}$$

Condition (C3.1) ensures that $S_{k2,1} \le \{E(\|X_{ik} - X_{jk}\|_1^2)\}^{1/2} \le \{4E(X_k^2)\}^{1/2}$ and $S_{k2,2} \le \big\{E(\|\mathbf{y}_i - \mathbf{y}_j\|_q^2)\big\}^{1/2} \le \big\{4E(\|\mathbf{y}\|_q^2)\big\}^{1/2}$ are uniformly bounded. That is,

$$\max\Big\{\max_{1\le k\le p} S_{k2,1}, S_{k2,2}\Big\} \le C,$$

for some constant C. Using (3.13) repetitively, we can easily prove that

$$\begin{aligned} &\Pr\big\{\big|(\widehat{S}_{k2,1} - S_{k2,1})S_{k2,2}\big| \ge \varepsilon\big\} \le \Pr\big(\big|\widehat{S}_{k2,1} - S_{k2,1}\big| \ge \varepsilon/C\big) \\ &\le 2\exp\left\{-\varepsilon^2 n^{1-2\gamma}/(16C^2)\right\} + 2nC\exp\left(-sn^{2\gamma}/4\right), \\ &\Pr\big(\big|S_{k2,1}(\widehat{S}_{k2,2} - S_{k2,2})\big| \ge \varepsilon\big) \le \Pr\big(\big|\widehat{S}_{k2,2} - S_{k2,2}\big| \ge \varepsilon/C\big) \\ &\le 2\exp\left\{-\varepsilon^2 n^{1-2\gamma}/(16C^2)\right\} + 2nC\exp\left(-sn^{2\gamma}/4\right), \end{aligned} \tag{3.14}$$

and

$$\begin{aligned}
&\Pr\big\{\big|(\widehat{S}_{k2,1}-S_{k2,1})(\widehat{S}_{k2,2}-S_{k2,2})\big|\geq\varepsilon\big\}\\
\leq&\Pr\big(\big|\widehat{S}_{k2,1}-S_{k2,1}\big|\geq\sqrt{\varepsilon}\big)+\Pr\big(\big|\widehat{S}_{k2,2}-S_{k2,2}\big|\geq\sqrt{\varepsilon}\big)\\
\leq&4\exp\left(-\varepsilon n^{1-2\gamma}/16\right)+4nC\exp\left(-sn^{2\gamma}/4\right).
\end{aligned}\tag{3.15}$$

It follows from Bonferroni's inequality, inequalities (3.14) and (3.15) that,

$$\begin{aligned}
&\Pr\left(\left|\widehat{S}_{k2}-S_{k2}\right|\geq3\varepsilon\right)=\Pr\left(\left|\widehat{S}_{k2,1}\widehat{S}_{k2,2}-S_{k2,1}S_{k2,2}\right|\geq3\varepsilon\right)\\
\leq&\Pr\left\{\left|(\widehat{S}_{k2,1}-S_{k2,1})S_{k2,2}\right|\geq\varepsilon\right\}+\Pr\left\{\left|S_{k2,1}(\widehat{S}_{k2,2}-S_{k2,2})\right|\geq\varepsilon\right\}\\
&+\Pr\left\{\left|(\widehat{S}_{k2,1}-S_{k2,1})(\widehat{S}_{k2,2}-S_{k2,2})\right|\geq\varepsilon\right\}\\
\leq&2\exp\left\{-\varepsilon^2n^{1-2\gamma}/(16C^2)\right\}+2nC\exp\left(-sn^{2\gamma}/4\right)\\
&+2\exp\left\{-\varepsilon^2n^{1-2\gamma}/(16C^2)\right\}+2nC\exp\left(-sn^{2\gamma}/4\right)\\
&+4\exp\left(-\varepsilon n^{1-2\gamma}/16\right)+4nC\exp\left(-sn^{2\gamma}/4\right)\\
\leq&8\exp\left\{-\varepsilon^2n^{1-2\gamma}/(16C^2)\right\}+8nC\exp\left(-sn^{2\gamma}/4\right),
\end{aligned}\tag{3.16}$$

where the last inequality holds for ε sufficiently small and C sufficiently large.

It remains to the uniform consistency of $\widehat{S}_{k3}$. We first study the following U-statistic:

$$\begin{aligned}
\widehat{S}^*_{k3} &= \frac{1}{n(n-1)(n-2)}\sum_{i<j<l}\Big\{\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_j-\mathbf{y}_l\|_q+\|X_{ik}-X_{lk}\|_1\|\mathbf{y}_j-\mathbf{y}_l\|_q+\\
&\qquad\|X_{ik}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_l\|_q+\|X_{lk}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_l\|_q+\\
&\qquad\|X_{lk}-X_{jk}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q+\|X_{lk}-X_{ik}\|_1\|\mathbf{y}_i-\mathbf{y}_j\|_q\Big\}\\
&=: \frac{6}{n(n-1)(n-2)}\sum_{i<j<l}h_3(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j;X_{lk},\mathbf{y}_l).
\end{aligned}\tag{3.17}$$

Here, $h_3(X_{ik},\mathbf{y}_i;X_{jk},\mathbf{y}_j;X_{lk},\mathbf{y}_l)$ is the kernel of U-statistic $\widehat{S}^*_{k3}$. Following the

arguments to deal with $\widehat{S}^*_{k1}$, we decompose h_3 into two parts: $h_3 = h_3\mathbf{1}(h_3 > M) + h_3\mathbf{1}(h_3 \le M)$. Accordingly,

$$\begin{aligned}
\widehat{S}^*_{k3} &= \frac{6}{n(n-1)(n-2)} \sum_{i<j<l} h_3\mathbf{1}(h_3 \le M) + \frac{6}{n(n-1)(n-2)} \sum_{i<j<l} h_3\mathbf{1}(h_3 > M) \\
&= \widehat{S}^*_{k3,1} + \widehat{S}^*_{k3,2}, \\
S_{k3} &= E\left\{h_3\mathbf{1}(h_3 \le M)\right\} + E\left\{h_3\mathbf{1}(h_3 > M)\right\} = S_{k3,1} + S_{k3,2}.
\end{aligned}$$

Following similar arguments for proving (3.8), we can show that

$$\Pr\big(\big|\widehat{S}^*_{k3,1} - S_{k3,1}\big| \ge \varepsilon\big) \le 2\exp\left(-2\varepsilon^2 m'/M^2\right), \tag{3.18}$$

where $m' = [n/3]$ because $\widehat{S}^*_{k3,1}$ is a third-order U-statistic.

Then we deal with $\widehat{S}^*_{k3,2}$. We observe that $h_3(X_{ik}, \mathbf{y}_i; X_{jk}, \mathbf{y}_j; X_{lk}, \mathbf{y}_l) \le 4(X^2_{ik} + X^2_{jk} + X^2_{lk} + \|\mathbf{y}_i\|^2_q + \|\mathbf{y}_j\|^2_q + \|\mathbf{y}_l\|^2_q)/6$, which will be smaller than M if $X^2_{ik} + \|\mathbf{y}_i\|^2_q \le M/2$ for all $1 \le i \le p$. Thus, for any $\varepsilon > 0$, the events satisfy

$$\big\{\big|\widehat{S}^*_{k3,2}\big| > \varepsilon/2\big\} \subseteq \big\{X^2_{ik} + \|\mathbf{y}_i\|^2_q > M/2, \text{ for some } 1 \le i \le p\big\}.$$

By using the similar arguments to prove (3.11), it follows that

$$\Pr\big(\big|\widehat{S}^*_{k3,2} - S_{k3,2}\big| > \varepsilon\big) \le \Pr\big(\big|\widehat{S}^*_{k3,2}\big| > \varepsilon/2\big) \le 2nC\exp(-sM/4). \tag{3.19}$$

Then, we combine the results (3.18) and (3.19) with $M = cn^\gamma$ for some $0 < \gamma < 1/2 - \kappa$ to obtain that

$$\Pr\left(\left|\widehat{S}^*_{k3} - S_{k3}\right| \ge 2\varepsilon\right) \le 2\exp\left(-2\varepsilon^2 n^{1-2\gamma}/3\right) + 2nC\exp\left(-sn^\gamma/4\right). \tag{3.20}$$

By the definition of $\widehat{S}_{k3}$,

$$\widehat{S}_{k3} = \frac{(n-1)(n-2)}{n^2}\left\{\widehat{S}^*_{k3} + \frac{1}{(n-2)}\widehat{S}^*_{k1}\right\}.$$

Thus, using similar techniques to deal with $\widehat{S}_{k1}$, we can obtain that

$$\begin{aligned}\Pr\left(\left|\widehat{S}_{k3} - S_{k3}\right| \geq 4\varepsilon\right) &= \Pr\left\{\left|\frac{(n-1)(n-2)}{n^2}\left(\widehat{S}^*_{k3} - S_{k3}\right) - \frac{3n-2}{n^2}S_{k3}\right.\right. \\ &\qquad \left.\left. + \frac{n-1}{n^2}\left(\widehat{S}^*_{k1} - S_{k1}\right) + \frac{n-1}{n^2}S_{k1}\right| \geq 4\varepsilon\right\}.\end{aligned}$$

Using similar arguments for dealing with S_{k1}, we can show that S_{k3} is uniformly bounded in p. Taking n large enough such that $\{(3n-2)/n^2\}S_{k3} \leq \varepsilon$ and $\{(n-1)/n^2\}S_{k1} \leq \varepsilon$, then

$$\begin{aligned}\Pr\big(\left|\widehat{S}_{k3} - S_{k3}\right| \geq 4\varepsilon\big) &\leq \Pr\big(\left|\widehat{S}^*_{k3} - S_{k3}\right| \geq \varepsilon\big) + \Pr\big\{\left|\widehat{S}^*_{k1} - S_{k1}\right| \geq \varepsilon\big\} \\ &\leq 4\exp\left(-\varepsilon^2 n^{1-2\gamma}/6\right) + 4nC\exp\left(-sn^{\gamma}/4\right).\end{aligned} \tag{3.21}$$

The last inequality follows from (3.12) and (3.20). This, together with (3.12), (3.16) and the Bonferroni's inequality, implies

$$\begin{aligned}&\Pr\big\{\left|(\widehat{S}_{k1} + \widehat{S}_{k2} - 2\widehat{S}_{k3}) - (S_{k1} + S_{k2} - 2S_{k3})\right| \geq \varepsilon\big\} \\ \leq&\Pr\big(\left|\widehat{S}_{k1} - S_{k1}\right| \geq \varepsilon/4\big) + \Pr\big(\left|\widehat{S}_{k2} - S_{k2}\right| \geq \varepsilon/4\big) + \Pr\big(\left|\widehat{S}_{k3} - S_{k3}\right| \geq \varepsilon/4\big) \\ =&O\left\{\exp\left(-c_1\varepsilon^2 n^{1-2\gamma}\right) + n\exp\left(-c_2 n^{\gamma}\right)\right\},\end{aligned} \tag{3.22}$$

for some positive constants c_1 and c_2. The convergence rate of the numerator of $\widehat{\omega}_k$ is now achieved. Following similar arguments, we can obtain the convergence rate of the denominator. In effect the convergence rate of $\widehat{\omega}_k$ has the same form of (3.22). We omit the details here. Let $\varepsilon = cn^{-\kappa}$, where κ satisfies $0 < \kappa + \gamma < 1/2$.

We thus have

$$\begin{aligned}\Pr\Big\{\max_{1\le k\le p}|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\Big\} &\le p\max_{1\le k\le p}\Pr\left\{|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\right\}\\ &\le O\left(p\left[\exp\left\{-c_1n^{1-2(\kappa+\gamma)}\right\}+n\exp\left(-c_2n^{\gamma}\right)\right]\right).\end{aligned}$$

The first part of Theorem 3.3.4 is proven.

Now we deal with the second part of Theorem 3.3.4. If $\mathcal{D}\nsubseteq\widehat{\mathcal{D}}^{\star}$, then there must exist some $k\in\mathcal{D}$ such that $\widehat{\omega}_k<cn^{-\kappa}$. It follows from condition (C3.2) that $|\widehat{\omega}_k-\omega_k|>cn^{-\kappa}$ for some $k\in\mathcal{D}$, indicating that the events satisfy $\left\{\mathcal{D}\nsubseteq\widehat{\mathcal{D}}^{\star}\right\}\subseteq\left\{|\widehat{\omega}_k-\omega_k|>cn^{-\kappa},\ \text{for some } k\in\mathcal{D}\right\}$, and hence $\mathcal{E}_n=\left\{\max_{k\in\mathcal{D}}|\widehat{\omega}_k-\omega_k|\le cn^{-\kappa}\right\}\subseteq\left\{\mathcal{D}\subseteq\widehat{\mathcal{D}}^{\star}\right\}$. Consequently,

$$\begin{aligned}\Pr(\mathcal{D}\subseteq\widehat{\mathcal{D}}^{\star}) &\ge \Pr(\mathcal{E}_n)=1-\Pr(\mathcal{E}_n^c)=1-\Pr\Big(\min_{k\in\mathcal{D}}|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\Big)\\ &= 1-s_n\Pr\left\{|\widehat{\omega}_k-\omega_k|\ge cn^{-\kappa}\right\}\\ &\ge 1-O\left(s_n\left[\exp\left\{-c_1n^{1-2(\kappa+\gamma)}\right\}+n\exp\left(-c_2n^{\gamma}\right)\right]\right),\end{aligned}$$

where s_n is the cardinality of $\mathcal{D}$. This completes the proof of the second part.

□

Chapter 4

Robust Feature Screening and Variable Selection for Ultrahigh Dimensional Heteroscedastic Single-Index Models

4.1 Introduction

To explore the relationship between a response variable $Y \in \mathbb{R}$ and a covariate vector $\mathbf{x} = (X_1, \ldots, X_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n}$, regression analysis often decomposes Y into two parts, namely, $Y = E(Y \mid \mathbf{x}) + \varepsilon$ where ε denotes an independent error. The linear regression which assumes that $E(Y \mid \mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n}$ represents an unknown parameter vector, has been extensively studied in the literature. Due to its simplicity and interpretability, the linear regression lays out a foundation of contemporary regression analysis. However, in usual practice the linear regression is insufficient to capture how the mean function of Y varies with the value of $\mathbf{x}$. To enhance the flexibility while maintaining the interpretability

of linear regression, Härdle, Hall and Ichimura (1993) proposed the single-index model which assumes that $E(Y \mid \mathbf{x}) = G(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ for some unknown function $G(\cdot)$. That is, the mean function of Y depends on $\mathbf{x}$ through a single linear combination $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. Consequently, the single-index model avoids the "curse of dimensionality", which makes it popular in high dimensional data analysis. Powell, Stock and Stoker (1989), Duan and Li (1991), Horowitz and Härdle (1996), Carroll et al. (1997) and Xia et al. (2002) systematically studied the parameter estimation of single-index models and the associated theoretical properties.

In this chapter, we consider the following heteroscedastic single-index model

$$Y = G(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \sigma(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\varepsilon. \tag{4.1}$$

For identification purpose, we assume the independent error term ε has zero mean and unit variance. Because both $G(\cdot)$ and $\sigma(\cdot)$ are unknown functions, the index parameter $\boldsymbol{\beta}$ is not identifiable. Thus, the direction of $\boldsymbol{\beta}$, rather than its true value, is of primary interest in the literature. Our goal is to identify the indices of the zero elements of $\boldsymbol{\beta}$ and to estimate the magnitudes of the nonzero elements of $\boldsymbol{\beta}$ up to a proportionality constant.

The precision of parameter estimation and the accuracy of response prediction will deteriorate substantially if there are a large number of irrelevant covariates included in the model (Altham, 1984; Fan and Li, 2001). In model (4.1), these truly irrelevant covariates correspond to the coordinates of $\boldsymbol{\beta}$ with values being exactly zero. When the covariate dimension is very high and yet smaller than the sample size, some regularization procedures for variable selection are prominent in terms of removing the irrelevant covariates from the single-index models. See, for example, Naik and Tsai (2001), Kong and Xia (2007), Zhu, Qian and Lin (2011) and Liang et al. (2010) and references therein. When the dimension of predictors is much larger

than the sample size, however, how to remove these truly irrelevant covariates still remains a challenging problem. In such an ultrahigh dimensional setting, the regularization methods for variable selection may fail to perform well due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009).

To ease the computational complexity of ultrahigh dimensional data analysis, Fan and Lv (2008) proposed the sure independence screening (SIS) procedure to reduce the ultrahigh dimensionality down to a relatively moderate scale. They proposed to rank the importance of each covariate through its marginal correlation with the response variable and select the covariates highly correlated with the response variable. The SIS is computationally expedient. In addition, Fan and Lv (2008) proved that the SIS processes the sure screening property when the covariates and the response are jointly normal. That is, in an asymptotic sense it guarantees to pick up all truly important covariates. Fan and Song (2010) generalized the idea of SIS and proposed to utilize the marginal maximum likelihood. Zhu, Li, Li and Zhu (2011) suggested a sure independent ranking and screening (SIRS) procedure and proved that the SIRS has the ranking consistency property. That is, it ensures to rank the truly important covariates in the top asymptotically.

In this chapter, we propose the two-stage feature screening and variable selection procedure for the heteroscedastic single-index model (4.1) with ultrahigh dimensional covariates. Our goal is to identify the truly important covariates and to estimate the direction of $\boldsymbol{\beta}$. In the first stage, we propose a robust independent ranking and screening (RIRS) procedure for feature screening in an ultrahigh dimensional space and show that the RIRS possesses the sure screening property in the terminology of Fan and Lv (2008) and the ranking consistency property in the terminology of Zhu, Li, Li and Zhu (2011). In addition, we demonstrate that the

RIRS is robust to the extreme values and outliers in the response variable. The RIRS procedure can reduce the ultrahigh dimensionality to a moderate scale, and it guarantees to select all truly important covariates. However, some unimportant covariates will be chosen as well. Thus, in the second stage, we propose to apply penalized linear quantile regression to further exclude the unimportant covariates and to estimate the direction of $\boldsymbol{\beta}$ in model (4.1). We also study the theoretical properties of the resultant estimator and show that it still consistent up to a proportionality constant, and has the oracle property in the terminology of Fan and Li (2001). The two-step estimation procedure avoids completely estimating the nonlinear function $G(\cdot)$ and $\sigma(\cdot)$, and is computationally expedient in ultrahigh dimensional setting. Aside from this, we demonstrate through comprehensive numerical studies that the whole procedure presents an outstanding finite sample performance.

