The Pennsylvania State University
The Graduate School

NONPARAMETRIC INDEPENDENCE SCREENING AND TEST-BASED SCREENING VIA THE VARIANCE OF THE REGRESSION FUNCTION

A Dissertation in
Statistics
by
Won Chul Song

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2015

The dissertation of Won Chul Song was reviewed and approved* by the following:

Michael G. Akritas
Professor of Statistics
Dissertation Advisor, Chair of Committee

Runze Li
Verne M. Willaman Professor of Statistics

Dennis K.J. Lin
Distinguished Professor of Statistics

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

Aleksandra B. Slavković
Associate Professor of Statistics
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

This dissertation develops procedures for screening variables, in ultrahigh-dimensional settings, based on their predictive significance. First, we review existing literature on the sure screening procedures for analyzing ultrahigh-dimensional data. Second, we develop a screening procedure by ranking the variables, according to the variance of their respective marginal regression functions (RV-SIS). This is in sharp contrast with existing literature on feature screening, which ranks the variables according to some correlation measures with the response, and hence select variables with no predictive power (e.g., variables that influence aspects of the conditional distribution of the response other than the regression function). The RV-SIS is easy to implement and does not require any model specification for the regression functions (such as linear or other semi-parametric modeling). We show that, under some mild technical conditions, the RV-SIS possesses a sure independence property, which is defined by Fan and Lv (2008). Numerical comparisons suggest that RV-SIS has competitive performance compared to other screening procedure and outperforms them in many different model settings. Third, we develop a test procedure for the hypothesis of a constant regression function, and also a test-based variable screening procedure. We study the asymptotic theory for the variance of the regression function and use it to introduce a new test procedure for testing the significance of a predictor. Using the set of p-values, we introduce a variable screening procedure with a specified desirable false discovery rate by using Benjamini and Hochberg (1995) approach.

Table of Contents

List of Figures viii

List of Tables ix

Acknowledgments xi

Chapter 1
Introduction 1

1.1 Background 1

1.2 Contribution 3

1.3 Organization 5

Chapter 2
Literature Review 6

2.1 Summary 6

2.2 Independence Screening Procedure 7

2.2.1 Sure Independence Screening 7

2.2.2 Generalized Correlation Ranking 9

2.2.3 Sure Independence Screening for GLM 11

2.2.4 Sure Independence Ranking and Screening 15

2.2.5 Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models 18

2.2.6 Feature Screening via Distance Correlation Learning 22

2.2.7 Principled sure independence screening for Cox models with ultra-high-dimensional covariates 25

2.2.8 Robust Rank Correlation Based Screening 27

Chapter 3
Independence Screening via the Variance of the Regression Function **30**

3.1 Introduction 30

3.2 Nonparametric Independence Screening via the Variance of the Regression Function 31

3.2.1 Preliminaries 31

3.2.2 Thresholding Rule 33

3.2.3 Sure Screening Properties 33

3.3 Numerical Studies 36

3.3.1 Simulation Studies 36

3.3.2 Thresholding Simulation 37

3.3.3 A Real Data Example 39

3.3.4 A Real Data Example II 40

3.4 Theoretical Properties 41

3.4.1 Some Lemmas 41

3.4.2 Proof of Theorem 3.2.2 46

Chapter 4
Hypothesis testing of a constant regression function and Test-based Screening **61**

4.1 Introduction 61

4.2 Hypothesis testings of a constant regression function for variable screening and test-based screening 62

4.2.1 Preliminaries 62
4.2.2 Asymptotic Properties 63
4.2.3 Estimating the Asymptotic Distribution of Theorem 4.4.2 . . . 64
4.2.4 Test-based Screening 65
4.3 Numerical Studies 66
4.3.1 Simulation Studies 66
4.3.2 A Real Data Example 67
4.4 Theoretical Properties 68
4.4.1 Proof of Theorem 4.2.2 68
4.4.2 Estimating the percentile of the limiting distribution 72

Chapter 5
Conclusion and Future Work 78

5.1 Conclusion 78
5.2 Future Work 79

Bibliography 81

List of Figures

3.1 The spline curve of Msa.2877.0 and Msa.741.0 39

List of Tables

3.1 The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$. 52

3.2 The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$. 53

3.3 The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5$. 54

3.4 The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$. 55

3.5 The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$. 56

3.6 The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.5$. 57

3.7 The comparison of execution time of DC-SIS and RV-SIS in seconds for Model 3.(d) when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$. 58

3.8 The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(e) 58

3.9 The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(f) 59

3.10 The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(g) 59

3.11 The 5%, 25%, 50%, 75%, and 95% quantiles of submodel size using soft thresholding rule for Models 3.(e), 3.(f), and 3.(g). 60

4.1 The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.3^{|i-j|}$ 74

4.2 The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.5^{|i-j|}$ 75

4.3 The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.7^{|i-j|}$ 75

4.4 The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.3^{|i-j|}$ 76

4.5 The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.5^{|i-j|}$ 76

4.6 The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.7^{|i-j|}$ 77

Acknowledgments

I would like to express my gratitude to my academic advisor, Dr. Michael G. Akritas, for his guidance throughout my graduate study. Without his help and support, it would not be possible for me to complete my dissertation. I am also very grateful to my committee members, Dr. Jingzhi Huang, Dr. Runze Li, and Dr. Dennis K. J. Lin, for their comments and help to improve the dissertation.

In addition, I am truly grateful for my incredibly supportive family - my wife, Sunkyung Son, my sons, Hayden Hoyoung Song and Luke Seyoung Song. Without their love, support, and understanding, I could not complete this.

Chapter 1

Introduction

1.1 Background

High-dimensional data has received a lot of attention in the recent statistical literatures. Various statistical methods were developed for variable selection procedure in high-dimensional data including the Lasso (Tibshirani, 1996), the smoothly slipped absolute deviation (SCAD) (Fan and Li, 2001), the least angle regression (LARS) (Efron *et al.*, 2004), the elastic net (Zou and Hastie, 2005), the adaptive Lasso (Zou, 2006), and the Dantzig selector (Candes and Tao, 2007). All these methods can be used for analyzing the data where the number of predictors is greater than the number of the observations.

With advancements in the data collection technology, ultrahigh-dimensional data can be collected easily in many research areas such as genetic data, microarray data, and high volumn financial data. In these examples, the number of predictors (p) is some exponential function of the the number of the observations (n). In other words, $\log p = O(n^a)$ for some $a > 0$. The sparsity assumption, according to which only a small set of covariates has an effect on the response, makes the inference possible in ultrahigh-dimensional data.