The rest of this chapter is organized as follows. In Section 4.2, we illustrate the rationale of the RIRS procedure for feature screening and establish its ranking consistency and sure screening properties. In Section 4.3, we introduce the penalized linear quantile regression and study the consistency and the oracle property of the resultant penalized estimator. In Section 4.4, we compare the finite sample performance of the RIRS with other competitors through comprehensive simulations and an application to a real dataset. We also assess the finite sample performance of linear quantile regression with different penalties. All technical proofs are given in the Section 4.5.

4.2 Robust Independent Ranking and Screening

4.2.1 Some Preliminaries

Let Y be the response with support Ψ_y, and $\mathbf{x} = (X_1, \ldots, X_{p_n})^{\mathrm{T}}$ be the predictor vector. We denote by $F(y \mid \mathbf{x})$ the conditional distribution function of Y given $\mathbf{x}$. Without specifying a regression model, we define the active and inactive predictors by

$$\begin{aligned} \mathcal{A} &= \{k : F(y \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(y \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } y \in \Psi_y\}. \end{aligned}$$

We further write $\mathbf{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$ and refer to $\mathbf{x}_{\mathcal{A}}$ as an active predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ as an inactive predictor vector. We consider that the conditional distribution function $F(y \mid \mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ for some parameter $\boldsymbol{\beta}$. That is, $F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. This model framework contains a large number of parametric and semiparametric models, where the response Y depends on the predictors $\mathbf{x}$ only through linear combination $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, including the heteroscedastic single-index model (4.1).

In an ultrahigh dimensional setting where the covariate dimension p_n greatly exceeds the available sample size n, it is natural to assume the sparsity principle that only a small number of covariates are truly relevant to Y. Accordingly, many coordinates of $\boldsymbol{\beta}$ are zero. Thus, in terms of model (4.1), $\mathcal{A} = \{k : \beta_k \neq 0\}$ and $\mathcal{I} = \{k : \beta_k = 0\}$. We further denote by $\boldsymbol{\beta}_{\mathcal{A}}$ the nonzero coordinates of $\boldsymbol{\beta}$, namely, $\boldsymbol{\beta}_{\mathcal{A}} = \{\beta_k : k \neq \mathcal{A}\}$. When the sparsity principle applies, model (4.1) indicates immediately that

$$Y = G(\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}) + \sigma(\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\varepsilon. \tag{4.2}$$

We observe that model (4.1) together with (4.2) implies that

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = F(y \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}), \quad \text{for } y \in \Psi_y \text{ and } \mathbf{x} \in \mathbb{R}^{p_n}, \tag{4.3}$$

It also indicates that $\mathbf{x}$ and Y are statistically independent when $\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}$ is given. Our primary goal in this section is to develop a robust marginal utility to rank the importance of each covariate. We anticipate the robust marginal utility to behave well when extreme values and outliers are present in the observed values of Y.

4.2.2 The Robust Marginal Utility

Suppose $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample of $(\mathbf{x}, Y)$. For the sake of notational clarity, we assume throughout that the covariates have been standardized marginally, namely, $n^{-1}\sum_{i=1}^{n} X_{ik} = 0$ and $n^{-1}\sum_{i=1}^{n} X_{ik}^2 = 1$ for $1 \leq k \leq p_n$. Motivated by the SIS procedure (Fan and Lv, 2008) which ranks the importance of X_k through the sample estimator of the marginal correlation between X_k and Y, we propose to rank the importance through the sample estimator of the marginal correlation between X_k and the rank of Y. The rank of Y at the sample level has the form of $R_i = \sum_{j=1}^{n} \mathbf{1}(Y_j \leq Y_i)$. Thus, up to a proportionality constant, the marginal correlation can be equivalently written as

$$\widehat{\omega}_k = n^{-2}\sum_{i=1}^{n} X_{ik}R_i.$$

We propose to rank the importance of X_k, $k = 1, \ldots, p_n$, through the magnitude of the marginal utility $\widehat{\omega}_k$. With a pre-specified threshold γ_n, we select a set of variables

$$\widehat{\mathcal{A}} = \left\{k : \widehat{\omega}_k^2 \geq \gamma_n\right\}. \tag{4.4}$$

We except $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ which ensures to select all truly active covariates.

We remark here that the marginal utility $\widehat{\omega}_k$ is equivalent to the sample estimator of the marginal correlation between X_k and the marginal distribution function of Y. To be precise, we denote by $F_n(Y_i) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(Y_j \leq Y_i)$ the empirical distribution function of Y. Then $\widehat{\omega}_k$ can be rewritten equivalently as $\widehat{\omega}_k = n^{-1} \sum_{i=1}^{n} \{X_{ik} F_n(Y_i)\}$. It can be easily seen that $\widehat{\omega}_k$ is a sample estimator of $\omega_k = E\{X_k F(Y)\}$, where $F(y) = E\{\mathbf{1}(Y \leq y)\}$ denotes the marginal distribution function of Y. We make the following two observations.

(i) Because we utilize a bounded transformation of Y instead of Y itself in calculating the marginal utility $\widehat{\omega}_k$, it is robust to the presence of extreme values and outliers in Y.

(ii) Let the notation $\perp\!\!\!\perp$ stand for statistical independence. Because the transformation function $F(\cdot)$ is monotone, model (4.1) implies that $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}} \boldsymbol{\beta}_{\mathcal{A}}$ which is equivalent to $F(Y) \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}} \boldsymbol{\beta}_{\mathcal{A}}$. We can expect naturally that ω_k captures all the regression information of model (4.1) and hence model (4.3).

4.2.3 Theoretical Properties

Next, we investigate the theoretical properties of the proposed robust independent ranking and screening (RIRS) procedure. We assume the following conditions to establish the ranking consistency property in the terminology of Zhu, Li, Li and Zhu (2011), which ensures that all the truly active covariates are ranked above the inactive ones with an overwhelming probability, and the sure screening property in the terminology of Fan and Lv (2008), which guarantees to select all truly active covariates in an asymptotic sense.

(A4.1) The following inequality holds uniformly in p:

$$\max_{k\in\mathcal{I}} \left\{\mathrm{cov}^2(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\right\} < \min_{k\in\mathcal{A}} \left\{\mathrm{cov}^2(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\right\}. \tag{4.5}$$

(A4.2) The linearity condition

$$E\{\mathbf{x} \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}\} = \mathrm{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_{\mathcal{A}} \left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_{\mathcal{A}}\right\}^{-1} \boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}. \tag{4.6}$$

(A4.3) The covariates $\mathbf{x}$ satisfy the sub-exponential tail probability uniformly in p. That is, there exist positive constants t_0 and C such that

$$\max_{1\le k\le p} E\left\{\exp(t|X_k|)\right\} \le C < \infty, \text{ for } 0 < t \le t_0. \tag{4.7}$$

(A4.4) The minimal signal of the truly active covariates satisfies that

$$\min_{k\in\mathcal{A}} \left\{\mathrm{cov}^2(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\right\} > c_1 n^{-\kappa}, \tag{4.8}$$

for some positive constants c_1 and $0 \le \kappa < 1/2$.

Condition (A4.1) is intuitive and it requires in spirit that the correlation between the truly active covariates and Y be stronger than the correlation between the truly inactive covariates and Y. It allows for arbitrary mean and variance functions in model (4.1), and hence retains the model-free flavor in the sense of Zhu, Li, Li and Zhu (2011). It is always true when there is only a single active covariate. This condition rules out the situations in which the truly active covariates are highly correlated with the truly inactive covariates. In this regard, the condition is parallel to conditions 3 and 4 in Fan and Lv (2008) and condition (C1) in Zhu, Li, Li and Zhu (2011). The linearity condition (A4.2) follows if $\mathbf{x}$ has an elliptically symmetric

distribution (Fang, Kotz and Ng, 1989). Hall and Li (1993) demonstrated that, no matter what the covariate distribution is, the linearity condition always offers an ideal approximation of the reality as long as p_n is sufficient large. Therefore, the linearity condition is typically regarded as mild in an ultrahigh dimensional setting. The sub-exponential tail condition (A4.3) assumes essentially that all moments of the covariates are uniformly bounded. It holds true when the covariates follow multivariate normal distribution or are bounded uniformly. This condition is widely assumed in analysis of ultrahigh dimensional data to derive exponential inequalities. See, for example, Bickel and Levina (2008) and Zhu, Li, Li and Zhu (2011). Condition (A4.4) is assumed in Fan and Lv (2008) and Fan and Song (2010). This condition is often considered as a minimum requirement for ensuring the sure screening property.

Theorem 4.2.1 states the ranking consistency property at the population level. That is, the marginal utility of an active covariate is always larger than that of an inactive covariate.

Theorem 4.2.1. *If conditions (A4.1) and (A4.2) hold, then the following inequality holds uniformly in p,*

$$\max_{k\in\mathcal{I}}\omega_k^2 < \min_{k\in\mathcal{A}}\omega_k^2.$$

Theorem 4.2.2 states the ranking consistency property at the sample level.

Theorem 4.2.2. (Ranking Consistency Property) *If condition (A4.3) holds, then for any positive ϵ sufficiently small, there exists some positive constant a_1 such that*

$$\Pr\left(\max_{1\le k\le p}\left|\widehat{\omega}_k - \omega_k\right| \ge \epsilon\right) \le 2(p_n+1)\exp\left(-a_1 n\epsilon^2\right), \tag{4.9}$$

If, in addition, conditions (A4.1) and (A4.2) hold, then

$$\Pr\left(\max_{k\in\mathcal{I}}\widehat{\omega}_k^2 \geq \min_{k\in\mathcal{A}}\widehat{\omega}_k^2\right) \leq 2(p_n+1)\exp\left(-a_2 n\delta^2\right), \tag{4.10}$$

where $\delta = \min_{k\in\mathcal{A}}\omega_k^2 - \max_{k\in\mathcal{I}}\omega_k^2 > 0$ *and* a_2 *is some positive constant.*

It is worthwhile to observe that, as long as $p_n = o\{\exp(an)\}$ with any constant $a > 0$,

$$\Pr\left(\max_{k\in\mathcal{I}}\widehat{\omega}_k^2 < \min_{k\in\mathcal{A}}\widehat{\omega}_k^2\right) \to 1, \text{ as } n \to \infty.$$

That is, the proposed RIRS guarantees to rank the truly active covariates above the inactive ones with probability approaching one as $n \to \infty$.

Next we investigate the sure screening property in the terminology of Fan and Lv (2008). The following lemma paves the road for establishing the sure screening property.

Lemma 4.2.3. *If condition (A4.4) holds, there exists a positive constant* $c_2 > 0$ *such that*

$$\min_{k\in\mathcal{A}}\omega_k^2 \geq c_2 n^{-\kappa}. \tag{4.11}$$

Lemma 4.2.3 indicates that the signal of an active covariate is at least as large as the order of $n^{-\kappa}$ for some $0 \leq \kappa < 1/2$. Theorem 4.2.4 presents the sure screening property, which guarantees that the RIRS procedure can select all truly active covariates in an asymptotic sense.

Theorem 4.2.4. (Sure Screening Property) *If conditions (A4.3)-(A4.4) hold and we set* $\gamma_n = c_3 n^{-\kappa}$ *with* $c_3 \leq c_2/2$, *there exists a positive constant* $a_3 > 0$ *such that, for* n *sufficiently large,*

$$\Pr\left(\mathcal{A} \subseteq \widehat{\mathcal{A}}\right) \geq 1 - 2(q_n + 1)\exp\left(-a_3 n^{1-2\kappa}\right), \tag{4.12}$$

where q_n *is the cardinality of* $\mathcal{A}$.

Theorem 4.2.4 implies that the RIRS maintains the sure screening property even when $\log p_n = o(n^{1-2\kappa})$ under some mild conditions. This is the same rate obtained by Fan and Lv (2008) in the linear model with Gaussian predictors and response. This result is also slightly stronger than Fan and Song (2010) who permits $\log p_n = o\{n^{(1-2\kappa)\alpha/(\alpha+2)}\}$. In addition, the proposed RIRS is robust to the presence of heteroscedasticity and outliers in the response. The robustness is an appealing property of the RIRS as an independence screening procedure.

4.3 Penalized Linear Quantile Regression

4.3.1 Motivations

In the previous section, we propose the RIRS procedure to remove the truly inactive covariates. The ranking consistency and the sure screening properties ensure that the RIRS procedure guarantees to select all active covariates, however, some inactive covariates may be remained in the selected model as well. In this section, we will discuss how to refine the selection by removing those truly inactive covariates remained in the preceding RIRS stage. We are also interested in estimating the nonzero coefficients of the truly active covariates in model (4.1) up to a proportionality constant.

In the second stage, we only retain the covariates selected by the first screening stage and let $\{(\mathbf{x}_{i,\widehat{\mathcal{A}}}, Y_i), i = 1, \ldots, n\}$ denote the working dataset, where $\mathbf{x}_{i,\widehat{\mathcal{A}}}$ represents the coordinates of $\mathbf{x}_i$ indexed by $\widehat{\mathcal{A}}$. We denote by $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ the coordinates of $\boldsymbol{\beta}$ indexed by $\widehat{\mathcal{A}}$, and d_n the cardinality of $\widehat{\mathcal{A}}$, namely, $d_n = |\widehat{\mathcal{A}}|$. Without loss of generality, we write $\mathbf{x}_{i,\widehat{\mathcal{A}}} = (X_{i,1}, \ldots, X_{i,d_n})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{\widehat{\mathcal{A}}} = (\beta_1, \ldots, \beta_{d_n})^{\mathrm{T}}$ in the sequel. Both the ranking consistency and the sure screening properties in Theorems 4.2.2 and 4.2.4 ensure that $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ holds asymptotically, which allows us to rewrite model (4.1), or equivalently model (4.2), as follows,

$$Y = G(\mathbf{x}_{\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}}) + \sigma(\mathbf{x}_{\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}})\varepsilon. \tag{4.13}$$

Because some truly inactive covariates may be selected after the RIRS procedure, some coordinates of $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ will be exactly zero. Therefore, it is important to further refine the selection by removing those inactive covariates and to estimate the magnitudes of the nonzero coefficients up to a proportionality constant.

REMARK. In the model (4.13), merely for the technical proofs, we assume that the index $\widehat{\mathcal{A}}$ is independent of the data $\{(\mathbf{x}_{i,\widehat{\mathcal{A}}}, Y_i), i = 1, \ldots, n\}$, although $\widehat{\mathcal{A}}$ is obtained from the same data in the first screening stage. For a rigorous theoretic development, one may randomly partition the data into two parts, the screening set and the cleaning set. In the first stage, the RIRS procedure is applied to the screening set, which completely determine the estimated index set $\widehat{\mathcal{A}}$. In the second stage, the penalized linear quantile regression is implemented to the cleaning set. Consequently, $\widehat{\mathcal{A}}$ is independent with the cleaning set, and thus the above technical assumption does not need any more. The comprehensive simulation studies in the latter section show that both the estimation using the all data for both two stages and that using the separate set for each stage have the similarly outstanding finite sample performance.

4.3.2 The Penalized Estimation

The primary interest in this section is to identify the index set $\mathcal{A}$ and estimate the direction of $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$. We first illustrate the rationale of estimating $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ through linear quantile regression at the population level. Let $\rho_\tau(r) = \tau r - rI(r < 0)$ be the check loss function at the τ-quantile. Define

$$\mathcal{L}_\tau(u, \mathbf{b}) = E\{\rho_\tau(Y - u - \mathbf{x}_{\widehat{\mathcal{A}}}^{\mathrm{T}}\mathbf{b})\}, \text{ and } (u_\tau^o, \boldsymbol{\beta}_\tau^o) = \underset{u,\mathbf{b}}{\operatorname{argmin}}\{\mathcal{L}_\tau(u, \mathbf{b})\}, \quad (4.14)$$

where $\mathbf{b} = (b_1, \ldots, b_{d_n})^{\mathrm{T}} \in \mathbb{R}^{d_n}$. Zhu, Huang and Li (2011) proved that, if the linearity condition (A4.2) holds and $\mathcal{A} \subseteq \widehat{\mathcal{A}}$, then $\boldsymbol{\beta}_\tau^o$ is proportional to $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ for arbitrary mean and variance functions $G(\cdot)$ and $\sigma(\cdot)$ in model (4.13). Accordingly, we call $\boldsymbol{\beta}_\tau^o$ the true linear quantile estimator. This observation motivates us to consider the following penalized linear quantile regression

$$Q(u, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - u - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\mathbf{b}) + \sum_{j=1}^{d_n} p_\lambda(|b_j|). \quad (4.15)$$

We implement the penalty function $p_\lambda(\cdot)$ to shrink some small values of $\mathbf{b}$ to zero. We advocate using the SCAD penalty for its unbiasedness, continuity and sparsity properties (Fan and Li, 2001). The SCAD penalty is defined as follows,

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

where $a = 3.7$ suggested by Fan and Li (2001). The penalized estimator $(\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau)$ at the τ-th quantile is defined through

$$(\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau) = \underset{u,\mathbf{b}}{\operatorname{argmin}}\{Q(u, \mathbf{b})\}. \tag{4.16}$$

The penalized quantile regression identifies the indices to nonzero components of $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ and simultaneously estimates its direction. In addition, it maintains the merit of the RIRS in that it is insensitive to the presence of extreme values and outliers in the response.

4.3.3 The Oracle Property

In this section we study the asymptotic property of the oracle estimator in the linear quantile regression. We define this oracle estimator at the population level as

$$\mathcal{L}_\tau(u, \mathbf{b}_1) = E\{\rho_\tau(Y - u - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mathbf{b}_1)\}, \quad \text{and} \quad (u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o) = \underset{u,\mathbf{b}_1}{\operatorname{argmin}}\{\mathcal{L}_\tau(u, \mathbf{b}_1)\}, \tag{4.17}$$

where $\mathbf{b}_1 = (b_1, \ldots, b_{q_n})^{\mathrm{T}} \in \mathbb{R}^{q_n}$. If $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ holds, without lose of generality, we can denote $\boldsymbol{\beta}_\tau^o = (\boldsymbol{\beta}_{\tau 1}^{o\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{\tau 1}^o$ represents a q_n-dimensional vector of nonzero components associated with the covariates indexed by $\mathcal{A}$ and $\mathbf{0}$ denotes a $(d_n - q_n)$-dimensional vector of zeros.

Accordingly, we define it at the sample level as

$$\mathcal{L}_{\tau n}(u, \mathbf{b}_1) = \frac{1}{n}\sum_{i=1}^{n}\{\rho_\tau(Y_i - u - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\mathbf{b}_1)\}, \quad \text{and} \quad (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_{\tau 1}^o) = \underset{u,\mathbf{b}_1}{\operatorname{argmin}}\{\mathcal{L}_{\tau n}(u, \mathbf{b}_1)\}. \tag{4.18}$$

In addition to conditions (A4.1)-(A4.4), we need further assume the following regularity conditions to facilitate the derivations of the consistency of the oracle

estimator and the oracle property of the penalized linear quantile regression.

(A4.5) There exist positive constants $0 < C_1 \leq C_2 < \infty$, such that

$$C_1 \leq \lambda_{\min}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \leq \lambda_{\max}\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\} \leq C_2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest eigenvalue and largest eigenvalues, respectively. In addition, suppose that $(\mathbf{x}_{i,\mathcal{A}}, Y_i)$ are in general positions (Koenker, 2005, Section 2.2), for $i = 1, 2, \ldots, n$.

(A4.6) The probability density function of $Y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau$ conditional on $\mathbf{x}$, denoted by $f(\cdot \mid \mathbf{x})$, is uniformly bounded away from 0 and ∞ in the neighborhood around u_τ^o.

(A4.7) The true model dimension q_n satisfies $q_n = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.

(A4.8) For $\boldsymbol{\beta}_{\tau 1}^o = (\beta_{\tau,1}^o, \beta_{\tau,2}^o, \ldots, \beta_{\tau,q_n}^o)^{\mathrm{T}}$, there exist positive constants c_2 and C such that $2c_1 < c_2 \leq 1$ and

$$\min_{1 \leq j \leq q_n} |\beta_{\tau,j}^o| \geq C n^{-(1-c_2)/2}.$$

Lemma 4.3.1 states the convergence rate of the oracle estimators $\widehat{u}_\tau^o$ and $\widehat{\boldsymbol{\beta}}_{\tau 1}^o$.

Lemma 4.3.1. *Suppose Conditions (A4.3) and (A4.5)-(A4.7) hold, then the oracle estimators $\widehat{u}_\tau^o$ and $\widehat{\boldsymbol{\beta}}_{\tau 1}^o$ satisfy*

$$\|\widehat{\boldsymbol{\beta}}_{\tau 1}^o - \boldsymbol{\beta}_{\tau 1}^o\| = O_p\big(\sqrt{q_n/n}\big) \quad \textit{and} \quad \|\widehat{u}_\tau^o - u_\tau^o\| = O_p\big(\sqrt{q_n/n}\big). \tag{4.19}$$

Next, we study the oracle property of the resulting estimator in the following Theorem 4.3.2.

Theorem 4.3.2. *Suppose Conditions (A4.3) and (A4.5)-(A4.8) hold,* $d_n = O(n)$ *and* $\lambda = o\{n^{-(1-c_2)/2}\}$. *Let* $\mathcal{B}_n(\lambda)$ *be the set of local minima* $\widehat{\boldsymbol{\beta}}_\tau$ *of the objective function* $Q(u, \mathbf{b})$ *with the SCAD penalty and tuning parameter* λ. *The oracle estimator* $\widehat{\boldsymbol{\beta}}_\tau^o = (\widehat{\boldsymbol{\beta}}_{\tau 1}^{oT}, \mathbf{0}^T)^T$ *satisfies*

$$\Pr\left\{\widehat{\boldsymbol{\beta}}_\tau^o \in \mathcal{B}_n(\lambda)\right\} \to 1, \quad as \;\; n \to \infty.$$

Theorem 4.3.2 implies that the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o$ is a local minimizer of the objective function (4.15) with the probability approaching one as $n \to \infty$. This is different from Theorem 2.4 of Wang, Wu and Li (2012) where the underlying true mode is linear whereas in our context the underlying true model is possible nonlinear. Therefore, together with the result that $\boldsymbol{\beta}_\tau^o$ is proportional to $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ and $\mathcal{A} \subseteq \widehat{\mathcal{A}}$, the direction of the estimator $\widehat{\boldsymbol{\beta}}_\tau$ of the objective function $Q(u, \mathbf{b})$ is asymptotically equivalent to that of the index parameter $\boldsymbol{\beta}_{\widehat{\mathcal{A}}}$ in model (4.13).

4.4 Numerical Studies

4.4.1 Simulations

Example 1. In this example, we assess the finite sample performance of the proposed RIRS by Monte Carlo simulation. Our simulation studies were conducted using R code. In this simulation example, we generate $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ from a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, where we set $\sigma_{ij} = 0.5^{|i-j|}$. We consider the dimensionality $p = 1000$ and the sample size $n = 200$. We repeat each experiment 500 times, and evaluate the performance through the following three criteria.

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the

5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 500 replications.

2. $\mathcal{P}_s$: the proportion that an individual active predictor is selected for a given model size d in the 500 replications.

3. $\mathcal{P}_a$: the proportion that all active predictors are selected for a given model size d in the 500 replications.

When the ranking consistency property holds, we expect $\mathcal{S}$ to be close to the number of truly active predictors. We also expect $\mathcal{S}$ to offer an approximation of the cardinality of $\mathcal{A}$ when the active predictors are ranked in the top. The sure screening property ensures that $\mathcal{P}_s$ and $\mathcal{P}_a$ are both close to one when the estimated model size d is sufficiently large. Followed the thresholding suggested by Fan and Lv (2008), we choose d to be $d_1 = [n/\log n]$ and $d_2 = 2[n/\log n]$, where $[a]$ denotes the integer part of a.

This example is designed to compare the performance of the proposed robust procedure RIRS with the SIS (Fan and Lv, 2008) and DC-SIS (Li, Zhong and Zhu, 2012). We consider the following true single index

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = 3X_1 + 1.5X_2 + 2X_7,$$

so X_1, X_2 and X_7 are truly important predictors out of the 1000 candidates. This setting was originated by Fan and Li (2001). Then, we generate the response from the following four single-index models including the linear model, and consider different error terms to obtain both homoscedastic and heteroscedastic data.

$$\textbf{Model (I)} \quad : \quad Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon \tag{4.20}$$

$$\textbf{Model (II)} \quad : \quad Y = \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - 1\} + \varepsilon \tag{4.21}$$

$$\textbf{Model (III)} \quad : \quad Y = \exp\{2 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2\} + \{2 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2\}^2 + \varepsilon \tag{4.22}$$

$$\textbf{Model (IV)} \quad : \quad Y = \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - 1\} * \sin^2\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\} + \varepsilon \tag{4.23}$$

where ε is the error term independent of $\mathbf{x}$. We consider three different error terms in the following.

Scenario 1: $\varepsilon \sim \mathcal{N}(0, 1)$, standard normal distribution;

Scenario 2: $\varepsilon \sim \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2)\mathcal{N}(0, 1)$;

Scenario 3: $\varepsilon \sim t(1)$, t distribution with degree freedom 1.

REMARK. The error term in Scenario 2 depends on the single index $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ and makes the response heteroscedastic. $t(1)$ in Scenario 3 is the well-known Cauchy distribution and has the heavy probability tails to produce outliers in the response easily.

Tables 4.1 and 4.2 depict the simulation results for $\mathcal{S}$, $\mathcal{P}_s$ and $\mathcal{P}_a$. We can see that the finite sample performances of four independence screening procedures are similarly good for the linear model with ordinary normal error. However, in the presence of heteroscedasticity in the response, the SIS and the DC-SIS do not perform well. On the other hand, the RIRS can perform very well and select the truly important predictors with very high probability in our model settings.

Table 4.1. The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications.

Model	Error	SIS					DC-SIS					RIRS				
	Scenario 1	3.0	3.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0
(I)	Scenario 2	6.0	90.5	337.0	721.0	945.2	3.0	3.0	3.0	5.0	30.0	3.0	3.0	3.0	5.0	24.0
	Scenario 3	3.0	10.0	119.5	585.2	942.2	3.0	3.0	3.0	4.0	41.2	3.0	3.0	3.0	3.0	4.0
	Scenario 1	5.0	28.0	75.5	196.0	641.1	3.0	6.0	24.0	78.8	298.2	3.0	3.0	3.0	3.0	4.0
(II)	Scenario 2	4.0	22.8	63.5	170.2	660.2	3.0	5.0	17.0	75.0	380.1	3.0	3.0	3.0	3.0	4.0
	Scenario 3	5.0	25.8	76.5	178.0	642.0	3.0	5.0	20.0	88.2	373.2	3.0	3.0	3.0	4.0	6.0
	Scenario 1	3.0	3.0	5.0	15.0	122.2	3.0	3.0	3.0	3.0	6.0	3.0	3.0	3.0	3.0	4.0
(III)	Scenario 2	3.0	3.0	5.0	19.0	198.3	3.0	3.0	3.0	4.0	11.0	3.0	3.0	3.0	3.0	5.0
	Scenario 3	3.0	3.0	5.0	16.2	286.1	3.0	3.0	3.0	4.0	7.0	3.0	3.0	3.0	3.0	4.0
	Scenario 1	5.0	27.0	76.0	202.0	696.0	3.0	7.0	25.0	101.0	479.8	3.0	3.0	3.0	4.0	8.0
(IV)	Scenario 2	5.0	26.8	83.5	243.0	737.0	3.0	6.0	28.0	103.0	377.8	3.0	3.0	4.0	7.0	32.1
	Scenario 3	6.0	31.8	93.5	242.2	610.3	3.0	9.0	32.0	118.2	386.1	3.0	3.0	4.0	6.0	39.1

Table 4.2. The empirical probabilities of each active predictor (denoted by $\mathcal{P}_s$) and all active predictors (denoted by $\mathcal{P}_a$) are chosen for a given model size.

		SIS				DC-SIS				RIRS			
		$\mathcal{P}_s$			$\mathcal{P}_a$	$\mathcal{P}_s$			$\mathcal{P}_a$	$\mathcal{P}_s$			$\mathcal{P}_a$
Model	Size	X_1	X_2	X_7	ALL	X_1	X_2	X_7	ALL	X_1	X_2	X_7	ALL
Scenario 1: $\varepsilon \sim \mathcal{N}(0,1)$													
(I)	d_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(II)	d_1	0.97	0.83	0.45	0.34	0.99	0.96	0.62	0.59	1.00	1.00	1.00	1.00
	d_2	0.99	0.90	0.58	0.50	1.00	0.98	0.76	0.74	1.00	1.00	1.00	1.00
(III)	d_1	1.00	1.00	0.87	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	d_2	1.00	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(IV)	d_1	0.96	0.83	0.44	0.33	0.99	0.94	0.60	0.56	1.00	1.00	0.99	0.99
	d_2	0.99	0.90	0.57	0.50	1.00	0.98	0.72	0.70	1.00	1.00	0.99	0.99
Scenario 2: $\varepsilon \sim \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2)\mathcal{N}(0,1)$													
(I)	d_1	0.56	0.44	0.27	0.15	1.00	1.00	0.96	0.96	1.00	1.00	0.97	0.97
	d_2	0.62	0.53	0.34	0.24	1.00	1.00	0.98	0.98	1.00	1.00	0.98	0.98
(II)	d_1	0.97	0.82	0.50	0.36	1.00	0.94	0.69	0.63	1.00	1.00	1.00	1.00
	d_2	0.99	0.89	0.63	0.53	1.00	0.97	0.77	0.75	1.00	1.00	1.00	1.00
(III)	d_1	1.00	0.97	0.82	0.81	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
	d_2	1.00	0.98	0.88	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(IV)	d_1	0.96	0.80	0.40	0.30	1.00	0.94	0.60	0.56	1.00	1.00	0.94	0.94
	d_2	0.99	0.89	0.55	0.47	1.00	0.97	0.71	0.68	1.00	1.00	0.97	0.97
Scenario 3: $\varepsilon \sim t(1)$													
(I)	d_1	0.67	0.59	0.42	0.36	0.99	0.98	0.95	0.95	1.00	1.00	1.00	1.00
	d_2	0.71	0.64	0.51	0.43	0.99	0.99	0.96	0.97	1.00	1.00	1.00	1.00
(II)	d_1	0.98	0.82	0.44	0.32	1.00	0.95	0.64	0.61	1.00	1.00	0.99	0.99
	d_2	0.99	0.90	0.57	0.50	1.00	0.97	0.74	0.72	1.00	1.00	1.00	1.00
(III)	d_1	0.98	0.97	0.85	0.84	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
	d_2	0.98	0.98	0.90	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(IV)	d_1	0.95	0.81	0.42	0.30	0.99	0.92	0.59	0.53	1.00	1.00	0.95	0.95
	d_2	0.99	0.87	0.56	0.45	1.00	0.97	0.69	0.67	1.00	1.00	0.98	0.98

Example 2. In this example, we will examine the finite sample performance of the proposed penalized linear quantile regression for the single-index models. We generate simulated data from the single-index Model (III) with the same sample size $n = 200$ and dimensionality $p = 1000$; That is,

$$Y = \exp(2 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2) + (2 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2)^2 + \varepsilon, \tag{4.24}$$

where we also consider three different error terms ε, which are defined in section 4.4.1 and the true direction of the single-index parameter is

$$\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\| \;=\; (3, 1.5, 0, 0, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}}_{p\times 1}/\sqrt{15.25} \tag{4.25}$$

In the first screening stage, we apply the proposed robust independence ranking and screening (RIRS) to reduce the dimensionality down to the reduced model $\widehat{\mathcal{A}}$ with the size $2[n/\log n]$. Consequently, we note that all three true predictors have been screened into $\widehat{\mathcal{A}}$. In the second cleaning stage, the proposed penalized linear quantile regression is implemented on the dataset indexed by $\widehat{\mathcal{A}}$ to estimate the direction of the true single-index parameter and select variables via shrinking some coefficients to zeros. For the conditional quantile regression, we consider three different quantiles $\tau = 0.25, 0.50$ and 0.75, which correspond to the 1st quartile, the median and 3rd quartile of the response conditional on the predictors. An additional independent data set of size $10n$ is generated to select the tuning parameter λ by minimizing the estimated prediction error based on the quantile check loss function.

In the penalized linear quantile regression procedure, we consider three popular penalty functions: LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). We denote the final estimate by $\widehat{\boldsymbol{\beta}}_\tau = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$. Based on 100 repetitions, we evaluate the simulation performance in terms of the following criteria.

Size: The average number of non-zero estimated regression coefficients $\widehat{\beta}_j \neq 0$ for $1 \leq j \leq p$;

C: The average number of truly non-zero coefficients correctly estimated to be non-zero;

IC: The average number of truly zero coefficients incorrectly estimated to be non-zero;

AE: The average of absolute estimation error of $\widehat{\boldsymbol{\beta}}_\tau$, which is defined by

$$\sum_{j=1}^{p} \left| \frac{\widehat{\beta}_j}{\|\widehat{\boldsymbol{\beta}}_\tau\|} - \frac{\beta_{0j}}{\|\boldsymbol{\beta}_0\|} \right|,$$

where the true parameter $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p})^{\mathrm{T}}$ defined in (4.25).

Table 4.3 displays the simulation results for **Size, C, IC** and **AE**. In each column, the value represents the mean of 100 replicates with its sample standard deviation in the parentheses. For each scenario, the first three columns demonstrate that the LASSO is relatively conservative and tends to select larger models while the SCAD and MCP are consistent to select the true model. The relatively small values in the column labeled "**AE**" shows that the proposed penalized linear quantile regression procedure can produce the consistent estimator and support the theoretical findings in Theorem 4.3.2. The outstanding simulation results of Scenario 2 and 3 demonstrate that the proposed two-stage procedure is robust to the presence of heteroscedasticity and outliers in the response. In conclusion, the simulation results confirm the excellent finite sample performance of the penalized linear quantile regression approach.

In addition, we mentioned in section 4.3.1 that one may partition the full dataset into two separate parts for two estimation stages to avoid the technical assumption that $\widehat{\mathcal{A}}$ is independently of the data. To sufficiently study the proposed two-stage procedure, we randomly partition the data into the two equal parts, the screening set and cleaning set. In the screening stage, we apply the proposed RIRS on the screening set to obtain the reduced model $\widehat{\mathcal{A}}$ with the size $[n/\log n]$, which is naturally independent of the cleaning set. Consequently, we note that

Table 4.3. Simulation Results for Penalized Linear Quantile Regression with difference τ's (25%, 50% and 75%) and difference penalty functions (LASSO, SCAD, MCP), when the full dataset is used in both screening and cleaning stages.