The aforementioned variable selection methods may have technical difficulties and performance issues for analyzing ultrahigh-dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan *et al.*, 2009). Motivated by this, Fan and Lv (2008) recommended that a variable screening procedure be performed prior to variable selection. Working with a linear model, they introduced sure independence screening (SIS), a variable screening procedure based on Pearson's correlation coefficient. Assuming Gaussian predictors and response variable, they showed that SIS possesses the sure screening property, which means that the true predictors will be chosen with probability one as the sample size approaches to infinity. Arguing that linear correlation may fail to detect important predictors which have a nonlinear effect, Hall and Miller (2009) proposed a screening procedure based on a notion of generalized correlation and showed that covariates with sufficiently large covariance with response Y are ranked ahead of those with smaller covariance. Fan and Song (2010) extended the screening procedure to a generalized linear model using the maximum marginal likelihood estimator. It has been shown that this method possesses the sure screening property with vanishing false positive rate, that is the maximum marginal likelihood estimator of inactive predictors will approach to zero as the sample size approaches to infinity. Fan, Feng and Song (2011) introduce a nonparametric screening procedure (NIS), which uses a spline-based nonparametric estimation of the marginal regression functions, and ranks predictors by the Euclidean norm of the estimated marginal regression function (evaluated at the data points). Under the assumption of an additive model, they show that this method also possesses the sure screening property. This method also possesses the sure screening property with the vanishing false positive rate. Li, Zhong, and Zhu (2012b) proposed a ranking procedure using the distance correlation (DC-SIS). DC-SIS can be used for grouped predictors and multivariate responses. They also show that DC-SIS holds the sure screening property under fairly general conditions. Li, Peng, Zhang, and Zhu (2012a) propose a robust rank correlation screening (RRCS), which uses a ranking based on Kendall's τ rank correlation coefficient. They show that this procedure can handle semiparametric models under monotonic constraint to the link function. This procedure can be also used when there exists outliers, influence points, or heavy tailed errors. Under mild condition, it has been shown that the RRCS possesses the sure screening property.

1.2 Contribution

Variables that are relevant for prediction purposes are of particular interest in most applications. The existing screening methods fail to discern between variables that have predictive significance from those that influence the variance function or other aspects of the conditional distribution of the response. We propose a method that screens out variables without (marginal) predictive significance. The basic idea is that if variable X_i has no predictive significance, the regression function $E(Y|X_i)$ has zero variance. This leads to a method which ranks the predictors according the sample variance (evaluated at the data points) of the p estimated regression functions. We refer to Sure Independence Screening procedure based on the variance of the regression function as the RV-SIS. This is a model free screening procedure. We present the theoretical properties of the RV-SIS show that RV-SIS possesses the sure independence screening property under a general nonparametric regression setting. While the proofs use Nadaraya-Watson estimators for the marginal regression functions, the proofs (with mild modifications) continue to hold for other nonparametric regressions such as local linear estimators.

We conduct numerical simulation studies to compare the RV-SIS to SIS, DC-SIS, RRCS and NIS. The RV-SIS outperforms SIS, DC-SIS, RRCS and NIS in many different model settings. The RV-SIS procedure shows that it takes less computing time than both DC-SIS and NIS. We also compare the performance of the RV-SIS, DC-SIS, and NIS on real data.

If the conditional expected value $E(Y|X_i)$ is constant and thus has zero variance, then a variable X_i has no predictive significance. We also propose a new test procedure for testing the significance of a predictor by testing the hypothesis of a constant regression function:

$$H_0 : m(x) = c \text{ for all } x$$

where $m(x) = E(Y|X = x)$ is the nonparametric regression function. Under the null hypothesis, the variance of the regression function $\sigma_m^2 = \text{var}(m(X))$ is zero. Under mild technical conditions and the null hypothesis, the asymptotic distribution of the estimated variance of the regression function is established, and used to calculate the p-value. We also introduce a variable screening procedure using multiple testing idea. Using the Benjamini and Hochberg (1995) approach, this test-based screening procedure can control the false discovery rate. We conduct numerical simulation studies.

1.3 Organization

This dissertation is organized as follow. In Chapter 2, the existing sure independence screening procedures for ultrahigh dimensional data are reviewed as well as their theoretical properties. In Chapter 3, we introduce the RV-SIS and show that the RV-SIS holds the sure screening property. Numerical studies are followed to show the performance of the RV-SIS compared to other existing procedures. In Chapter 4, we introduce a new test procedure for testing the significance of a predictor. Using the set of p-values, we introduce a variable screening procedure using multiple testing ideas. In Chapter 5, we discuss the future research projects. The curriculum vitae is included in the end.

Chapter 2

Literature Review

2.1 Summary

High dimensional data has become popular in recent scientific research, and various statistical methods were developed for variable selection procedure in high-dimensional data. With advancements in the data collection technology, ultra-high dimensional data can be collected easily in many research areas such as genetic data, microarray data, and high volume financial data. The ultra-high dimensional data is defined by the number of predictors is some exponential function of the the number of the observations. In other words, $\log p = O(n^a)$ for some $a > 0$. The sparsity assumption, according to which only a small set of covariates has an effect on the response, makes the inference possible in ultra-high dimensional data.

In this chapter, we briefly review the existing independence screening methods as well as their theoretical properties.

2.2 Independence Screening Procedure

2.2.1 Sure Independence Screening

For analyzing the ultrahigh dimensional model, Fan and Lv (2008) introduced the Sure Independence Screening (SIS), a variable screening procedure based on Pearson's correlation coefficient.

Consider the following linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2.1}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ is an n-vector responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is an $n \times p$ random design matrix with i.i.d $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^T$ is an n-vector of i.i.d. random errors.

Denote the true model as $\mathcal{M} = \{1 \leq j \leq p : \beta_j \neq 0\}$ under sparsity assumption with the true model size $s = |\mathcal{M}|$. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$ be a p-vector that is obtained by componentwise regression, i.e

$$\boldsymbol{\omega} = \mathbf{X}^T\mathbf{y} \tag{2.2.2}$$

where $\mathbf{X}$ is the columwise standardized $n \times p$ matrix. Then $\boldsymbol{\omega}$ can be viewed as a vector of marginal Pearson's correlation of predictors with the response variable, rescaled by the standard deviation of the response.

The SIS procedure ranks the importance of predictors according to the magnitude of ω_j and selects predictors that have a high correlation with the response Y. More specifically, for any given $\gamma \in (0, 1)$, the SIS selects a submodel $(\widehat{\mathcal{M}})$ that contains the first $[\gamma n]$ largest $|\omega_j|$

$$\widehat{M_\gamma} = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.2.3}$$

where $[\cdot]$ denotes the floor function.

Fan and Lv (2008) defined the sure screening property, that is the true predictors will be chosen with probability 1 as the sample size n approaches to ∞. The following conditions are required to establish the sure screening property.

Define

$$\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{x}, \text{ and } \mathbf{Z} = \mathbf{X}\mathbf{\Sigma}^{-1/2} \tag{2.2.4}$$

where $\mathbf{x} = (X_1, ..., X_p)^T$ and $\mathbf{\Sigma} = \text{cov}(\mathbf{x})$.