Scenario 1: $\varepsilon \sim \mathcal{N}(0,1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	16.40(5.44)	3.00(0.00)	13.40(5.44)	0.30(0.16)
LASSO($\tau = 0.50$)	12.35(5.85)	3.00(0.00)	9.35(5.85)	0.59(0.31)
LASSO($\tau = 0.75$)	9.20(4.85)	2.99(0.10)	6.21(4.85)	0.84(0.45)
SCAD($\tau = 0.25$)	3.67(1.17)	3.00(0.00)	0.67(1.17)	0.07(0.04)
SCAD($\tau = 0.50$)	3.46(0.91)	3.00(0.00)	0.46(0.91)	0.20(0.12)
SCAD($\tau = 0.75$)	3.57(1.25)	2.91(0.29)	0.66(1.17)	0.43(0.23)
MCP($\tau = 0.25$)	3.44(0.95)	3.00(0.00)	0.44(0.95)	0.07(0.04)
MCP($\tau = 0.50$)	3.47(0.92)	3.00(0.00)	0.47(0.92)	0.21(0.12)
MCP($\tau = 0.75$)	3.60(1.20)	2.91(0.29)	0.69(1.14)	0.43(0.23)
Scenario 2: $\varepsilon \sim \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2)\mathcal{N}(0,1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	20.25(6.17)	3.00(0.00)	17.25(6.17)	0.33(0.14)
LASSO($\tau = 0.50$)	16.69(6.32)	3.00(0.00)	13.69(6.32)	0.66(0.29)
LASSO($\tau = 0.75$)	13.05(6.42)	2.97(0.17)	10.08(6.45)	1.11(0.60)
SCAD($\tau = 0.25$)	3.67(1.41)	3.00(0.00)	0.67(1.41)	0.06(0.05)
SCAD($\tau = 0.50$)	3.21(0.59)	3.00(0.00)	0.21(0.59)	0.15(0.10)
SCAD($\tau = 0.75$)	3.58(1.44)	2.90(0.30)	0.68(1.39)	0.41(0.31)
MCP($\tau = 0.25$)	3.48(1.28)	3.00(0.00)	0.48(1.28)	0.06(0.05)
MCP($\tau = 0.50$)	3.20(0.59)	3.00(0.00)	0.20(0.59)	0.15(0.10)
MCP($\tau = 0.75$)	3.47(1.23)	2.89(0.31)	0.58(1.18)	0.41(0.30)
Scenario 3: $\varepsilon \sim t(1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	20.02(4.68)	3.00(0.00)	17.02(4.68)	0.50(0.20)
LASSO($\tau = 0.50$)	15.43(6.30)	3.00(0.00)	12.43(6.30)	0.80(0.38)
LASSO($\tau = 0.75$)	11.85(6.08)	2.96(0.20)	8.89(6.06)	1.07(0.53)
SCAD($\tau = 0.25$)	3.46(0.82)	3.00(0.00)	0.46(0.82)	0.07(0.04)
SCAD($\tau = 0.50$)	3.29(0.62)	3.00(0.00)	0.29(0.62)	0.18(0.11)
SCAD($\tau = 0.75$)	3.51(1.69)	2.92(0.27)	0.59(1.65)	0.41(0.29)
MCP($\tau = 0.25$)	3.40(0.88)	3.00(0.00)	0.40(0.88)	0.07(0.04)
MCP($\tau = 0.50$)	3.26(0.56)	3.00(0.00)	0.26(0.56)	0.19(0.11)
MCP($\tau = 0.75$)	3.57(1.80)	2.92(0.27)	0.65(1.76)	0.42(0.30)

all three true predictors have been screened into $\widehat{\mathcal{A}}$. In the cleaning stage, the proposed penalized linear quantile regression is implemented on the cleaning set indexed by $\widehat{\mathcal{A}}$ to estimate the direction of the true single-index parameter and select variables via shrinking some coefficients to zeros. The similar simulation results are presented in Table 4.4, which demonstrates the excellent performance

of the proposed two-stage approach again.

Table 4.4. Simulation Results for Penalized Linear Quantile Regression with difference τ's (25%, 50% and 75%) and difference penalty functions (LASSO, SCAD, MCP), when the dataset is partitioned into the screening set and the cleaning set.

Scenario 1: $\varepsilon \sim \mathcal{N}(0,1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	17.26(4.93)	2.98(0.14)	14.28(4.92)	0.50(0.30)
LASSO($\tau = 0.50$)	13.88(6.46)	2.98(0.14)	10.90(6.44)	0.85(0.43)
LASSO($\tau = 0.75$)	9.46(5.39)	2.89(0.31)	6.57(5.32)	1.05(0.51)
SCAD($\tau = 0.25$)	3.53(1.36)	2.97(0.17)	0.56(1.34)	0.11(0.11)
SCAD($\tau = 0.50$)	3.45(1.19)	2.96(0.20)	0.49(1.16)	0.31(0.20)
SCAD($\tau = 0.75$)	3.18(1.67)	2.68(0.51)	0.50(1.55)	0.63(0.40)
MCP($\tau = 0.25$)	3.46(1.02)	2.98(0.14)	0.48(1.00)	0.11(0.11)
MCP($\tau = 0.50$)	3.45(1.15)	2.96(0.20)	0.49(1.18)	0.31(0.20)
MCP($\tau = 0.75$)	3.26(2.03)	2.70(0.50)	0.56(1.90)	0.63(0.40)
Scenario 2: $\varepsilon \sim \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/2)\mathcal{N}(0,1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	16.52(5.48)	2.92(0.27)	13.60(5.41)	0.45(0.33)
LASSO($\tau = 0.50$)	13.64(5.85)	2.92(0.27)	10.72(5.82)	0.73(0.37)
LASSO($\tau = 0.75$)	10.10(5.58)	2.84(0.37)	7.26(5.52)	1.18(0.63)
SCAD($\tau = 0.25$)	3.47(1.18)	2.91(0.32)	0.56(1.11)	0.15(0.22)
SCAD($\tau = 0.50$)	3.42(1.01)	2.90(0.30)	0.52(0.94)	0.29(0.21)
SCAD($\tau = 0.75$)	3.43(2.03)	2.63(0.56)	0.80(1.87)	0.65(0.43)
MCP($\tau = 0.25$)	3.29(0.90)	2.91(0.32)	0.38(0.83)	0.15(0.22)
MCP($\tau = 0.50$)	3.43(1.10)	2.90(0.30)	0.53(1.04)	0.29(0.21)
MCP($\tau = 0.75$)	3.49(2.09)	2.64(0.56)	0.85(1.92)	0.66(0.43)
Scenario 3: $\varepsilon \sim t(1)$				
Method	Size	C	IC	AE
LASSO($\tau = 0.25$)	15.53(4.58)	2.96(0.20)	12.57(4.58)	0.58(0.33)
LASSO($\tau = 0.50$)	13.41(6.20)	2.96(0.20)	10.45(6.18)	0.88(0.44)
LASSO($\tau = 0.75$)	9.54(5.66)	2.88(0.33)	6.66(5.61)	1.11(0.56)
SCAD($\tau = 0.25$)	3.56(1.04)	2.96(0.20)	0.60(1.03)	0.15(0.19)
SCAD($\tau = 0.50$)	3.54(1.30)	2.91(0.29)	0.63(1.24)	0.34(0.24)
SCAD($\tau = 0.75$)	3.64(2.94)	2.66(0.48)	0.98(2.80)	0.69(0.42)
MCP($\tau = 0.25$)	3.32(0.72)	2.96(0.20)	0.36(0.72)	0.15(0.18)
MCP($\tau = 0.50$)	3.42(1.08)	2.91(0.29)	0.51(1.02)	0.33(0.23)
MCP($\tau = 0.75$)	3.59(2.31)	2.67(0.47)	0.92(2.14)	0.69(0.42)

4.4.2 Real Data Analysis

In this subsection, we examine the proposed two-stage feature screening and variable selection procedure on the Cardiomyopathy microarray dataset. This dataset has been analyzed by Segal, Dahlquist and Conklin (2003) , Hall and Miller (2009) and Li, Zhong and Zhu (2012). The primary interest is to determine the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice. G protein-coupled receptors "comprise a large protein family of transmembrane receptors that sense molecules outside the cell and activate inside signal transduction pathways and, ultimately, cellular responses " (Wikipedia). In this data analysis, the Ro1 expression level is the response Y and genetic expression levels are considers as the predictors X_k's. The dimension of predictors is 6319, denoted by p, while the number of observed specimens is only 30, denoted by n. Thus, $p \gg n$ and this is a ultrahigh dimensional data analysis problem.

First, we conduct some exploratory data analysis on this microarray dataset. The histogram of Y on the left side of Figure 4.1 reveals that the distribution of the response is highly skewed. The right side of Figure 4.1 depicts the boxplot of Y, where the distance the plot whiskers extend out from the box is set to be 1 in order to detect the potential outliers. The boxplot shows that the distribution is positively skewed and there exist some potential outliers in the response. This motivates us to use the model (4.1) with the proposed two-stage procedure to conduct empirical analysis of this dataset.

In the screening stage, we implement the proposed RIRS on this microarray dataset to reduce the ultrahigh dimension to the reduced model $\widehat{\mathcal{A}}$ with the size $2[n/\log n] = 16$. The RIRS selects the two genes, labeled Msa.2877.0 and Msa.2134.0, in the top, which are same as the DC-SIS (Li, Zhong and Zhu, 2012). The gene, Msa.1166.0, identified by generalized correlation ranking (Hall and

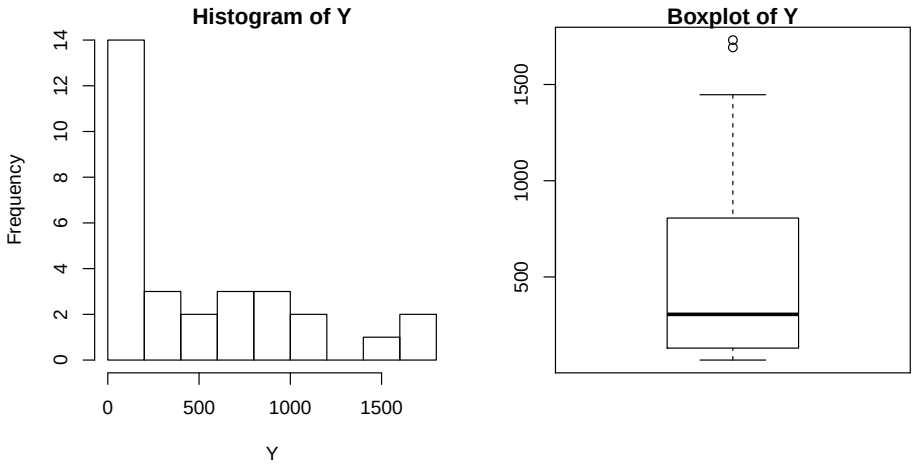

Figure 4.1. Exploratory Data Analysis: Histogram and Boxplot of Ro1.

Miller, 2009) is also ranked in the top by the RIRS. Figure 4.2 depicts scatter plots of Y against all 16 selected gene expression levels with cubic spline fitted curves, which indicate the marginal relationship between the response and each selected predictor.

In the cleaning stage, we implement the proposed penalized linear quantile regression on the reduced model $\widehat{\mathcal{A}}$ to estimate the direction of the single-index parameter and select important variables for different conditional quantiles of the response. Like the simulation studies, we consider three quantiles, $\tau = 0.25, 0.50$ and 0.75, and three different penalty functions, LASSO, SCAD and MCP. We use leave-one-out cross validation to select the tuning parameter for each method. After obtaining the estimated the single index, we propose to use the cubic splines to estimate the function $G(\cdot)$ in the model (4.1). As a benchmark, we also consider the model with 16 selected genes by the RIRS, denoted by NONE in Table 4.5. To compare the performances of different methods with different quantiles, we report the the number of nonzero coefficients selected by each method, denoted by "Size" in the first column of Table 4.5. In addition, to evaluate the goodness of fit for each model, we follow the idea of R^2 for the linear model and define the

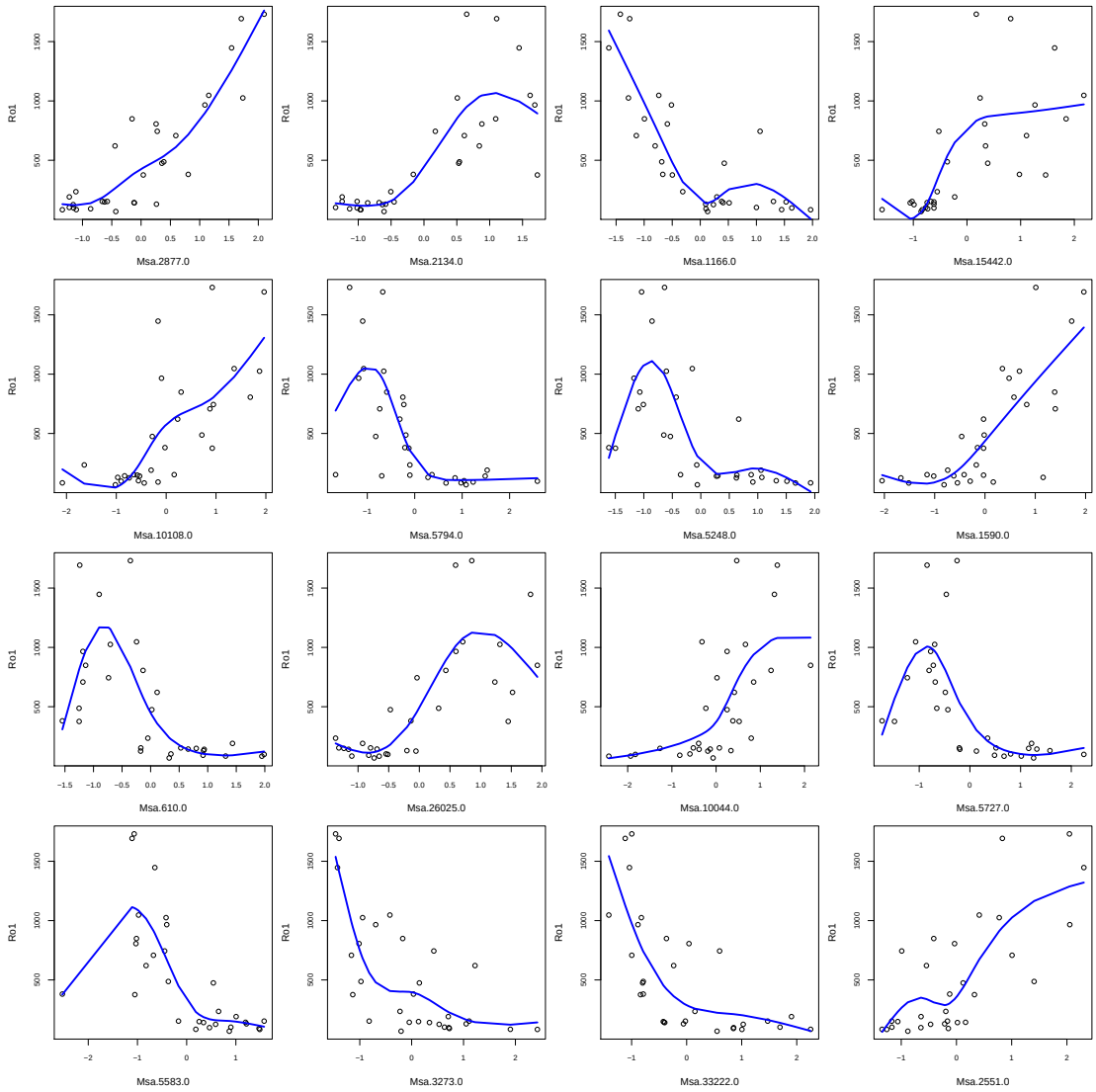

Figure 4.2. The scatter plots of Y versus top 16 genes expression levels identified by the proposed RIRS.

quantile-adjusted R^2 (Q-R^2) as follows

$$\text{Q-}R^2 = \left\{ 1 - \frac{\sum_{i=1}^{n} \rho_\tau(Y_i - \widehat{G}(X_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau))^2}{\sum_{i=1}^{n} \rho_\tau(Y_i - \widehat{Y}_\tau)^2} \right\} \times 100\%, \tag{4.26}$$

where $\rho_\tau(\cdot)$ the τth quantile check loss function, $\widehat{G}(\cdot)$ is the estimate of $G(\cdot)$ by cubic splines and thus $\widehat{G}(X_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau)$ is the fitted value of Y_i, and $\widehat{Y}_\tau$ is the sample τth

quantile of Y. The larger Q-R^2 is, the better the model fit is. For example, for $\tau = 0.75$, SCAD selected 5 predictors, which can explain 93.3% variance of the response in terms of the defined Q-R^2.

In addition, we also report the estimated coefficients of 16 gene expression levels in Table 4.6, and the scatter plots of Y versus the estimated single index with cubic splines fitted curves in Figure 4.3. It is interesting to note that the important predictors selected by each method are different at different quantiles. The top ranked genes, Msa.2877.0 and Msa.2134.0, are selected as the important predictors for the conditional median ($\tau = 0.5$) of Y; However, Msa.2134.0 is not important for the conditional third quartile ($\tau = 0.75$) of Y and Msa.2877.0 is only very slightly important for the conditional first quartile ($\tau = 0.25$) of Y (see the first two rows of Table 4.6).

Next, we conduct 50 random partitions to examine the prediction performance of each method. For each partition, we random select 90% of the data (27 observations) as the training set and the rest 10% (3 observations) as the test set. Leave-one-out cross validation is implemented on the training set to select the tuning parameter. The average of the model sizes selected by each method with its standard error across 50 partitions in the parenthesis is reported in the third column ("Ave Size") of Table 4.5. The column labeled by "PE" denotes the prediction errors based on the quantile check loss function and the corresponding standard errors in the parentheses. In conclusion, the penalized linear quantile regression improve both the model interpretability in terms of the model size and the model predictability in terms of the prediction errors.

Table 4.5. Empirical analysis of Cardiomyopathy microarray dataset.

	All Data		Partitioned Data	
Method	Size	Q-R^2	Ave Size	PE
NONE($\tau = 0.25$)	16	77.9	16(0)	0.63(0.36)
NONE($\tau = 0.50$)	16	91.9	16(0)	0.66(0.37)
NONE($\tau = 0.75$)	16	93.1	16(0)	0.56(0.39)
LASSO($\tau = 0.25$)	9	60.9	8.64(2.25)	0.48(0.17)
LASSO($\tau = 0.50$)	11	89.1	10.28(1.92)	0.50(0.21)
LASSO($\tau = 0.75$)	7	94.3	6.42(1.64)	0.42(0.17)
SCAD($\tau = 0.25$)	3	55.5	7.90(2.97)	0.49(0.22)
SCAD($\tau = 0.50$)	10	91.0	8.28(2.82)	0.48(0.32)
SCAD($\tau = 0.75$)	5	93.3	4.50(2.29)	0.45(0.21)
MCP($\tau = 0.25$)	3	53.7	7.88(3.59)	0.53(0.11)
MCP($\tau = 0.50$)	10	91.1	8.82(3.09)	0.55(0.26)
MCP($\tau = 0.75$)	4	93.9	4.66(2.43)	0.47(0.25)

Table 4.6. Estimated coefficients of 16 gene expression levels for Cardiomyopathy microarray dataset.

Quantile	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
Method	LASSO	SCAD	MCP	LASSO	SCAD	MCP	LASSO	SCAD	MCP
Msa.2877.0	0	0.032	0.033	0.370	0.549	0.557	0.493	0.729	0.698
Msa.2134.0	0.308	0.463	0.463	0.214	0.366	0.361	0	0	0
Msa.1166.0	0	0	0	0	0	0	-0.006	-0.069	0
Msa.15442.0	0.192	0	0	0.139	0.078	0.081	0.010	0	0
Msa.10108.0	0.243	0.260	0.260	0.100	0.047	0.041	0.027	0	0
Msa.5794.0	-0.054	0	0	-0.031	-0.023	-0.030	0	0	0
Msa.5248.0	0.077	0	0	0.076	0.193	0.203	0	0	0
Msa.1590.0	0	0	0	0.085	0.078	0.086	0.166	0.041	0.140
Msa.610.0	0.021	0	0	0	0	0	0	0	0
Msa.26025.0	0	0	0	0.098	0.036	0.039	0.136	0.098	0.104
Msa.10044.0	0.088	0	0	0.037	0.065	0.064	0	0.073	0.010
Msa.5727.0	0	0	0	0	0	0	0	0	0
Msa.5583.0	0	0	0	0	0	0	0	0	0
Msa.3273.0	0	0	0	0	0	0	0	0	0
Msa.33222.0	0.053	0	0	0.167	0.260	0.266	0	0	0
Msa.2551.0	0.033	0	0	0.050	0	0	0.078	0	0

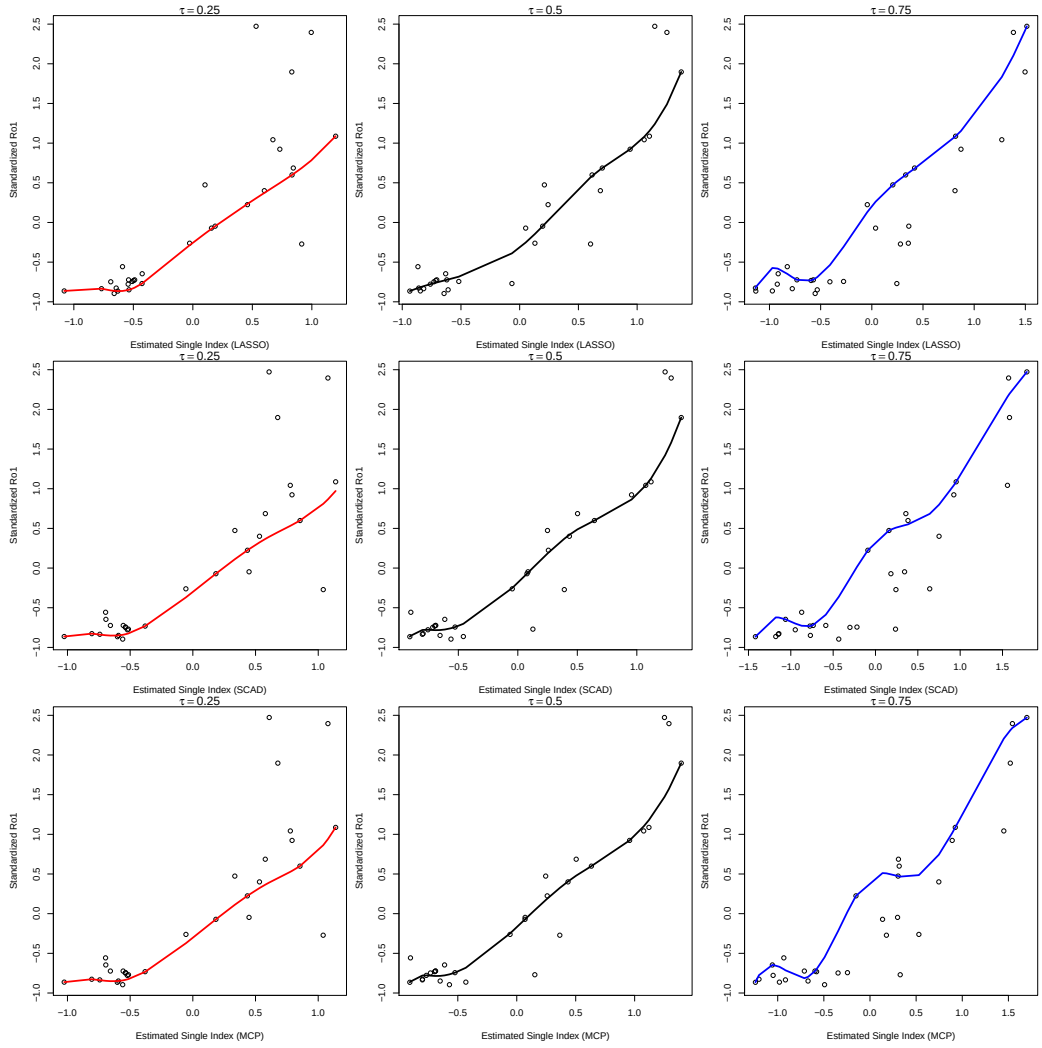

Figure 4.3. The scatter plots of Y versus the estimated single index with cubic splines fitted curves for LASSO, SCAD and MCP at three different quantiles ($\tau = 0.25, 0.50$ and 0.75).

4.5 Theoretical Proofs

4.5.1 Preliminary Lemmas

Lemma 4.5.1. (Bernstein's Inequality) *Let $X_1, \ldots, X_n$ be independent zero-mean random variables such that $E|X_i|^m \leq m! M^{m-2} \nu_i/2$, for every $m \geq 2$, all i*

and some constants M and ν_i. Then

$$Pr\left(\left|\sum_{i=1}^{n} X_i\right| > \varepsilon\right) \leq 2\exp\left\{-\frac{\varepsilon^2}{2(\nu + M\varepsilon)}\right\}, \tag{4.27}$$

for any $\varepsilon > 0$ and $\nu \geq \nu_1 + \ldots + \nu_n$.

For details, See Lemma 2.2.11, van der Vaart and Wellner (1996).

Lemma 4.5.2. (Dvoretzky-Kiefer-Wolfowitz Inequality)
Let $Y_1, Y_2, \ldots, Y_n$ be real-valued independent and identically distributed random variables with distribution function $F(\cdot)$. Let $F_n(\cdot)$ denote the associated empirical distribution function, then the following inequality holds for any $\varepsilon > 0$,

$$Pr\left(\sup_{y\in\mathbb{R}} |F_n(y) - F(y)| > \varepsilon\right) \leq 2\exp\{-2n\varepsilon^2\}. \tag{4.28}$$

For details, see Dvoretzky, Kiefer and Wolfowitz (1956).

4.5.2 Proof of Theorem 4.2.1

We recall that $\omega_k = E\{X_k F(Y)\}$, which, by the law of iterated expectations, is equal to $E\left[E\left\{X_k F(Y) \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}\right\}\right]$. In addition, we observe that $\mathbf{x}$ is independent of Y when $\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}$ is given. These facts, together with the linearity condition (A4.2) and the law of iterated expectations, indicate that

$$\begin{aligned}
\omega_k &= E\left[E\left(X_k \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}\right) E\left\{F(Y) \mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}\right\}\right] \\
&= \mathrm{cov}(X_k, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_{\mathcal{A}} \left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_{\mathcal{A}}\right\}^{-1} E\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}} F(Y)\right\}.
\end{aligned}$$

Consequently, it follows that

$$\begin{aligned}\min_{k\in\mathcal{A}}\omega_k^2-\max_{k\in\mathcal{I}}\omega_k^2 &= \left\{\min_{k\in\mathcal{A}}\mathrm{cov}^2(X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})-\max_{k\in\mathcal{I}}\mathrm{cov}^2(X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}})\right\}\\ &\quad\times\frac{E\left\{\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}F(Y)\right\}^2}{\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\boldsymbol{\beta}_{\mathcal{A}}\right\}^2},\end{aligned}$$

where must be positive invoking condition (A4.1). The proof of Theorem 3.3.4 is completed. □

4.5.3 Proof of Theorem 4.2.2

At first, we prove the first part (4.9) of the Theorem 4.2.2. Note that

$$\begin{aligned}\widehat{\omega}_k-\omega_k &= \frac{1}{n}\sum_{i=1}^n X_{ik}F_n(Y_i)-E\{X_kF(Y)\}\\ &= \frac{1}{n}\sum_{i=1}^n X_{ik}[F_n(Y_i)-F(Y_i)]+\frac{1}{n}\sum_{i=1}^n[X_{ik}F(Y_i)-E\{X_kF(Y)\}]\end{aligned}$$

Therefore, for any $\epsilon>0$

$$\begin{aligned}&\Pr(\max_{1\le k\le p_n}|\widehat{\omega}_k-\omega_k|\ge\epsilon)\\ =\;&\Pr\left(\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^n X_{ik}[F_n(Y_i)-F(Y_i)]+\frac{1}{n}\sum_{i=1}^n[X_{ik}F(Y_i)-E\{X_kF(Y)\}]\right|\ge\epsilon\right)\\ \le\;&\Pr\left(\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^n X_{ik}[F_n(Y_i)-F(Y_i)]\right|\ge\epsilon/2\right)\\ +\;&\Pr\left(\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^n[X_{ik}F(Y_i)-E\{X_kF(Y)\}]\right|\ge\epsilon/2\right)\\ =:\;&A+B.\end{aligned}\tag{4.29}$$

Then, we consider the part (A) first,

$$\begin{aligned}
& \Pr\left(\max_{1\le k\le p_n}\left|\frac{1}{n}\sum_{i=1}^{n}X_{ik}[F_n(Y_i)-F(Y_i)]\right|\ge\epsilon/2\right)\\
\le\ & \Pr\left(\max_{1\le k\le p_n}\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}X_{ik}^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}[F_n(Y_i)-F(Y_i)]^2\right\}}\ge\epsilon/2\right)\\
\le\ & \Pr\left(\frac{1}{n}\sum_{i=1}^{n}[F_n(Y_i)-F(Y_i)]^2\ge\epsilon^2/4\right)\\
\le\ & \Pr\left(\max_{0\le i\le n}|F_n(Y_i)-F(Y_i)|\ge\epsilon/2\right)\\
\le\ & \Pr\left(\sup_{y\in\mathbb{R}}|F_n(y)-F(y)|\ge\epsilon/2\right)\\
\le\ & 2\exp(-n\epsilon^2/2), \qquad (4.30)
\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz inequality and the fact that $n^{-1}\sum_{i=1}^{n}X_{ik}^2=1$, and the last inequality follows by Lemma 4.5.2 (Dvoretzky, Kiefer and Wolfowitz, 1956).

For the second part (B), let $T_{ik}=X_{ik}F(Y_i)-E\{X_kF(Y)\}$, then $E(T_{ik})=0$ and T_{ik} are identical and independently distributed for all $i=1,\ldots,n$.

For every $m\ge 2$ and any $0<t\le t_0$,

$$\begin{aligned}
E\left|T_{ik}\right|^m &\le E\left|X_{ik}F(Y_i)-E\{X_kF(Y)\}\right|^m\\
&\le m!t^{-m}E\exp\left\{t\left|X_{ik}F(Y_i)-E\{X_kF(Y)\}\right|\right\}\\
&\le m!t^{-m}C\le\frac{1}{2}m!\left(\frac{1}{t}\right)^{m-2}\left(\frac{2C}{t^2}\right),
\end{aligned}$$

where the second inequality follows the fact $x^m\le m!\exp(x)$ for any $x>0$ and the third inequality follows condition (A4.3) and Jensen's inequality. By Bernstein's

inequality (Lemma 4.31), for any $\gamma > 0$,

$$\Pr\left(\left|\sum_{i=1}^{n} T_{ik}\right| > \gamma\right) \;\leq\; 2\exp\left(-\frac{1}{2}\frac{\gamma^2}{2nC/t^2 + \gamma/t}\right).$$

Then,

$$\begin{aligned}
&\Pr\left(\max_{1\leq k\leq p_n}\left|\frac{1}{n}\sum_{i=1}^{n}[X_{ik}F(Y_i) - E\{X_kF(Y)\}]\right| \geq \epsilon/2\right)\\
\leq\;& p_n \max_{1\leq k\leq p_n}\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}[X_{ik}F(Y_i) - E\{X_kF(Y)\}]\right| \geq \epsilon/2\right)\\
=\;& p_n \max_{1\leq k\leq p_n}\Pr\left(\left|\sum_{i=1}^{n} T_{ik}\right| \geq n\epsilon/2\right)\\
\leq\;& 2p_n\exp\left(-\frac{1}{4}\frac{n\epsilon^2}{4C/t^2 + \epsilon/t}\right)\\
\leq\;& 2p_n\exp\left(-C'n\epsilon^2\right), \qquad (4.31)
\end{aligned}$$

where the last inequality holds for $\epsilon > 0$ sufficiently small and some positive constant C'. Therefore, (4.30) and (4.31) together entails that

$$\begin{aligned}
\Pr\left(\max_{1\leq k\leq p_n}|\widehat{\omega}_k - \omega_k| \geq \epsilon\right) \;&\leq\; 2\exp\{-n\epsilon^2/2\} + 2p_n\exp\left(-C'n\epsilon^2\right)\\
&=\; 2(p_n+1)\exp\left(-a_1n\epsilon^2\right) \to 0, \qquad (4.32)
\end{aligned}$$

with some positive constant a_1, as $n \to \infty$.

Next, we show the second part of Theorem 4.2.2. Let $\delta = \min_{k\in\mathcal{A}}\omega_k^2 - \max_{k\in\mathcal{I}}\omega_k^2$, then by Theorem 4.2.1, $\delta > 0$ holds uniformly for p_n under conditions (A4.1)-(A4.3). Therefore,

$$\begin{aligned}
&\Pr\left(\max_{k\in\mathcal{I}}\widehat{\omega}_k^2 \geq \min_{k\in\mathcal{A}}\widehat{\omega}_k^2\right)\\
=\;&\Pr\left(\max_{k\in\mathcal{I}}\widehat{\omega}_k^2 - \max_{k\in\mathcal{I}}\omega_k^2 \geq \min_{k\in\mathcal{A}}\widehat{\omega}_k^2 - \min_{k\in\mathcal{A}}\omega_k^2 + \delta\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \Pr\left(\left|\max_{k\in\mathcal{I}}\widehat{\omega}_k^2-\max_{k\in\mathcal{I}}\omega_k^2\right|+\left|\min_{k\in\mathcal{A}}\widehat{\omega}_k^2-\min_{k\in\mathcal{A}}\omega_k^2\right|\geq\delta\right)\\
&\leq \Pr\left(\left|\max_{k\in\mathcal{I}}\widehat{\omega}_k^2-\max_{k\in\mathcal{I}}\omega_k^2\right|\geq\delta/2\right)+\Pr\left(\left|\min_{k\in\mathcal{A}}\widehat{\omega}_k^2-\min_{k\in\mathcal{A}}\omega_k^2\right|\geq\delta/2\right)\\
&\leq \Pr\left(\max_{k\in\mathcal{I}}\left|\widehat{\omega}_k^2-\omega_k^2\right|\geq\delta/2\right)+\Pr\left(\max_{k\in\mathcal{A}}\left|\widehat{\omega}_k^2-\omega_k^2\right|\geq\delta/2\right) \qquad (4.33)
\end{aligned}
$$

We observe that $\widehat{\omega}_k^2 \leq \{n^{-1}\sum_{i=1}^n X_{ik}^2\}\{n^{-1}\sum_{i=1}^n F_n^2(Y_i)\} \leq 1$ and $\omega_k^2 \leq EX_k^2 = 1$. Then (4.33) entails that

$$
\begin{aligned}
\Pr\Big(\max_{k\in\mathcal{I}}\widehat{\omega}_k^2\geq\min_{k\in\mathcal{A}}\widehat{\omega}_k^2\Big) &\leq \Pr\Big(\max_{k\in\mathcal{I}}|\widehat{\omega}_k-\omega_k|\geq\delta/4\Big)+\Pr\Big(\max_{k\in\mathcal{A}}|\widehat{\omega}_k-\omega_k|\geq\delta/4\Big)\\
&\leq 2(p_n+1)\exp\left(-a_2n\delta^2\right)\to 0, \qquad (4.34)
\end{aligned}
$$

with some constant $a_2 > 0$, as $n \to \infty$. The last inequality follows the same arguments for proving (4.32). This completes the proof of Theorem 4.2.2. □