(A1) $p > n$ and $\log p = O(n^\xi)$ for some $\xi \in (0, 1 - 2\kappa)$, where κ is given by (A3).

(A2) $\mathbf{z}$ has a spherically symmetric distribution and for the random matrix $\mathbf{Z}$, there exist some constants $c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P(\lambda_{max}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T) > c_1 \text{ and } \lambda_{min}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T) > 1/c_1) \leq e^{-C_1 n}$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \tilde{p} \leq p$, where $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ denote the largest and smallest eigenvalues of a matrix, respectively.

Also, $\epsilon \sim N(0, \sigma^2)$ for some $\sigma > 0$

(A3) $\text{var}(Y) = O(1)$ and for some $\kappa \geq 0$ and $c_2, c_3 > 0$,

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

(A4) For some $\tau \geq 0$ and $c_4 > 0$ such that

$$\lambda_{max}(\mathbf{\Sigma}) \leq c_4 n^{\tau}$$

Condition (A1) shows that the proposed SIS works for the ultrahigh dimensional setting. By setting a lower bound, Condition (A3) removes the situation where a significant variable is marginally uncorrelated with the response Y, but jointly correlated with Y. κ controls the rate of probability error in recovering the true spare model. Condition (A4) excludes the situation of strong collinearity among predictors.

Theorem 2.2.1. Under conditions (A1) - (A4), assuming the true model size $s \leq [\gamma n]$, if $2\kappa + \tau < 1$ then there exist some $\theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have for some $C > 0$,

$$\mathrm{P}(M \subset \widehat{M_{\gamma}}) = 1 - O(\exp(-Cn^{1-2\kappa}/\mathrm{log}n)),$$

therefore,

$$\mathrm{P}(M \subset \widehat{M_{\gamma}}) \to 1, \text{ as } n \to \infty$$

The above theorem shows that the SIS has the sure independence screening property and reduce the dimension of predictors from p down to $d = [\gamma n]$.

2.2.2 Generalized Correlation Ranking

Pearson's correlation can perform effectively in the linear relationship case between each predictor X_j and the response Y. However, nonliearity of response can result in significant predictors being overlooked. Since the SIS is based on the Pearson's correlation, if an active predictor X_j is nonlinearly related to the response Y, the SIS is likely to fail to detect this predictor. In order to overcome the weakness of the SIS, Hall and Miller (2009) proposed an approach based on ranking generalized empirical correlations between the response variable and components of the predictors. This procedure can capture both linear and nonlinear relationship between the predictor and the response. Assume that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d observed pairs of p-vectors X_i and scalars Y_i. The generalized correlation between Y and X_j is defined as

$$\psi_j = \sup_{h \in \mathcal{H}} \frac{\text{cov}(\{h(X_j)\}, Y)}{\sqrt{var\{h(X_j)\}var(Y)}} \tag{2.2.5}$$

where $\mathcal{H}$ is the vector space generated by any given set of functions h. If $\mathcal{H}$ is set to be the class of linear functions, then it is the Pearson's correlation.

The generalized correlation between Yi and the jth component X_{ij} of X_i can be estimated by

$$\widehat{\psi}_j = \sup_{h \in H} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n \{h(X_{ij})^2 - \bar{h}^2\} \sum_{i=1}^n (Y_i - \bar{Y})^2}} \tag{2.2.6}$$

where $\bar{h}_j = n^{-1} \sum_{i=1}^n h(X_{ij})$.

The proposed method orders the estimators $\widehat{\psi}_j$ at (2.2.6) as $\widehat{\psi}_{\widehat{j}_1} \geq \widehat{\psi}_{\widehat{j}_2} \geq ... \geq \widehat{\psi}_{\widehat{j}_p}$ and take

$$\widehat{j}_1 \succeq \widehat{j}_2 \succeq ... \succeq \widehat{j}_p \tag{2.2.7}$$

to represent an empirical ranking of the predictor indices of X in order of their influence expressed through a generalized coefficient of correlation. The notation

$j \succeq j'$ represents $\widehat{\psi}_j \geq \widehat{\psi}_{j'}$.

The following assumptions are required to establish the theoretical properties of the generalized correlation ranking.

(B1) the pairs $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d

(B2) $\mathcal{H}$ is the class of polynomial functions up to but not exceeding a given degree $d \geq 1$;

(B3) $\text{var}\{h(X_{ij})\} = \text{var}\{Y_i\} = 1$ for all i and j;

(B4) for constants $\gamma, c > 0$, and sufficiently large n, $p \leq O(n^\gamma)$;

(B5) for a constant $C > 4d(\gamma + 1)$, $\sup_n \max_{j \leq p} E|X_{1j}|^C < \infty$, and $\sup_n E|Y_1|^C < \infty$.

Given constant $0 < c_1 < c_2 < \infty$, define $I_1(c_1) = \{j : |\text{cov}(X_i, Y_i)| \leq c_1\sqrt{\log n/n}\}$ and $I_2(c_2) = \{j : |\text{cov}(X_i, Y_i)| \geq c_2\sqrt{\log n/n}\}$

Theorem 2.2.2. Under above assumptions, for sufficiently small c_1 and sufficiently large c_2, in the correlation-based ranking $\widehat{j}_1 \succeq \widehat{j}_2 \succeq ... \succeq \widehat{j}_p$, with the probability converging to 1 as $n \to \infty$, all indices in $I_2(c_2)$ are listed before any of the indices in $I_1(c_1)$.

Theorem 2.2.2 shows that for predictors that have sufficiently large covariance are ranked before predictors that have smaller covariance.

2.2.3 Sure Independence Screening for GLM

Since the SIS possesses a sure independent screening property in the context of the linear model, Fan and Song (2010) proposed a more general version of the independent learning using the maximum marginal likelihood in generalized linear model. Thus, the SIS is a special case of this proposed procedure.

Assume that Y is from an exponential family with the probability density function taking the canonical form

$$f_Y(y|\theta(x)) = \exp\{y\theta(x) - b(\theta(x)) + c(y)\} \tag{2.2.8}$$

for some known functions $b(\cdot), c(\cdot)$ and $\theta(x) = \sum_{j=1}^{p} \beta_j x_j$. The mean response is $b'(\theta)$, the first derivative of $b(\theta)$ respect to θ. The dispersion parameter ϕ is set to be 1. Then, we estimate a $(p+1)$-vector of parameter $\beta = (\beta_0, ..., \beta_p)$ from the following generalized linear model:

$$E(Y|X = x) = b'(\theta(x)) = g^{-1}(\sum_{j=0}^{p} \beta_j x_j)$$

where $\mathbf{x} = \{1, ..., x_p\}^T$ is a (p+1) dimensional covariate. We assume that the covariates are standardized.