4.5.4 Proof of Lemma 4.2.3

According to the proof of Theorem 4.2.1, we have that

$$
\begin{aligned}
\min_{k\in\mathcal{A}}\omega_k^2 &= \left\{\min_{k\in\mathcal{A}}\mathrm{cov}^2(X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta})\right\}\left\{\mathrm{cov}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})E[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}F(Y)]\right\}^2\\
&\geq (c_1n^{-\kappa})\left\{\mathrm{cov}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})E[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}F(Y)]\right\}^2=c_2n^{-\kappa}, \qquad (4.35)
\end{aligned}
$$

where we denote $c_2 = \{\mathrm{cov}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}})E[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}F(Y)]\}^2 c_1 > 0$. □

4.5.5 Proof of Theorem 4.2.4

We follow the uniform convergence rate of $\widehat{\omega}_k$ in (4.9) and set $\epsilon = c_4 n^{-\kappa}$ with any $c_4 > 0$. Consequently, for n sufficiently large, there exists a constant $a_1' > 0$,

$$\Pr\left\{\max_{1\le k\le p_n} |\widehat{\omega}_k - \omega_k| \ge c_3 n^{-\kappa}\right\} \le 2(p_n+1)\exp(-a_1' n^{1-2\kappa}). \tag{4.36}$$

Let $\mathcal{E}_n = \left\{\max_{k\in\mathcal{A}} |\widehat{\omega}_k^2 - \omega_k^2| \le c_3 n^{-\kappa}\right\}$, where $c_3 \le c_2/2$. On the event $\mathcal{E}_n$, by Lemma 4.2.3, we have that

$$c_3 n^{-\kappa} \ge \max_{k\in\mathcal{A}} \omega_k^2 - \min_{k\in\mathcal{A}} \widehat{\omega}_k^2 \ge c_2 n^{-\kappa} - \min_{k\in\mathcal{A}} \widehat{\omega}_k^2,$$

then $\min_{k\in\mathcal{A}} \widehat{\omega}_k^2 \ge c_3 n^{-\kappa}$, which entails that $\mathcal{A} \subseteq \widehat{\mathcal{A}}_{\gamma_n}$ by the choice of $\gamma_n = c_3 n^{-\kappa}$. Therefore,

$$\begin{aligned}\Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}_{\gamma_n}) &\ge \Pr(\mathcal{E}_n) = 1 - \Pr(\mathcal{E}_n^c) = 1 - \Pr\left\{\max_{k\in\mathcal{A}} \left|\widehat{\omega}_k^2 - \omega_k^2\right| \ge c_3 n^{-\kappa}\right\} \\ &\ge 1 - \Pr\left\{\max_{k\in\mathcal{A}} |\widehat{\omega}_k - \omega_k| \ge c_3 n^{-\kappa}/2\right\} \ge 1 - 2(q_n+1)\exp\{-a_3 n^{1-2\kappa}\},\end{aligned}$$

where q_n is the cardinality of $\mathcal{A}$, the second inequality follows that that $\widehat{\omega}_k^2 \le \{n^{-1}\sum_{i=1}^n X_{ik}^2\}\{n^{-1}\sum_{i=1}^n F_n^2(Y_i)\} \le 1$ and $\omega_k^2 \le EX_k^2 = 1$, the last inequality holds for n large enough and some positive constant $a_3 > 0$, which follows the same arguments for proving (4.36). This completes the proof of Theorem 4.2.4. □

4.5.6 Proof of Lemma 4.3.1

First, we need an additional lemma in the following.

Lemma 4.5.3. *According to (4.17), we have* $E\{I(Y - \mathbf{x}_{\mathcal{A}}^T \boldsymbol{\beta}_{\tau 1}^o \le u_\tau^o) \mid \mathbf{x}_{\mathcal{A}}\} = \tau$.

Proof: Let ξ_τ be the τth quantile of $Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o$ conditional on $\mathbf{x}_{\mathcal{A}}$. Then, $E\left\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le \xi_\tau) \mid \mathbf{x}_{\mathcal{A}}\right\} = \tau$. It is enough to show, $\mathcal{L}_\tau(\xi_\tau, \boldsymbol{\beta}_{\tau 1}^o) \le \mathcal{L}_\tau(u, \boldsymbol{\beta}_{\tau 1}^o)$ holds for any u. To be specific,

$$
\begin{aligned}
& \mathcal{L}_\tau(u, \boldsymbol{\beta}_{\tau 1}^o) - \mathcal{L}_\tau(\xi_\tau, \boldsymbol{\beta}_{\tau 1}^o) \\
= \;& E\{\rho_\tau(Y - u - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o)\} - E\{\rho_\tau(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o)\} \\
= \;& E\left\{(u - \xi_\tau)[I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le 0) - \tau]\right\} \\
+ \;& E\left\{\int_0^{u-\xi_\tau} [I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le t) - I(Y - \xi_\tau - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le 0)]\, dt\right\} \\
\ge \;& 0,
\end{aligned}
$$

where the second equality follows the identity (Knight, 1998). In the second equality, the first term is zero based on the definition of ξ_τ and the second term is always nonnegative. Therefore, $E\left\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le u_\tau^o) \mid \mathbf{x}_{\mathcal{A}}\right\} = \tau$. □

To prove Lemma 4.3.1, it suffices to show that for any fixe $\eta > 0$, there exists two constants Δ_1 and Δ_2 such that for all sufficiently large n,

$$
\Pr\left\{\inf_{\substack{\|\gamma\|=\Delta_1 \\ |u|=\Delta_2}} \mathcal{L}_{\tau n}(u_\tau^o + n^{-1/2}q_n^{1/2}u, \boldsymbol{\beta}_{\tau 1}^o + n^{-1/2}q_n^{1/2}\gamma) > \mathcal{L}_{\tau n}(u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o)\right\} \ge 1 - \eta.
$$

We define that

$$
\begin{aligned}
G_n(u, \gamma) \;=:\; & nq_n^{-1}\left\{\mathcal{L}_{\tau n}(u_\tau^o + n^{-1/2}q_n^{1/2}u, \boldsymbol{\beta}_{\tau 1}^o + n^{-1/2}q_n^{1/2}\gamma) - \mathcal{L}_{\tau n}(u_\tau^o, \boldsymbol{\beta}_{\tau 1}^o)\right\} \\
=\; & q_n^{-1}\sum_{i=1}^n n^{-1/2}q_n^{1/2}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)\left\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le u_\tau^o) - \tau\right\} \\
+\; & q_n^{-1}\sum_{i=1}^n \int_0^{n^{-1/2}q_n^{1/2}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)} \left\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le u_\tau^o + s) - I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^o \le u_\tau^o)\right\} ds \\
=:\; & I_{n1} + I_{n2},
\end{aligned}
$$

where the second equality follows from Knight (1998)'s identity. Note that

$E\left\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o} \le u_{\tau}^{o}) \mid \mathbf{x}_{\mathcal{A}}\right\} = \tau$ by Lemma (4.5.3) and hence $E(I_{n1}) = 0$. In addition,

$$\begin{aligned}
\mathrm{var}(I_{n1}) &= E\{\mathrm{var}(I_{n1} \mid \mathbf{x}_{\mathcal{A}})\} + \mathrm{var}\{E(I_{n1} \mid \mathbf{x}_{\mathcal{A}})\} \\
&= \tau(1-\tau)q_n^{-1}E\Big\{n^{-1}\sum_{i=1}^{n}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)^2\Big\} \\
&\le 2\tau(1-\tau)q_n^{-1}[u^2 + \lambda_{\max}\left\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\right\}\|\gamma\|^2] \\
&\le Cq_n^{-1}(\Delta_1^2 + \Delta_2^2),
\end{aligned}$$

which together with condition (A4.5) implies that $I_{n1} = O_p\big(q_n^{-1/2}\big)\sqrt{\Delta_1^2 + \Delta_2^2}$.

Next we evaluate I_{n2}. We denote by $F(\cdot \mid \mathbf{x}_{\mathcal{A}})$ and $f(\cdot \mid \mathbf{x}_{\mathcal{A}})$ the respective conditional distribution and density of $Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}$ given $\mathbf{x}_{\mathcal{A}}$.

$$\begin{aligned}
E\left(I_{n2}\right) &= q_n^{-1}E\left[\sum_{i=1}^{n}\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)}\left\{F(u_{\tau}^{o} + s \mid \mathbf{x}_{i,\mathcal{A}}) - F(u_{\tau}^{o} \mid \mathbf{x}_{i,\mathcal{A}})\right\}ds\right] \\
&= q_n^{-1}E\left[\sum_{i=1}^{n}\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)} f(u_{\tau}^{o} + s' \mid \mathbf{x}_{i,\mathcal{A}})sds\right] \\
&\ge Cq_n^{-1}E\left[\sum_{i=1}^{n}\left\{n^{-1/2}q_n^{1/2}(u + \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)\right\}^2\right] \\
&= CE(u + \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma)^2 \ge C\left(1 + \lambda_{\min}\left\{E(\mathbf{x}_{\mathcal{A}}\mathbf{x}_{\mathcal{A}}^{\mathrm{T}})\right\}\right)(u^2 + \|\gamma\|^2) \\
&\ge C(\Delta_1^2 + \Delta_2^2),
\end{aligned}$$

where the first inequality follows from condition (A6) and the last inequality follows from condition (A4.5). Therefore, $E\left(I_{n2}\right) = O(1)(\Delta_1^2 + \Delta_2^2)$. Next we consider the variance of I_{n2},

$$\begin{aligned}
\mathrm{var}\left(I_{n2}\right) &\le nq_n^{-2}E\left[\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma)}\left\{I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o} \le u_{\tau}^{o} + s) - I(Y - \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o} \le u_{\tau}^{o})\right\}ds\right]^2 \\
&\le nq_n^{-2}n^{-1/2}q_n^{1/2}E\left[|u + \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma|\int_0^{n^{-1/2}q_n^{1/2}(u+\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\gamma)}\left\{F(u_{\tau}^{o} + s|\mathbf{x}_{\mathcal{A}}) - F(u_{\tau}^{o}|\mathbf{x}_{\mathcal{A}})\right\}ds\right]
\end{aligned}$$

$$\leq \quad nq_n^{-2}n^{-3/2}q_n^{3/2}E\big(\,|u+\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma|^3\,\big).$$

Next we study the order of $E\big(\,|\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma|^3\,\big)$. Condition (A4.3) ensures that

$$E\big(\,|\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\gamma|^3\,\big) \leq E\big(\|\mathbf{x}_{\mathcal{A}}\|^3\big)\|\gamma\|^3 \leq \big\{E\big(\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}}\big)^3\big\}^{1/2}\|\gamma\|^3 = O(q_n^{3/2})\|\gamma\|^3,$$

and hence

$$\mathrm{var}\,(I_{n2}) = O\big(nq_n^{-2}n^{-3/2}q_n^{3/2}q_n^{3/2}\big)(|u|^3+\|\gamma\|^3) = O(q_n n^{-1/2})(|u|^3+\|\gamma\|^3)$$

which converges to zero as $q_n^2/n \to 0$. This indicates that $|I_{n2} - E(I_{n2})| = o_p(1)$ by Chebyshev's inequality. Furthermore, since I_{n2} is always nonnegative,

$$I_{n2} = E(I_{n2}) + o_p(1) \geq C(\Delta_1^2+\Delta_2^2) + o_p(1).$$

For sufficiently large Δ_1 and Δ_2, I_{n2} will dominate I_{n1} asymptotically as $n \to \infty$. Therefore, for any fixed $\eta > 0$, there exists two constants Δ_1 and Δ_2 such that for all sufficiently large n, we have $G_n(u,\gamma) > 0$ with probability at least $1-\eta$.

□

4.5.7 Proof of Theorem 4.3.2

The proof of Theorem 4.3.2 is parallel to the proof of Theorem 2.4 in Wang, Wu and Li (2012). With slightly notational abuse, we write $\mathbf{x}_{\mathcal{A}} = (1,\mathbf{x}_{\mathcal{A}})^{\mathrm{T}}$, $\mathbf{x}_{\widehat{\mathcal{A}}} = (1,\mathbf{x}_{\widehat{\mathcal{A}}})^{\mathrm{T}}$ $\boldsymbol{\beta}_\tau^o = (u_\tau^o, \boldsymbol{\beta}_\tau^{o\mathrm{T}})^{\mathrm{T}}$, $\widehat{\boldsymbol{\beta}}_\tau = (\widehat{u}_\tau, \widehat{\boldsymbol{\beta}}_\tau^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_\tau^o = (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_\tau^{o\mathrm{T}})^{\mathrm{T}}$. where $\widehat{\boldsymbol{\beta}}_\tau$ denotes the penalized linear quantile estimator and $\widehat{\boldsymbol{\beta}}_\tau^o = (\widehat{\boldsymbol{\beta}}_{\tau 1}^{o\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ is the oracle estimator. Accordingly, we write $\boldsymbol{\beta}_{\tau 1}^o = (u_\tau^o, \boldsymbol{\beta}_{\tau 1}^{o\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_{\tau 1}^o = (\widehat{u}_\tau^o, \widehat{\boldsymbol{\beta}}_{\tau 1}^{o\mathrm{T}})^{\mathrm{T}}$.

The objective function (4.15) with the SCAD penalty can be written in terms

of $\boldsymbol{\beta} = (u, \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}$,

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{j=1}^{d_n} p_\lambda(|b_j|),$$

which is the difference of two convex functions in $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}),$$

where $g(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{d_n}|\beta_j|$, and $h(\boldsymbol{\beta}) = \sum_{j=1}^{d_n} H_\lambda(\beta_j)$, with

$$H_\lambda(\beta_j) = \begin{cases} 0, & 0 \le |\beta_j| < \lambda; \\ (\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)/\{2(a-1)\}, & \lambda \le |\beta_j| \le a\lambda; \\ \lambda|\beta_j| - (a+1)\lambda^2/2, & |\beta_j| > a\lambda. \end{cases}$$

Thus, the subdifferential of $h(\boldsymbol{\beta})$ at any $\boldsymbol{\beta}$ is

$$\partial h(\boldsymbol{\beta}) = \left\{ \boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{d_n})^{\mathrm{T}} \in \mathbb{R}^{d_n+1} : \mu_0 = 0, \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, j = 1, 2, \ldots, d_n \right\}.$$

The subdifferential of $g(\boldsymbol{\beta})$ at any $\boldsymbol{\beta}$ is

$$\begin{aligned} \partial g(\boldsymbol{\beta}) \;=\; & \left\{ \boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_{d_n})^{\mathrm{T}} \in \mathbb{R}^{d_n+1} : \xi_j = (1-\tau)n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta} < 0) \right. \\ & \left. -\tau n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta} > 0) - n^{-1}\sum_{i=1}^{n} X_{ij} v_i + \lambda l_j \right\}, \end{aligned}$$

where $v_i = 0$ if $Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise; $l_0 = 0$; $l_j = \mathrm{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise, for $1 \le j \le d_n$.

Let $s(\widehat{\boldsymbol{\beta}}) = \left\{ s_0(\widehat{\boldsymbol{\beta}}), s_1(\widehat{\boldsymbol{\beta}}), \ldots, s_{d_n}(\widehat{\boldsymbol{\beta}}) \right\}^{\mathrm{T}}$ be the set of the subgradient functions

for the unpenalized quantile regression, where

$$\begin{aligned} s_j(\boldsymbol{\beta}) &= (1-\tau)n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta} < 0) \\ &\quad -\tau n^{-1}\sum_{i=1}^{n} X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\boldsymbol{\beta} > 0) - n^{-1}\sum_{i=1}^{n} X_{ij} v_i, \end{aligned}$$

where $v_i = 0$ if $Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \neq 0$ and $v_i \in [\tau - 1, \tau]$ otherwise.

We first provide Lemmas 4.5.4, 4.5.5 and 4.5.6 to facilitate the proof of Theorem 4.3.2. Tao and An (1997) proposed the numerical algorithm based on the convex difference representation and we present the result in the following Lemma 4.5.4. The following Lemmas 4.5.5 and 4.5.6 characterize the properties of the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o$ and the associated subgradient functions $s(\widehat{\boldsymbol{\beta}}_\tau^o)$ corresponding to the active and inactive variables, respectively.

Lemma 4.5.4. (Difference Convex Program) *$g(\mathbf{x})$ and $h(\mathbf{x})$ are two convex functions. Let $\mathbf{x}^*$ be a point that admits a neighborhood U such that $\partial h(\mathbf{x}) \cap \partial g(\mathbf{x}^*) \neq \emptyset$, $\forall \mathbf{x} \in U \cap dom(g)$. Then $\mathbf{x}^*$ is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$.*

Lemma 4.5.5. *Assume the conditions (A4.7)-(A4.8) holds and $\lambda = o(n^{-(1-c_2)/2})$. For the oracle estimator $\widehat{\boldsymbol{\beta}}_\tau^o$, there exist v_i^* which satisfies $v_i^* = 0$ if $Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{T}\widehat{\boldsymbol{\beta}}_\tau^o \neq 0$ and $v_i^* \in [\tau - 1, \tau]$ otherwise, such that, with probability approaching one, we have*

$$s_j(\widehat{\boldsymbol{\beta}}_\tau^o) = 0, j = 0, 1, \ldots, q_n, \quad \textit{and} \quad |\widehat{\beta}_j^{\,o}| \geq (a + 1/2)\lambda, j = 1, \ldots, q_n.$$

Proof of Lemma 4.5.5: The unpenalized quantile loss objective function is convex. By the convex optimization theory,

$$\mathbf{0} \in \partial \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_\tau^o).$$

Therefore, there exists v_i^* such that $s_j(\widehat{\boldsymbol{\beta}}_\tau^o) = 0$ with $v_i = v_i^*$ for $j = 0, 1, \ldots, q_n$. On the other hand,

$$\min_{1\le j\le q_n} |\widehat{\beta}_j^{\,o}| \ge \min_{1\le j\le q_n} |\beta_{\tau,j}^o| - \max_{1\le j\le q_n} |\widehat{\beta}_j^{\,o} - \beta_{\tau,j}^o|.$$

Condition (A4.8) requires that $\min_{1\le j\le q_n} |\beta_{\tau,j}^o| \ge Cn^{-(1-c_2)/2}$. In addition, $\max_{1\le j\le q_n} |\widehat{\beta}_j^{\,o} - \beta_{\tau,j}^o| \le \|\widehat{\boldsymbol{\beta}}_\tau^o - \boldsymbol{\beta}_{\tau 1}^o\| = O_p(\sqrt{q_n/n}) = O_p(n^{-(1-c_1)/2}) = o_p(n^{-(1-c_2)/2})$. Therefore, $\min_{1\le j\le q_n} |\widehat{\beta}_j^{\,o}| \ge Cn^{-(1-c_2)/2} - o_p(n^{-(1-c_2)/2})$, where c_1 and c_2 are defined in conditions (A4.7) and (A4.8) respectively. For $\lambda = o(n^{-(1-c_2)/2})$, we have that, with probability approaching one, $|\widehat{\beta}_j^{\,o}| \ge (a + 1/2)\lambda, j = 1, \ldots, q_n$, which completes the proof.

□

Lemma 4.5.6. *Assume the conditions (A4.3) and (A4.5)-(A4.8) hold and* $\lambda = o(n^{-(1-c_2)/2})$, $d_n = O\big(n\big)$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}_\tau^o$ *and the* $s_j(\widehat{\boldsymbol{\beta}}_\tau^o)$, *with probability approaching one, we have*

$$|s_j(\widehat{\boldsymbol{\beta}}_\tau^o)| \le \lambda, \quad \textit{and} \quad |\widehat{\beta}_j^{\,o}| = 0, j = q_n + 1, \ldots, d_n.$$

Proof of Lemma 4.5.6: Since the $\widehat{\boldsymbol{\beta}}_\tau^o$ is the oracle estimator, $|\widehat{\beta}_j^{\,o}| = 0, j = q_n + 1, \ldots, d_n$. It remains to show that

$$\Pr\Big(|s_j(\widehat{\boldsymbol{\beta}}_\tau^o)| > \lambda, \ \text{ for some } j = q_n + 1, \ldots, d_n\Big) \to 0, \ \text{ as } n \to \infty.$$

Let $\mathcal{D} = \{i : Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\tau^o = 0\} = \{i : Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\tau 1}^o = 0\}$, then for $j = q_n + 1, \ldots, d_n$,

$$\begin{aligned} s_j(\widehat{\boldsymbol{\beta}}_\tau^o) &= (1-\tau)n^{-1}\sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\tau^o < 0) - \tau n^{-1}\sum_{i=1}^n X_{ij} I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\tau^o > 0) - n^{-1}\sum_{i=1}^n X_{ij} v_i, \\ &= n^{-1}\sum_{i=1}^n X_{ij}\big\{I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\tau^o \le 0) - \tau\big\} - n^{-1}\sum_{i=1}^n X_{ij}\big\{v_i + (1-\tau) I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\tau^o = 0)\big\} \end{aligned}$$

$$= \quad n^{-1}\sum_{i=1}^{n} X_{ij}\big\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} \le 0) - \tau\big\} - n^{-1}\sum_{i\in\mathcal{D}} X_{ij}[v_i^* + (1-\tau)],$$

where $v_i^* \in [\tau - 1, \tau]$ with $i \in \mathcal{D}$ satisfies $s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o}) = 0$ with $v_i = v_i^*$, for $j = 1, \ldots, q_n$ by Lemma 4.5.5.

$$\begin{aligned}
& \Pr(|s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o})| > 2\lambda, \quad \text{for some } j = q_n + 1, \ldots, d_n) \\
\le \quad & \Pr\Big(\Big|n^{-1}\sum_{i=1}^{n} X_{ij}\big\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} \le 0) - \tau\big\}\Big| > \lambda, \quad \text{for some } j = q_n + 1, \ldots, d_n\Big) \\
& + \quad \Pr\Big(\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\{v_i^* + (1-\tau)\}\Big| > \lambda, \quad \text{for some } j = q_n + 1, \ldots, d_n\Big) \\
=: \quad & T_{n1} + T_{n2}.
\end{aligned}$$

First, we deal with T_{n2}. Let $M = O(n^{\tau})$ with a carefully chosen $0 < \tau < 1/2$, we have that

$$\begin{aligned}
T_{n2} \quad \le \quad & \Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| \le M\}\{v_i^* + (1-\tau)\}\Big| > \lambda/2\Big) \\
& + \quad \Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}[v_i^* + (1-\tau)]\Big| > \lambda/2\Big) \\
=: \quad & T_{n21} + T_{n22}.
\end{aligned}$$

Since $(\mathbf{x}_{i,\mathcal{A}}, Y_i)$ are in general positions (Koenker, 2005, Section 2.2), with probability approaching one there exists exactly $q_n + 1$ elements in $\mathcal{D}$. Thus, with probability approaching one,

$$\begin{aligned}
& \max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| \le M\}\{v_i^* + (1-\tau)\}\Big| \\
& \le M(q_n + 1)n^{-1} = O(n^{\tau + c_1 - 1}) = o(\lambda),
\end{aligned}$$

where the last equality holds for $\lambda = o(n^{-(1-c_2)/2})$ and $0 < \tau < 1/2$. Therefore,

$T_{n21} \to 0$ as $n \to \infty$. Next, we deal with T_{n22}. Note that the events satisfy

$$\left\{\left|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}\right| > \lambda/2\right\} \subseteq \{|X_{ij}| > M, \text{ for some } i \in \mathcal{D}\},$$

because that if $|X_{ij}| \leq M$ for all $i \in \mathcal{D}$, then $n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\} = 0$. Therefore,

$$\begin{aligned}
T_{n22} &\leq d_n \Pr\left(\left|n^{-1}\sum_{i\in\mathcal{D}} X_{ij}\mathbf{1}\{|X_{ij}| > M\}[v_i^* + (1-\tau)]\right| > \lambda/2\right) \\
&\leq d_n(q_n+1)\max_{i\in\mathcal{D}, q_n+1\leq j\leq d_n} \Pr\left(|X_{ij}| > M\right) \\
&\leq d_n(q_n+1)\exp(-tM)E\{\exp(t|X_{ij}|)\} \\
&\leq C d_n(q_n+1)\exp(-tM) \\
&= O(n)O(n^{c_1})\exp(-tn^{\tau}) \to 0,
\end{aligned}$$

as $n \to \infty$, where the second inequality holds from Markov's inequality and the third inequality follows from Condition (A4.3). Therefore,

$$T_{n2} = T_{n21} + T_{n22} \to 0, \text{ as } n \to \infty.$$

Therefore, it is enough to show that

$$\Pr\left(\left|n^{-1}\sum_{i=1}^{n} X_{ij}\{I(Y_i - \mathbf{x}_{i,\widehat{\mathcal{A}}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau}^{o} < 0) - \tau\}\right| > \lambda, \quad \text{for some } j = q_n+1, \ldots, d_n\right) \to 0,$$

as $n \to \infty$.

Next, we consider

$$\begin{aligned}
&\Pr\left(\max_{q_n+1,\ldots,d_n}\left|n^{-1}\sum_{i=1}^{n} X_{ij}\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o} \leq 0) - \tau\}\right| > \lambda\right) \\
\leq\ &\Pr\left(\max_{q_n+1,\ldots,d_n}\left|n^{-1}\sum_{i=1}^{n} X_{ij}\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o} \leq 0) - \tau\}\right| > \frac{\lambda}{2}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\tau 1}^{o}\le 0)-I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)\}\Big|>\frac{\lambda}{2}\Big)\\
\le\;& \Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)-\tau\}\Big|>\frac{\lambda}{2}\Big)\\
&+\Pr\Big(\max_{q_n+1,\ldots,d_n}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{\tau 1}^{o}\|\le\Delta\sqrt{q_n/n}}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\Big[I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1\le 0)-I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)\\
&\qquad-\big\{\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1\le 0)-\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)\big\}\Big]\Big|>\frac{\lambda}{4}\Big)\\
&+\Pr\Big(\max_{q_n+1,\ldots,d_n}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{\tau 1}^{o}\|\le\Delta\sqrt{q_n/n}}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\\
&\qquad\Big\{\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_1\le 0)-\Pr(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)\Big\}\Big|>\frac{\lambda}{4}\Big)\\
=:\;& J_{n1}+J_{n2}+J_{n3}.
\end{aligned}
$$

First, let us consider J_{n1}. We choose a $M=O(n^{\tau})$ with $0<\tau<1/2$, then

$$
\begin{aligned}
J_{n1}\;\le\;& \Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}\{|X_{ij}|\le M\}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}<0)-\tau\}\Big|>\lambda/4\Big)\\
&+\Pr\Big(\max_{q_n+1,\ldots,d_n}\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}\{|X_{ij}|> M\}\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)-\tau\}\Big|>\lambda/4\Big)\\
=:\;& J_{n11}+J_{n12}.
\end{aligned}
$$

By Hoeffding's inequality, we have that

$$
\Pr\Big(\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}(|X_{ij}|\le M)\{I(Y_i-\mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\tau 1}^{o}\le 0)-\tau\}\Big|>\lambda/4\Big)\le 2\exp\left(-\frac{n\lambda^2}{8M^2}\right).
$$

Thus, $J_{n11}\le 2d_n\exp\{-n\lambda^2/(8M^2)\}\to 0$, as $n\to\infty$. On the other hand, we can similarly follow the arguments that deal with T_{n22} and have that

$$
\begin{aligned}
J_{n12}\;&\le\; d_n\Pr\Big(\Big|n^{-1}\sum_{i=1}^{n}X_{ij}\mathbf{1}\{|X_{ij}|>M\}\Big|>\lambda/4\Big)\\
&\le\; d_n n\max_{n+1\le i\le 2n}\Pr\left(|X_{ij}|>M\right)\to 0,
\end{aligned}
$$

as $n \to \infty$. Therefore,

$$J_{n1} = J_{n11} + J_{n12} = o(1).$$

Following similar arguments for proving Lemma 4.3 of Wang, Wu and Li (2012), we can show that $J_{n2} = o(1)$. It remains to deal with J_{n3}. For a fixed $M = O(n^{\tau})$ with $0 < \tau < 1/2$,

$$\begin{aligned}
J_{n3} &\leq \Pr\Big(\max_{q_n+1,\ldots,d_n} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \leq \Delta\sqrt{q_n/n}} \Big| n^{-1} \sum_{i=1}^{n} X_{ij} \mathbf{1}\{|X_{ij}| \leq M\} \\
&\qquad \Big\{ \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}^o_{\tau 1} \leq 0) \Big\} \Big| > \frac{\lambda}{8} \Big) \\
&+ \Pr\Big(\max_{q_n+1,\ldots,d_n} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \leq \Delta\sqrt{q_n/n}} \Big| n^{-1} \sum_{i=1}^{n} X_{ij} \mathbf{1}\{|X_{ij}| > M\} \\
&\qquad \Big\{ \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}^o_{\tau 1} \leq 0) \Big\} \Big| > \frac{\lambda}{8} \Big) \\
&=: J_{n31} + J_{n32}.
\end{aligned}$$

To handle J_{n31}, we observe that

$$\begin{aligned}
&\max_{q_n+1,\ldots,d_n} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \leq \Delta\sqrt{q_n/n}} \Big| n^{-1} \sum_{i=1}^{n} X_{ij} \mathbf{1}\{|X_{ij}| > M\} \\
&\qquad \Big\{ \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}_1 \leq 0) - \Pr(Y_i - \mathbf{x}^{\mathrm{T}}_{i,\mathcal{A}} \boldsymbol{\beta}^o_{\tau 1} \leq 0) \Big\} \Big| \\
&\leq M \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \leq \Delta\sqrt{q_n/n}} \Big| E\Big\{ f(\zeta|\mathbf{x}_{\mathcal{A}}) \mathbf{x}^{\mathrm{T}}_{\mathcal{A}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}) \Big\} \Big| \\
&\leq M \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \leq \Delta\sqrt{q_n/n}} \lambda^{1/2}_{\max} \left\{ E(\mathbf{x}_{\mathcal{A}} \mathbf{x}^{\mathrm{T}}_{\mathcal{A}}) \right\} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1}\| \\
&\leq O(\sqrt{q_n/n}) = O(n^{-(1-c_1)})
\end{aligned}$$

where $f(\cdot|\mathbf{x}_{\mathcal{A}})$ is defined in Condition (A4.6) with ζ is between $u^o_{\tau} + \mathbf{x}^{\mathrm{T}}_{\mathcal{A}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}^o_{\tau 1})$ and u^o_{τ} and thus the second inequality follows Condition (A4.6) and Cauchy-Schwartz inequality, and the third inequality follows Condition (A4.5). Consequently, to-

gether with $\lambda = o\{n^{-(1-c_2)/2}\}$, we have that $J_{n31} \leq \Pr\{O(n^{-(1-c_1)}) > \lambda/8\} = o(1)$. We can also follow similar arguments for handling J_{n12} and obtain that $J_{n32} = o(1)$. Therefore, $J_{n3} = J_{n31} + J_{n32} = o(1)$.

Hence,

$$\begin{aligned} & \Pr\Big(\max_{q_n+1,\ldots,d_n} \Big| n^{-1} \sum_{i=1}^{n} X_{ij}\{I(Y_i - \mathbf{x}_{i,\mathcal{A}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\tau 1}^{o} < 0) - \tau\} \Big| > \lambda \Big) \\ \leq \;& J_{n1} + J_{n2} + J_{n3} = o(1), \end{aligned}$$

which implies that

$$\Pr\Big(|s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o})| > \lambda, \;\; \text{for some } j = q_n + 1, \ldots, d_n \Big) \to 0, \;\; \text{as } n \to \infty.$$

That is, $\Pr\Big(|s_j(\widehat{\boldsymbol{\beta}}_{\tau}^{o})| \leq \lambda, \;\; \text{for} \;\; j = q_n + 1, \ldots, d_n \Big) \to 1, \;\; \text{as } n \to \infty$. This completes the proof.

□

Recall that

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}),$$

where $h(\boldsymbol{\beta}) = \sum_{j=1}^{d_n} H_\lambda(\beta_j)$, and for the SCAD penalty function,

$$H_\lambda(\beta_j) = \begin{cases} 0, & 0 \leq |\beta_j| < \lambda; \\ (\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)/\{2(a-1)\}, & \lambda \leq |\beta_j| \leq a\lambda; \\ \lambda|\beta_j| - (a+1)\lambda^2/2, & |\beta_j| > a\lambda. \end{cases}$$

Thus, $\partial h(\boldsymbol{\beta})/\partial\beta_0 = 0$, and for $j = 1, \ldots, d$

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \begin{cases} 0, & 0 \le |\beta_j| < \lambda; \\ \{\beta_j - \lambda \mathrm{sgn}(\beta_j)\}/(a-1), & \lambda \le |\beta_j| \le a\lambda; \\ \lambda \mathrm{sgn}(\beta_j), & |\beta_j| > a\lambda. \end{cases}$$

By the definition of the set $\partial g(\boldsymbol{\beta})$, we obtain

$$\partial g(\widehat{\boldsymbol{\beta}}_\tau) = \left\{ \boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_d)^{\mathrm{T}} \in \mathbb{R}^{d_n+1} : \xi_j = s_j(\widehat{\boldsymbol{\beta}}_\tau) + \lambda l_j \right\}.$$

By Lemma 4.5.5, there exist v_i^* such that $s_j(\widehat{\boldsymbol{\beta}}_\tau) = 0$ for $j = 1, 2, \ldots, q_n$. By Lemma 4.5.6, $\Pr(|s_j(\widehat{\boldsymbol{\beta}}_\tau)| \le \lambda, \text{ for } j = q_n+1, \ldots, d_n) \to 1$, as $n \to \infty$. Thus, there exist $l_j^* \in [-1, 1]$, such that $\Pr\left\{ s_j(\widehat{\boldsymbol{\beta}}_\tau) + \lambda l_j = 0, \text{ for } j = q_n + 1, \ldots, d_n \right\} \to 1$, as $n \to \infty$.