Define $\mathcal{M} = \{1 \leq j \leq p_n : \beta_j^* \neq 0\}$ be the true sparse model with nonsparsity size $|M|$. The maximum marginal likelihood estimator (MMLE) $\widehat{\beta}_j^M$, for $j = 1, ..., p_n$ is defined as the minimizer of the componentwise regression

$$\widehat{\beta}_j^M = (\widehat{\beta}_{j,0}^M, \widehat{\beta}_j^M) = \arg\min_{\beta_0, \beta_j} P_n l(\beta_0 + \beta_j X_j, Y), \tag{2.2.9}$$

where $l(\theta, Y)$ is the log likelihood for the natural parameter, $l(\theta, Y) = -[\theta Y - b(\theta) - \log c(Y)]$, and $P_n f(X, Y) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i)$. Select a submodel

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p_n : |\widehat{\beta}_j^M| \geq \gamma_n\},$$

where γ_n is a predefined threshold value.

Although the interpretation of the marginal model is biased from the joint model, the non sparse information about the joint model be passed along to the marginal model under minor conditions.

Following conditions are required to establish the sure screening property

(C1) The Fisher information,

$$I(\beta) = \mathrm{E}\{[\frac{\partial}{\partial\beta} l(X^T\beta, Y)][\frac{\partial}{\partial\beta} l(X^T\beta, Y)]^T\}$$

is finite and positive definite at $\beta = \beta_0$.

(C2) The second derivative of $b(\theta)$ is continuous and positive. There exists and $\epsilon_1 > 0$ such that for all $j = 1, \ldots, p_n$

$$\sup_{\beta \in \mathcal{B}, ||\beta - \beta_j^M|| \leq \epsilon_1} |\mathrm{E}b(\mathbf{X}_j^T \boldsymbol{\beta}) I(|X_j| > K_n)| \leq o(n^{-1})$$

(C3) For all $\beta_j \in \boldsymbol{\mathcal{B}}$, we have $\mathrm{E}(l(\mathbf{X}_j^T \boldsymbol{\beta}_j, Y) - l(\mathbf{X}_j^T \boldsymbol{\beta}_j^M, Y)) \geq V||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^M||^2$, for some positive V, bounded from below uniformly over $j = 1, \ldots, p_n$.

(C4) There exists some positive constants m_0, m_1, s_0, s_1 and α, such that for sufficiently large t,

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

(C5) $|cov(b'(\mathbf{X}^T\boldsymbol{\beta}^*), X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}$

Let $k_n = b'(K_n B + B) + m_0 K_n^\alpha / s_0$. The following theorem gives a uniform convergence result of MMLEs and the sure screening property.

Theorem 2.2.3 Suppose that conditions (C1), (C2), (C3), and (C4) hold,

(i) If $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, then for any $c_3 0$, there exists a positive constant c_4, such that

$$P(\max_{1\le j\le p_n} |\widehat{\beta}_j^M - \beta_j^M| \ge c_3 n^{-\kappa}) \le p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)\}$$

(ii) If, in addition, condition (C5) holds, then by taking $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \le c_2/2$ we have

$$P(\mathcal{M} \subset \widehat{\mathcal{M}}_{\gamma_n}) \le 1 - s_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)\},$$

where s_n is the size of non-sparse elements.

Fan and Song (2010) also discuss the controlling the false selection rates.

Under some mild conditions, it has been shown that

$$\max_{j\notin M} |\widehat{\beta}_j^M| = c_3 n^{-\kappa} \text{ for any } c_3 > 0$$

with probability approach to 1. By choosing $\gamma_n = c_5 n^{-\kappa}$, model consistency can be achieve

$$P(\widehat{M}_{\gamma_n} = M) = 1 - o(1)$$

2.2.4 Sure Independence Ranking and Screening

Zhu, Li, Li, and Zhu (2011) proposed a model free feature screening approach for ultrahigh-dimensional data, called sure independence ranking and screening (SIRS). The SIR impose a general model framework including commonly used parametric and semi parametric methods instead of a specific model.

Let Y be the response variable with support Ψ_y, and Y can be both univariate and multivariate. Let $\mathbf{x} = (X_1, \ldots, X_p)^T$ be a predictr vector. Define the notation of active predictors and inactive predictors without specifying a regression model. Consider the conditional distribution function $F(y|\mathbf{x}) = P(Y < y|\mathbf{x})$. Define two index sets:

$\mathcal{A} = \{k : F(y|\mathbf{x})$ functionally depends on X_k for some $y \in \Psi_y\}$,

$\mathcal{I} = \{k : F(y|\mathbf{x})$ does not functionally depends on X_k for some $y \in \Psi_y\}$,

where $\mathcal{A}$ is the index of active predictors and $\mathcal{I}$ is the index of inactive predictors. Consider that the conditional distribution of Y given $\mathbf{x}$ depends only through $\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}}$ where $\mathbf{x}_{\mathcal{A}}$ is an an active predictor vector

$$F(y|\mathbf{x}) = F_0(y|\boldsymbol{\beta}^T\mathbf{x}_A), \tag{2.2.10}$$

where $F_0(\cdot|\boldsymbol{\beta}^T\mathbf{x}_A)$ is an unknown conditional distribution function given $\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}}$,

Without a loss of generality, assume that $\mathrm{E}(X_k) = 0$ and $var(X_k) = 1$ for $k = 1, \ldots, p$. Define $\boldsymbol{\Omega}(y) = \mathrm{E}\{\mathbf{x}F(y|\mathbf{x})\}$. Then by the law of iterated expectations that $\boldsymbol{\Omega}(y) = \mathrm{E}[\mathbf{x}\mathrm{E}\{1(Y < y)|\mathbf{x}\}] = cov\{\mathbf{x}, 1(Y < y)\}$. Let $\Omega_k(y)$ be the kth element of

$\boldsymbol{\Omega}(y)$ and define

$$\omega_k = \mathrm{E}\{\Omega_k^2(Y)\} \quad k = 1, \ldots, p \tag{2.2.11}$$

Then ω_k is the population quantity of proposed marginal utility measure for predictor ranking.

Given a random sample $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ from $\{\mathbf{x}, Y\},$, the sample estimator of ω_k is

$$\tilde{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\{\frac{1}{n}\sum_{i=1}^{n} X_{ik}1(Y_i < Y_j)\}^2, \quad k = 1, \ldots, p. \tag{2.2.12}$$

where X_{ik} is the kth element of $\mathbf{x}_i$.

Following conditions are require for performing the sure independence ranking and screening

(D1) The following inequality condition holds uniformly for p

$$\frac{K^2\lambda_{max}\{cov(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_I^T)cov(\mathbf{x}_I, \mathbf{x}_{\mathcal{A}}^T)\}}{\lambda_{min}^2\{cov(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^T)\}} < \frac{\min_{k\in\mathcal{A}}\omega_k}{\lambda_{max}\{\boldsymbol{\Omega}_{\mathcal{A}}\}} \tag{2.2.13}$$

where $\boldsymbol{\Omega}_{\mathcal{A}} = \mathrm{E}\{\boldsymbol{\Omega}_{\mathcal{A}}(T)\boldsymbol{\Omega}_{\mathcal{A}}^T(Y)\}$, and and $\lambda_{max}\{\mathbf{B}\}, \lambda_{min}\{\mathbf{B}\}$ denote the largest and smallest eigenvalues of a matrix $\mathbf{B}$.

(D2) The linearity condition:

$$\mathrm{E}\{\mathbf{x}|\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}}\} = cov(\mathbf{x}, \mathbf{x}_{\mathcal{A}}^T)\boldsymbol{\beta}\{cov(\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}})\}^{-1}\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}} \tag{2.2.14}$$

(D3) The moment condition: there exists a positive constant t_0 such that

$$\max_{1\leq k\leq p} \mathrm{E}\{\exp(tX_k)\} < \infty \text{ for } 0 < t \leq t_0. \tag{2.2.15}$$

condition (D1) dictates the correlations among the predictors, and is the key assumption to ensure that the proposed screening procedure works properly.

Theorem 2.2.4 Under conditions (D1)-(D3), the following inequality holds uniformly for p:

$$\max_{k\in\mathcal{I}} \omega_k < \min_{k\in\mathcal{A}} \omega_k \tag{2.2.16}$$

Theorem 2.2.4 shows that the proposed marginal utility measure ω_k for inactive predictors are smaller than ω_k for active predictors.

Theorem 2.2.5 In addition to the conditions in Theorem 1.2.3, assume that $p = o\{exp(an)\}$ for any fixed $a > 0$. Then for any $\epsilon > 0$, there exists a sufficiently small constant $s_\epsilon \in (0, 2/\epsilon)$ such that

$$P(\sup_{k=1,\ldots,p} |\hat{\omega}_k - \omega_k| > \epsilon) \leq 2p\exp\{n\log(1-\epsilon s_\epsilon/2)/3\}$$

In addition, if we write $\delta = \min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{I}} \omega_k$, then there exists a sufficiently small constant $s_{\delta/2} \in (0, 4/\delta)$ such that

$$P(\max_{k\in\mathcal{I}} \hat{\omega}_k < \min_{k\in\mathcal{A}} \hat{\omega}_k) \geq 1 - 4p\exp\{n\log(1-\delta s_{\delta/2}/4)/3\}$$

Theorem 2.2.5 demonstrates that the proposed marginal utility measure estimate $\hat{\omega}_k$ ranks active predictors ahead of inactive ones with the probability approaches to 1.

Compare to the hard thresholding rule for selecting a submodel by Fan and Lv (2008), Zhu, Li, Li, and Zhu (2011) suggested a soft thresholding rule based on adding auxiliary variables. Generate randomly and independently d auxiliary variables where $\mathbf{z} \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$, and $\mathbf{z}$ is independent of both $\mathbf{x}$ and Y. Since $\mathbf{z}$ is a vector of inactive predictors, we have $\min_{k\in A} \omega_k > \max_{k\in\mathbf{z}} \omega_k$. Define $C_d = \max_{k\in\mathbf{z}} \widehat{\omega}_k$, which can be viewed as a benchmark that separates the active predictors from the inactive ones. This leads to the selection

$$\hat{\mathcal{A}}_1 = \{k : \hat{\omega}_k > C_d\} \tag{2.2.17}$$

Theorem 2.2.6 Assume that the inactive predictors $\{X_j, j \in \mathcal{I}\}$ and the auxiliary variables$\{Z_j, j = 1, \ldots, d\}$are exchangeble in the sense that both the inactive and auxiliary variables are equally likely to be recruited by the soft thresholding procedure. Then

$$P(|\hat{\mathcal{A}} \cap \mathcal{I}| \geq r) \leq (1 - \frac{r}{p+d})^d$$

where $|\cdot|$ denotes the cardinaility of a set.

2.2.5 Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models

Fan, Feng, and Song (2011) proposed nonparametric independence screening in ultrahigh dimensional additive models (NIS). By using a more flexible class of nonparametric models, this procedure increases the flexibility of the ordinary linear model proposed by Fan and Lv (2008).

Suppose that we have a random sample $(\mathbf{X}_1, Y_2), \ldots, (\mathbf{X}_n, Y_n)$, from the population

$$Y = m(\mathbf{X}) + \epsilon = \sum_{j=1}^{p} m_j(X_j) + \epsilon \tag{2.2.18}$$

where $\mathbf{X}_i = (X_1, ..., X_p)^T$, ϵ is the random error with conditional mean 0. We assume that the true regression function admits the additive structure. $m_j(X_j)$ also assumes to have mean 0 for $j = 1, \ldots, p$. Define the set of true predictors as $\mathcal{M} = \{j : \mathrm{E}m_j(X_j)^2 > 0\}$. We consider the following p marginal nonparametric regression problems:

$$\min_{f_j \in L_2(P)} E(Y - f_j(X_j))^2 \tag{2.2.19}$$

The minimizer of (2.2.19) is $f_j = E(Y|X_j)$. This method ranks the marginal utility of covariates according to $Ef_j^2(X_j)$ and select a submodel by predefined thresholding. For an estimate of the marginal nonparametric regression, a B-spline basis is used. Let S_n be the space of polynomial spline of degree l and $\{\Psi_{jk}, k = 1, ..., d_n\}$ denote a normalized B-spline basis with $||\Psi_{jk}||_\infty \leq 1$. For any $f_{nj} \in S_n$, we have,

$$f_{nj}(x) = \sum_{k=1}^{d_n} \beta_{jk} \Psi_{jk}(x), \quad 1 \leq j \leq p \tag{2.2.20}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_n}$. Under some smoothness conditions, the nonparametric projections $\{f_j\}_{j=1}^{p}$ can be well approximated by functions in S_n. The estimate of the marginal regression is

$$\min_{f_{nj} \in S_n} P_n(Y - f_{nj}(X_j))^2 = \min_{\beta_j \in R^{d_n}} P_n(Y - \Psi_j^T \beta_j)$$

where Ψ_j denotes the d_n-dimensional basis functions and $P_n g(X, Y)$ is the sample average of $\{g(X_i, Y_i)\}_{i=1}^n$. Then select a set of variables

$$\widehat{M}_{v_n} = \{1 \leq j \leq p : ||\widehat{f}_{nj}||_n^2 \geq v_n\},$$

where $||\widehat{f}_{nj}||_n^2 = \frac{1}{n}\sum_{i=1}^n \widehat{f}_{nj}(X_{ij})^2$

For establishing the theoretical property, following conditions are required

(E1) The nonparametric marginal projections $\{f_j\}$ for $j = 1, \ldots, p$ belong to a class of functions F, whose rth derivative $f^{(r)}$ exists and is Lipschitz of order α,

$F = \{f(\cdot) : |f^{(r)}(s) - f^{(r)}(t)| \leq K|s-t|^{|\alpha|} \text{ for } s, t \in [a, b]\}$

for some positive constant K, where r is a nonnegative integer and $\alpha \in (0, 1]$ such that $d = r + \alpha > 0.5$

(E2) The marginal density function g_j of X_j satisfies $0 < K_1 \leq g_j(X_j) \leq K_2 < \infty$ on $[a, b]$ for $1 \leq j \leq p$ for some constants K_1 and K_2.