We denote $\boldsymbol{\xi}^*$ be the vector $\boldsymbol{\xi}$ in $\partial g(\widehat{\boldsymbol{\beta}}_\tau)$ with $v_i = v_i^*$ and $l_j = l_j^*$, then we have

$$\Pr(\boldsymbol{\xi}^* \subseteq \partial g(\widehat{\boldsymbol{\beta}}_\tau)) \to 1, \text{ as } n \to \infty,$$

where $\boldsymbol{\xi}^* = (\xi_0^*, \xi_1^*, \ldots, \xi_{d_n}^*)^{\mathrm{T}}$ satisfies

$$\xi_0^* = 0; \quad \xi_j^* = \lambda \mathrm{sgn}(\widehat{\beta}_j^{\,o}), j = 1, 2, \ldots, q_n; \quad \xi_j^* = 0, j = q_n + 1, \ldots, d_n.$$

Consider any $\boldsymbol{\beta}$ in a ball $U(\widehat{\boldsymbol{\beta}}_\tau, \lambda/2)$ in $\mathbb{R}^{d_n+1}$ with the center $\widehat{\boldsymbol{\beta}}_\tau$ and radius $\lambda/2$. It suffices to show that

$$\Pr\left(\boldsymbol{\xi}^* = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, \text{ for } j = 0, 1, \ldots, d_n \right) \to 1, \text{ as } n \to \infty.$$

(1) For $j = 0$, it is obvious that $\xi_0^* = \partial h(\boldsymbol{\beta})/\partial\beta_0 = 0$;

(2) For $j = 1, \ldots, q_n$, $\xi_j^* = \lambda \text{sgn}(\widehat{\beta}_j^o)$. Lemma 4.5.5 states that, with portability approaching one, for any $\boldsymbol{\beta} = (\beta_0, \beta_j, \ldots, \beta_{d_n}) \in U(\widehat{\boldsymbol{\beta}}_\tau, \lambda/2)$,

$$\min_{1 \le j \le q_n} |\beta_j| \ge \min_{1 \le j \le q_n} |\widehat{\beta}_j^{\,o}| - \max_{1 \le j \le q_n} |\widehat{\beta}_j^{\,o} - \beta_j| \ge (a + 1/2)\lambda - \lambda/2 = a\lambda.$$

Thus,

$$\Pr\left\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \lambda \text{sgn}(\beta_j), j = 1, \ldots, q_n \right\} \to 1, \text{ as } n \to \infty.$$

Since $\boldsymbol{\beta} \in U(\widehat{\boldsymbol{\beta}}_\tau, \lambda/2)$, for n sufficiently large, $\widehat{\beta}_j^o$ and β_j has the same sign. Therefore,

$$\Pr(\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j^*, j = 1, \ldots, q_n) \to 1, \text{ as } n \to \infty.$$

(3) For $j = q_n + 1, \ldots, d_n$, $\xi_j^* = 0$. Lemma 4.5.6 states that, with portability approaching one, for any $\boldsymbol{\beta} = (\beta_0, \beta_j, \ldots, \beta_{d_n}) \in U(\widehat{\boldsymbol{\beta}}_\tau, \lambda/2)$,

$$\max_{q_n+1 \le j \le d_n} |\beta_j| \le \max_{q_n+1 \le j \le d_n} |\widehat{\beta}_j^{\,o}| + \max_{q_n+1 \le j \le d_n} |\widehat{\beta}_j^{\,o} - \beta_j| \le \lambda/2.$$

Thus,

$$\Pr\left\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0, j = q_n + 1, \ldots, d_n \right\} \to 1, \text{ as } n \to \infty.$$

Then,

$$\Pr\left\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j^*, j = q_n + 1, \ldots, d_n \right\} \to 1, \text{ as } n \to \infty.$$

By above (1), (2) and (3), we have

$$\Pr\left\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j^*, j = 0, 1, \ldots, d_n \right\} \to 1, \text{ as } n \to \infty.$$

That is, for $\forall \boldsymbol{\beta} \in U(\widehat{\boldsymbol{\beta}}_\tau, \lambda/2)$

$$\Pr(\partial h(\boldsymbol{\beta}) \cap \partial g(\widehat{\boldsymbol{\beta}}_\tau^o) \neq \emptyset) \to 1, \text{ as } n \to \infty.$$

Hence, by Lemma 4.5.4,

$$\Pr(\widehat{\boldsymbol{\beta}}_\tau^o \in \mathcal{B}_n^*(\lambda)) \to 1.$$

This completes the proof. □

Chapter 5

Conclusion and Future Research

5.1 Conclusion Remarks

In this dissertation, we systematically reviewed the existing variable selection methods for the high dimensional regressions and independence screening procedures for the ultrahigh dimensional problems. We proposed a novel sure independence screening procedure using distance correlation (DC-SIS, for short) for the ultrahigh dimensional data analysis in Chapter 3 and the two-stage robust feature screening and variable selection estimation procedure for the ultrahigh dimensional heteroscedastic single-index models.

In Chapter 3, we proposed the DC-SIS to select potential predictors when the number of predictors, p, greatly diverging as the sample size n. Then, we theoretically studied that the proposed DC-SIS processes the desirable sure screening property in the terminology of Fan and Lv (2008). That is, with an appropriate threshold, it can select all important predictors with probability approaching to one as n goes to the infinity. The proposed DC-SIS has several appealing properties which are unique among the existing independence screening procedures. Based on the fact that the distance correlation is defined for predictors $\mathbf{x}$ and responses $\mathbf{y}$

in arbitrary dimensions, it allows to consider independent screening for groupwise predictors and multivariate responses. Because the distance correlation characterizes the dependence between $\mathbf{x}$ and $\mathbf{y}$, the proposed DC-SIS built on distance correlation imposes little assumption on regression structure. Therefore, it allows arbitrary regression relationship between $\mathbf{x}$ and $\mathbf{y}$ and thus is robust to model misspecification. In the end of the chapter, we examined the finite-sample performance of the proposed procedure via Monte Carlo studies and three real data examples, which supported that the DC-SIS perform quit well in the ultrahigh dimensional regressions.

In Chapter 4, in the first stage, we proposed the new robust independent ranking and screening procedure and demonstrated that the robust RIRS enjoys both the ranking consistency property (Zhu, Li, Li and Zhu, 2011) and the sure screening property (Fan and Lv, 2008) under mild conditions. In the second stage, we applied the penalized linear quantile regression to further select the important variables and estimate the direction of the index parameter. It maintains the appealing robustness property of the RIRS in that it is insensitive to the presence of extreme values and outliers in the response. We demonstrated that the resulting estimator is consistent and processes the oracle property (Fan and Li, 2001), even the single-index model structure is mis-specified. Numerical studies confirmed the outstanding finite sample performances of the proposed two-stage estimation procedure.

5.2 Future Research

5.2.1 False Positive Rate Controlling

Independence screening procedures are commonly used for feature selection. However, there always exist positive selection false and negative deletion false. That is, to select the truly irrelevant predictors into the model and to exclude the truly important predictors out of the model, respectively. How to control the false positive and false negative rates is important. Therefore, it becomes an open issue to theoretically study both false rates.

5.2.2 Criteria to Independence Screening

There are many well-established model-based and model-free independence screening procedures in literature (Fan and Lv, 2008; Fan and Song, 2010; Fan, Feng and Song, 2011; Zhu, Li, Li and Zhu, 2011; Li, Zhong and Zhu, 2012). However, there are no commonly-used criteria to judge the independence screening procedures. When more than one independence screening procedures are available, it is difficult to choose a proper one, especially for the real data analysis. In the future, we will examine the existing screening procedures and further design criteria to choose an appropriate independence screening procedure.

5.2.3 Application to Genome-Wide Association Studies

The modern development of genotyping technologies allows the fast and accurate collection of genotype data throughout the entire genome. Genome-wide association studies (GWASs) can be used to test the associations between Single nucleotide polymorphism (SNPs) and diseases and estimate genetic effects of SNPs on traits. Both cases involve hundreds of thousands of SNPs collected from hun-

dreds or thousands of subjects. It is impossible for traditional regression methods to analyze the data where the number of SNPs highly exceeds the sample size. As a direct application, the proposed independence screening procedures in this dissertation can be used to select the SNPs which have significant genetic effects on traits.

On the other hand, the interaction effects among SNPs always exist in the genetics. To consider the interaction terms into the regression model will exponentially increase the number of predictors. We may further consider to apply the proposed independence screening procedures to detect both important main predictors and interaction effects in the GWASs.

Bibliography

Akaike, H. (1973) "Maximum likelihood identification of Gaussian autoregressive moving average models", *Biometrika*, **60**, 255–265.

Altham, P. M. E. (1984), "Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood," *Journal of the Royal Statistical Society, Series B*, **46**, 118–119.

Allen, D.(1974), "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction," *Technometrics*, **16**, 125–127.

Antoniadis, A. (1997), "Wavelets in Statistics: A Review (with discussion)", *Journal of the Italian Statistical Association*, **6**, 97–144.

Antoniadis, A. and Fan, J. (2001), "Regularization of wavelets approximations (with discussion)", *Journal of the American Statistical Association*, **96**, 939–967.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000), "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nature Genetics*, **25**, 25-29.

Bickel, P., and Levina E. (2008), "Regularized estimation of large covariance matrices," *Annals of Statistics*, **36**, 199–227.

Bild, A., Yao, G., Chang, J. T., Wang, Q., Potti, A., et al. (2006), "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature* **439** 353–357.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, **24**, 2350–2383.

Candes, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)," *Annals of Statistics*, **35**, 2313–2404.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997), " Generalized partially linear single-index models," *Journal of the American Statistical Association*, **92**, 477–489.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011), "A regularized Hotelling's T^2 test for pathway analysis in proteomic studies," *Journal of the American Statistical Association* **106** 1345–1360.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M. and Shefield, V. C. (2006), "Homozygosity Mapping with SNP Arrays Identifies a Novel Gene for Bardet-Biedl Syndrome (BBS10)," *Proceeding of the National Academy of Sciences*, **103**, 6287–6292.

Chmielewski, M.A., "Elliptically Symmetric Distributions: A Review and Bibliography", *International Statistical Review*, **49**, 67-74.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numer. Math*, **31**, 377–403.

Duan, N. H. and Li, K. C. (1991), "Slicing regression: A link-free regression method," *Annals of Statistics*, **19**, 505–530.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956), "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator", *Annals of Mathematical Statistics*, **27**, 642–669.

Efron, B., Hastie T., Johnstone, I. and Tibshirani, R.(2004), "Least angle regression (with discussion)," *Annals of Statistics*, **32**, 409–499.

Efron, B., and Tibshirani, R. (2007), "On Testing the Significance of Sets of Genes," *The Annals of Applied Statistics*, **1**, 107–129.

Fan, J. (1997), "Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis", *Journal of the Italian Statistical Association*, **6**, 131–38.

Fan, J., Feng, Y. and Song, R. (2011), "Nonparametric independence screening in sparse ultra-high dimensional additive models," *Journal of the American Statistical Association*, **106**, 544–557

Fan, J., Feng, Y. and Wu, Y. (2010), "Ultrahigh dimensional variable selection for Cox's proportional hazards model," IMS Collection, **6**, 70–86.

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

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

Fan, J. and Lv, J. (2010) "A selective overview of variable selection in high dimensional feature space". *Statistica Sinica* **20** 101–148.

Fan, J., Samworth, R. and Wu, Y. (2009), "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of Machine Learning Research*, **10**, 1829–1853.

Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Fang, K.-T., Kotz, S. and Ng, K.W. (1989), "Symmetric Multivairate and Related Distributions," *Chapman and Hall,* London.

Foster, D. P. and George, E. I. (1994) "The risk inflation criterion for multiple regression," *Annals of Statistics* **22** 1947–1975.

Frank, I. E. and Friedman, J. H. (1993) "A statistical view of some chemometrics regression tools" *Technometrics* **35** 109–148.

Härdle, W., Hall, P., Ichimura, H. (1993), "Optimal smoothing in single-index models," *Annals of Statistics*, **21**, 157–178.

Hall, P. and Li, K. C. (1993). "On almost linearity of low dimensional projection from high dimensional data," *Annals of Statistics*, **21**, 867–889.

Hall, P. and Miller, H. (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, **18**, 533–550.

Hoerl, A. and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, **12**, 55–67.

Horowitz, J. L. and Härdle, W. (1996), "Direct semiparametric estimation of single-index models with discrete covariates," *Journal of the American Statistical Association*, **91**, 1632–1639.

Huang, J., Ma, S. G. and Zhang, C. H. (2008), "Adaptive Lasso for sparse high-dimensional regression models," *Statistica Sinica*, **18**, 1603–1618

Ji, P. and Jin, J.(2012), "UPS delivers optimal phase diagram in high dimensional variable selection," *Annals of Statistics*, **1**, 73–103.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., et al. (2008), "Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses," *Science*, **321** 1801.

Kim, Y., Choi, H. and Oh, H. S. (2008), "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association,* **103**, 1665–1673.

Knight, K. (1998), "Limiting distributions for L_1 regression estimators under general conditions," *The Annals of Statistics*, **26**, 755–770.

Koenker, R. (2005). *Quantile Regression. Cambridge University Press.*

Koenker, R. and Bassett, G.W. (1978),"Regression quantiles," *Econometrica*, **46**, 33–50.

Kong, E. and Xia, Y.(2007), "Variable Selection for the single-index model," *Biometrika*, **94**, 217–229.

Li, J., Zhong, W., Li, R. and Wu, R.(2012), "A fast algorithm for detecting gene-gene interactions in Genome-Wide Association Studies," *manuscript.*

Li, R., Zhong, W. and Zhu, L.P.(2012), "Feature screening via distance correlation learning," *Journal of the American Statistical Association,* forthcoming.

Liang, H., Liu, X., Li, R. and Tsai, C.(2010), "Estimation and testing for partially linear single-index models," *Annals of Statistics*, **38**, 3811–3836.

Mallows, C. L. (1973), "Some comments on Cp", *Technometrics*, **15**, 661–675.

Miller, A. J. (2002), *Subset Selection in Regression (2nd edition),* Chapman & HALL/CRC, New York.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., et al. (2003), "PGC-1-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes," *Nature Genetics*, **34**, 267-273.

Naik, P. A. and Tsai, C.-L. (2001), "Single-index model selections," *Biometrika*, **88**, 821–832.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989), "Semiparametric estimation of index coefficient," *Econometrica,* **51**, 1403–1430.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression approach for microarray data analysis," *Journal of Computational Biology,* **10**, 961–980.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Shefield, V. C. and Stone, E. M. (2006), "Regulation of gene expression in the mammalian eye and its relevance to eye disease," *Proceeding of the National Academy of Sciences*, **103**, 14429–14434.

Schwartz, G. (1978) "Estimating the dimension of a model," *Annals of Statistics*, **6**, 461–464.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics,* New York: John Wiley & Sons Inc.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, **7**, 221–264.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., et al. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences of the USA*, **102**, 15545–15505.

Székely, G. J. and Rizzo, M. L. (2009), "Brownian distance covariance," *Annals of Applied Statistics*, **3**, 1233–1303.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, **35**, 2769–2794.

Tao, P. D. and An, L.T.H. (1997), "Convex analysis approach to D.C. programming: theory, algorithms and applications," *Acta Mathematica Vietnamica*, **22**, 289–355.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005), "Discovering Statistically Significant Pathways in Expression Profiling Studies," *Proceedings of the National Academy of Sciences of the USA*, **102**, 13544–13549.

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

van der Vaart, A. W. and Wellner, J. A. (1996), " Weak Convergence and Empirical Processes," *Springer*, New York.

Wang, H. (2009), " Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, **104**, 1512–1524.

Wang, L., Wu, Y. and Li, R.(2012), "Quantile regression for analyzing heterogeneity in ultra-high dimension," *Journal of the American Statistical Association*, in press.

Wasserman, L. and Roeder, K. (2009), "High-dimensional variable selection," *Annals of Statistics*, **37**, 2178–2201.

Xia, Y. C., Tong, H., Li, W. K., Zhu, L. X. (2002). "An adaptive estimation of optimal regression subspace," *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhang, C. (2010), "Nearly unbiased variable selection under minimax concave penalty", *Annals of Statistics*, **38**, 894–942.

Zhu, L., Huang, M. and Li, R.(2011), "Semiparametric quantile regression with high-dimensional covariates," *Statistica Sinica*, Advance online publication. doi: 10.5705/ss.2010.199.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011) "Model-free feature screening for ultrahigh dimensional data," *Journal of the American Statistical Association*, **106**, 1464–1475.

Zhu, L. P., Qian, L. and Lin, J., "Variable selection in a class of single-index models," *Annals of the Institute of Statistical Mathematics*, **63**, 1277–1293.

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

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

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models (with discussions)," *Annals of Statistics*, **36**, 1509–1533.

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

Wei Zhong

333 Thomas Building
Department of Statistics
The Pennsylvania State University
State College, PA, 16802

Phone: (814)-321-7555
Email:wxz123@psu.edu
Webpage:www.personal.psu.edu/wxz123

Education

Ph.D. Candidate in Statistics 2008-2012 (expected)
Department of Statistics, The Pennsylvania State University (PSU) GPA: 3.97/4.0
PhD Dissertation: Feature Screening and Variable Selection for Ultrahigh Dimensional Data Analysis.

B.S. in Statistics and Financial Mathematics 2004-2008
School of Mathematics, Beijing Normal University (BNU) GPA: 3.76/4.0

Research Interests

- Ultrahigh/high Dimensional Data Analysis: large p small n problems
- Large Covariance Matrix Estimation with its Application in Portfolio Theory
- Application of Statistics in Operation Research, Economics and Finance

Research Papers

Li, R., **Zhong, W.** and Zhu, L.P.(2012), Feature screening via distance correlation learning, *Journal of the American Statistical Association,* forthcoming.

Coffman, D. and **Zhong, W.**(2012), Assessing Mediation using Marginal Structural Models in the Presence of Confounding and Moderation, *Psychological Methods,* forthcoming.

Li, J., **Zhong, W.**, Li, R. and Wu, R.(2012), A fast algorithm for detecting gene-gene interactions in Genome-Wide Association Studies, *manuscript.*

Zhong, W., Li, R. and Zhu, L.P.(2012), Robust Feature Screening and Selection for Ultrahigh Dimensional Heteroscedastic Single-Index Models, *manuscript.*

Research Experience

Research Assistant, Department of Statistics and the Methodology Center, PSU
Summer 2009, 2010 and May-Dec. 2011

Marketing Analyst, Promotion Response Analytics, Merck & Co., PA (Internship)
Jun.-Aug. 2010

Statistical Consultant, Statistical Consulting Center, PSU Aug. 2010-May 2011

Teaching Experience

Instructor, Department of Statistics, PSU. Summer 2011 & Spring 2012
Courses: STAT 480 (Introduction to SAS) and STAT 401 (Experimental Methods).

Teaching Assistant, Department of Statistics, PSU. Aug. 2008-May 2011
Courses: STAT 301 (Statistical Analysis), STAT 250 (Introduction to Biostatistics) and STAT 318 (Elementary Probability).

Conference Presentation

ENAR Annual Statistical Meetings, Miami, FL. March 2011.
The 2011 Rao Prize Conference, PSU, State College, PA, May 2011.
Joint Statistical Meetings, Miami Beach, FL, August 2011.