(E3) $\min_{j \in M} \mathrm{E}\{\mathrm{E}(Y|X_j)^2\} \geq c_1 d_n n^{-2\kappa}$, for some $0 < \kappa < d/(2d+1)$ and $c_1 > 0$.

(E4) $||m||_\infty < B_1$ for some positive constant B_1, where $||\cdot||_\infty$ is the sup norm.

(E5) The random error ϵ_i are i.i.d with conditional mean 0 for $i = 1, \ldots, n$, and for any $B_2 > 0$, there exists a positive constant B_3 such that $\mathrm{E}[\exp(B_2|\epsilon_i|)|\mathbf{X}_i] < B_3$.

(E6) There exist positive constants c_1 and $\psi \in (0, 1)$ such that $d_n^{-2d-1} \leq c_1(1-\psi)n^{-2\kappa}/C_1$.

Lemma 2.2.7 Under conditions (E1) - (E3), we have

$$\min_{j \in M} ||f_{nj}||^2 \geq c_1 \psi d_n n^{-2\kappa}$$

Theorem 2.2.8 Under conditions (E1), (E2), (E4), and (E5),

(i) For any $c_2 > 0$, there exist some positive constants c_3 and c_4 such that

$$P(\max_{1\leq j\leq p_n} | \, ||\widehat{f}_{nj}||_n^2 - ||f_{nj}||^2| > c_2 d_n n^{-2\kappa})$$
$$\leq p_n d_n \{(8+2d_n) exp(-c_3 n^{1-4\kappa} d_n^{-3}) + 6d_n \exp(-c_4 n d_n^{-3})\}$$

(ii) In addition, under condition (E3) and (E6), by taking $v_n = c_5 d_n n^{-2\kappa}$ with $c_5 \leq c_1\psi/2$, we have

$$P(M \subset \widehat{M}_{v_n}) \geq 1 - s_n d_n \{(8+2d_n)\exp(-c_3 n^{1-4\kappa} d_n^{-3}) + 6d_n \exp(-c_4 n d_n^{-3})\}$$

Fan, Feng, and Song (2011) also discuss about the controlling the false selection rate. The ideal case for the vanishing false positive rate is that

$$\max_{j\notin M} ||f_{nj}||^2 = o(d_n n^{-2\kappa})$$

so there is a gap between active and inactive predictors under the marginal nonparametric screening. By theorem 1.2.7 (i), if theorem tends to 0 with probability tending to 1 that

$$\max_{j\notin M} ||\hat{f}_{nj}||^2 \leq c_2 d_n n^{-2\kappa} \text{ for any } c_2 > 0 \tag{2.2.21}$$

Thus, by the choice of v_n in theorem 1(ii), we can achieve model selection consistency

$$P(\widehat{M}_{v_n} = M) = 1 - o(1) \tag{2.2.22}$$

2.2.6 Feature Screening via Distance Correlation Learning

Li, Zhong, and Zhu (2012b) proposed a feature screening procedure for ultrahigh dimensional data based on distance correlation (DC-SIS). The proposed DC-SIS is the model free procedure which can handle grouped predictors and multivariate responses.

Székely *et al.* (2009) defined the distance covariance between **u** and **v** with finite first moments to be the nonnegative number dcov(u, v) given by

$$dcov^2(\mathbf{u}, \mathbf{v}) = \int_{R^{d_u+d_v}} ||\Phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \Phi_{\mathbf{u}}(\mathbf{t})\Phi_{vv}(\mathbf{s})||^2 w(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s} \tag{2.2.23}$$

where d_s and d_v are the dimensions of **u** and **v**, and $\Phi_u(t)$ and $\Phi_v(s)$ be the respective characteristic functions of the random vectors **u** and **v**, and $\varphi_{u,v}(t, s)$ be the joint characteristic function of **u** and **v**, and $w(t, s) = \{c_{d_u} c_{d_v} ||t||_{d_u}^{1+d_u} ||s||_{d_v}^{1+d_v}\}^{-1}$ Székely *et al.* (2009) stated that

$$dcov^2(\mathbf{u}, \mathbf{v}) = S_1 + S_2 - 2S_3 \tag{2.2.24}$$

where S_1, S_2, and S_3 are defined as

$$\begin{aligned} S_1 &= \mathrm{E}\{||\mathbf{u} - \tilde{\mathbf{u}}||_{d_u} ||\mathbf{v} - \tilde{\mathbf{v}}||_{d_v}\} \\ S_2 &= \mathrm{E}\{||\mathbf{u} - \tilde{\mathbf{u}}||_{d_u}\}\mathrm{E}\{||\mathbf{v} - \tilde{\mathbf{v}}||_{d_v}\} \\ S_3 &= \mathrm{E}\{\mathrm{E}(||\mathbf{u} - \tilde{\mathbf{u}}||_{d_u}|\mathbf{u})\mathrm{E}(||\mathbf{v} - \tilde{\mathbf{v}}||_{d_v}|\mathbf{v})\} \end{aligned} \tag{2.2.25}$$

Then the estimates of S_1, S_2, and S_3 are using the usual moment estimation,

$$\begin{aligned}\hat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{u}-\tilde{\mathbf{u}}||_{d_u}||\mathbf{v}-\tilde{\mathbf{v}}||_{d_v} \\ \hat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{u}-\tilde{\mathbf{u}}||_{d_u}||\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{v}-\tilde{\mathbf{v}}||_{d_v} \\ \hat{S}_3 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}||\mathbf{u}-\tilde{\mathbf{u}}||_{d_u}||\mathbf{v}-\tilde{\mathbf{v}}||_{d_v}\end{aligned} \tag{2.2.26}$$

The estimate of distance correlation between $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$\widehat{dcorr}(\mathbf{u},\mathbf{v}) = \frac{\widehat{dcov}(\mathbf{u},\mathbf{v})}{\sqrt{\widehat{dcov}(\mathbf{u},\mathbf{u})\widehat{dcov}(\mathbf{v},\mathbf{v})}} \tag{2.2.27}$$

Let $\mathbf{y} = (Y_1, \ldots, Y_q)^T$ be the response vector with support Ψ_y and $\mathbf{x} = (X_1, \ldots, X_p)^T$ be the predictor vector. Without specifying a regression model, we define the index set of the active and inactive predictors by

$\mathcal{A} = \{k : F(\mathbf{y}|\mathbf{x})$ functionally depends on X_k for some $y \in \Psi_y\}$,

$\mathcal{I} = \{k : F(\mathbf{y}|\mathbf{x})$ does not functionally depends on X_k for some $y \in \Psi_y\}$,

We further write $x_D = \{X_k : k \in D\}$ and $x_I = \{X_k : k \in I\}$,and refer to x_D as an active predictor vector and its complement x_I as an inactive predictor vector.

The DC-SIS is a model free procedure that it allows for arbitrary regression relationship of y onto x, both linear or nonlinear and also permits univariate and multivariate response, including continuous, discrete and categorical. Select a set of a important predictors $\widehat{\omega}_k = \widehat{dcorr}^2(X_k, \mathbf{y})$. That is, define a submodel as

$$\widehat{\mathcal{D}} = \{k : \hat{\omega}_k \geq cn^{\kappa}, \text{ for } 1 \leq k \leq p\} \tag{2.2.28}$$

where c and κ are pre-specified threshold values.

Following conditions are required for establishing the sure screening property for the DC-SIS

(F1) Both x and y satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0 < s \leq 2s_0$,

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

(F2) The minimum distance correlation of active predictors satisfies

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

Condition (F1) follows when x and y are bounded uniformly, or when they have multivariate normal distribution. Condition (F2) requires that the marginal distance correlation of active predictors cannot be too smal.

Theorem 2.2.9 Under condition (F1), for any $0 < \gamma < 1/2 - \kappa$, there exists positive constant $c_1, c_2 > 0$ such that

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

Under condition (F1) and (F2)

$$Pr(D \subseteq \widehat{D}) \geq 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n \exp(-c_2 n^{\gamma})])$$

where s_n is the cardinality of D. The sure screening property holds for the DC-SIS.

2.2.7 Principled sure independence screening for Cox models with ultra-high-dimensional covariates

Zhao and Li proposed a sure independence screening for Cox model. This procedure provides a tool for censored data in the survival setting. This procedure assumes that the underlying survival times follow the Cox model with the true hazard function. To perform the screening, the marginal Cox model is fitted for each Z_{ij}, namely $\lambda_0^*(x)\exp(\beta Z_{ij}$. $N_i(t) = I(X_i \leq t, \delta_i = 1)$ be independent counting processes for each subject i and $Y_i(t) = I(X_i \geq t)$ be the at-risk processes. For $\kappa = 0, 1, \ldots,$ define

$$S_j^{(k)}(\beta, x) = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}^k Y_i(x)\exp(\beta Z_i), \quad s_j^{(k)}(x) = E\{S^{(k)}(x)\} \tag{2.2.29}$$

Then the maximum marginal partial likelihood estimator $\hat{\beta}_j$ solves the estimating equation

$$U_j(\beta) = \sum_{i=1}^{n}\int_0^\tau \{Z_{ij} - \frac{S_j^{(1)}(\beta, x)}{S_j^{(0)}(\beta, x)}\} dN_i(x) = 0. \tag{2.2.30}$$

Finally, let β_{0j} be the solution to the limiting estimation equation

$$u_j(\beta) = \int_0^\tau \{s_j^{(1)}(x) - \frac{S_j^{(1)}(\beta, x)}{S_j^{(0)}(\beta, x)}\} dN_i(x) = 0. \tag{2.2.31}$$

Define the information matrix to be $I_j(\beta) = -\delta U_j/\delta\beta$ at $\hat{\beta}_j$. Denote the final screened model by $\hat{M} = \{j : I_j(\hat{\beta}_j)^{1/2}|\hat{\beta}_j| \geq \gamma_n\}$. The thresholding value γ_n such that we can achieve the sure screening property while controlling the false positive rate. If the true model M has size s then the expected false positive rate can be written as

$$E(\frac{|\hat{M} \cap M^c|}{|M^c|}) = \frac{1}{p-s} \sum_{j \in M^c} P\{I_j(\hat{\beta}_j^{1/2}|\hat{\beta}| \geq \gamma_n\}.$$

Since $I_j(\hat{\beta}_j)^{1/2}\hat{\beta}_j$ has an asymptotically standard normal distribution, γ_n controls the expected false positive rate at $2\{1-\Phi(\gamma_m)\}$. If b is the number of false positive that can be allowed, then the false positive rate is $b/(p-s)$. Since s is unknown, choose the conservative $\gamma_n = \Phi^{-1}\{1-b/2p_n\}$, so the expected false positive rate is less than $b/(p-s)$.

Following conditions are required for establishing the theoretical property.

(G1) There exists a neighborhood B of β_{0j} such that for each $t < \infty$,

$$\sup_{x \in [0,t], \beta \in B} |S_j^{(0)}(\beta, x) - s_j^{(0)}(\beta, x)| \to 0$$

(G2) For each $t < \infty$ and $j = 1, \ldots, p_n$, $\int_0^t s_j^{(2)}(x)dx < \infty$.

(G3) The true parameter vector α_0 belongs to a compact set such that each component α_{0j} is bounded by a constant $A > 0$. Furthermore, $||\alpha_0||_1$ is bounded by a constant $L > 0$.

(G4) $\lambda_0(\tau)$ is bounded by a positive constant.

(G5) There is some constant $C > 0$ such that $n^{-1}|U_j(\hat{\beta}_j) - U_j(\beta_{0j})| \leq C|\hat{\beta} - \beta_{0j}|$ for all $j = 1, \ldots, p_n$.

(G6) The Z_{ij} are independent of time and bounded by a constant $T > 0$, and $E(Z_{ij}) = 0$ for all j.

(G7) If $F_T(x; Z_i)$ is the cumulative distribution function of T_i given Z_i, then for constants $c_1 > 0$ and $\kappa < 1/2$, $\min_{j \in M} |cov[Z_{ij}, EF_T(C_i; Z_i)|Z_i]| \leq c_1 n^{-\kappa}$.

(G8) The $Z_{ij}, j \in M^c$ are independent of the $Z_{ij}, j \in M$? and of C_i.

Theorem 2.2.10 Under assumptions (G1) $\sim$ (G8), if we choose $\gamma_n = \Phi(1 - b/2p)$, then then for $\kappa < 1/2$ and $log(p) = O(n^{1/2-\kappa})$, there exists a constant $c_1 > 0$ such that

$$P(M \subseteq \hat{M}) \geq 1 - s\exp(-c_1 n^{1-2\kappa})$$

This procedure possesses the sure independent property. **Theorem 2.2.11** Under some mild assumptions, if we choose $\gamma_n = \Phi(1 - b/2p)$, then then for $\kappa < 1/2$ and $log(p) = O(n^{1/2-\kappa})$, there exists a constant $c_2 > 0$ such that

$$E(\frac{|\hat{M} \cap M^c|}{|M^c|}) \leq b/p + c_2 n^{-1/2}$$

This procedure controls the false positive rate.

2.2.8 Robust Rank Correlation Based Screening

Li, Peng, Zhang, and Zhu (2012a) proposed a procedure that is based on the Kendall τ correlation coefficient between response and predictor variable. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$ is an n-vector of response and $\mathbf{X}$ is $n \times p$ random matrix. Let $\omega =$

$(\omega_1, \ldots, \omega_p)^T$ being

$$\omega_k = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} I(X_{ik} < X_{jk}) I(Y_i < Y_j) - 1/4, \; k = 1, \ldots, p, \tag{2.2.32}$$

a p-vector of the one fourth of the Kendall τ between Y and X_k. Then we rank ω_k by its magnitude and select a submodel

$$\hat{M}_{\gamma_n} = \{1 \leq k \leq p : |\omega_k| > \gamma_n\}$$

Since the Kendall τ is robust against heavy-tailed distribution, this procedure is more robust than the SIS. We assume that

$$M_* = \{1 \leq k \leq p : \beta_k \neq 0\}$$

The following conditions are required for the sure screening property.

(H1) As n approaches to infinity, the dimension of $\mathbf{X}$ satisfies $p = O(exp(n^\delta))$ for some $\delta \in (0,1)$, satisfying $\delta + 2\kappa < 1$ for any $\kappa \in (0, 1/2)$.

(H2) $c_{M_*} = \min_{k \in M_*} E|X_{1k}|$ is a positive constant and free of p.

(H3) The predictors $\mathbf{X_i}$ and the error ϵ_i, for $\imath = 1, \ldots, n$, are independent.

Theorem 2.2.12 Under the condition (H2) and the marginal symmetrical conditions, we have

(i) $E(\omega_k) = 0$ if and only if $\rho_k = 0$.

(ii) If $|\rho_k| > c_1 n^{-\kappa}$ for $k \in M_*$ with a positive constant c_1, then there exists a positive constant c_2 such that $\min_{k \in M_*} |E(\omega_k)| > c_2 n^{-\kappa}$.

Theorem 2.2.13 Under the conditions (H1) $\sim$ (H3) for some $0 < \kappa < 1/2$ and $c_3 > 0$, there exists a positive constant $c_4 > 0$ such that

$$P(\max_{1 \leq j \leq p} |\omega_j - E(\omega_j)| \geq c_3 n^{-\kappa}) \leq p\{exp(-c_4 n^{1-2\kappa})\}.$$

Furthermore, by taking $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \leq c_2/2$, if $|\rho_k| > c_1 n^{-\kappa}$ for $j \in M_*$, we have

$$P(M_* \subset \hat{M}_{\gamma_n}) \geq 1 - 2|M_*|\{\exp(-c_4 n^{1-2\kappa})\}$$

Thus, the sure screening property holds for the RRCS.

Chapter 3

Independence Screening via the Variance of the Regression Function

3.1 Introduction

In this chapter, we propose a new nonparametric screening procedure based on the variance of the regression function (RV-SIS). We show that the RV-SIS possesses the sure screening property that is defined by Fan and Lv (2008). We conduct numerical simulation studies to compare the performance of the RV-SIS to other procedures.

This chapter is organized as follows. In section 3.2, we introduce the RV-SIS for ultra high dimensional data, and we establish the sure screening property for the RV-SIS. In section 3.3, the result of the numerical studies is presented. Technical proofs are given in section 3.4.

3.2 Nonparametric Independence Screening via the Variance of the Regression Function

3.2.1 Preliminaries

Consider a random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of iid $(p+1)$-dimensional random vectors, where Y_i is univariate and $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})^T$ is a p-dimensional, $i = 1, \ldots, n$. Let $m(\mathbf{X}) = E(Y|\mathbf{X})$ and write

$$Y = m(\mathbf{X}) + \epsilon, \tag{3.2.1}$$

where $\epsilon = Y - m(\mathbf{X})$. For $k = 1, \ldots, p$, we consider p marginal nonparametric regression functions

$$m_k(x) = \mathrm{E}(Y_i | X_{ki} = x) \tag{3.2.2}$$

of Y on each variable X_k, and define the set of active and inactive predictors by

$$\mathcal{D} = \{k : m_k(x) \text{ is not a constant function}\}, \quad \mathcal{D}^c = \{1, \ldots, p\} - \mathcal{D}, \tag{3.2.3}$$

respectively. The proposed screening procedure relies on ranking the significance of the p covariates according to the magnitude of the variance of their respective marginal regression functions,

$$\sigma^2_{m_k} = \mathrm{var}(m_k(x)) \quad \text{for} \quad k = 1, \ldots, p. \tag{3.2.4}$$

Note that $\sigma^2_{m_k} > 0$ for $k \in \mathcal{D}$, while $\sigma^2_{m_k} = 0$ for $k \in \mathcal{D}^c$, making $\sigma^2_{m_k}$ a natural quantity to discriminate between the two classes of predictors. In addition, the variance of the regression function appears as the mean shift, under local alternatives, of the procedure for testing the significance of a covariate proposed in Wang, Akritas and Van Keilegom (2008). This suggests that $\sigma^2_{m_k}$ is the best quantity to discriminate between the two classes of predictors.

If $\widehat{m}_k$ denotes an estimator of m_k, $\sigma^2_{m_k}$ can be estimated by the sample variance of $\widehat{m}_k(X_{k1}), \ldots,\ \widehat{m}_k(X_{kn})$. The methodology described here works with any type of nonparametric estimator of m_k, but the theory has been developed for Nadaraya-Watson type estimators.

For a kernel function $K(\cdot)$ and bandwidth h, set $\widehat{m}_k(X_{ki}) = \sum_{j=1}^n Y_j W_{k;i,j}$, where $W_{k;i,j} = K(\frac{X_{kj}-X_{ki}}{h})/\sum_{j=1}^n K(\frac{X_{kj}-X_{ki}}{h})$, and

$$\widetilde{S}^2_{m_k} = \frac{1}{n}\sum_{i=1}^n \left(\hat{m}(X_{ki}) - \frac{1}{n}\sum_{l=1}^n \hat{m}(X_{kl}) \right)^2 \tag{3.2.5}$$

for the estimator of $\sigma^2_{m_k}$. The bandwidth will be of the order $h = cn^{-1/5}$, throughout this paper. The RV-SIS estimates $\mathcal{D}$ by

$$\widehat{\mathcal{D}} = \{k : \widetilde{S}^2_{m_k} \geq C_d, \quad \text{for } 1 \leq k \leq p\} \tag{3.2.6}$$

for some threshold parameter C_d. Thus, the RV-SIS procedure reduces the dimension of covariate vector from p to $|\hat{\mathcal{D}}|$, where $|\cdot|$ refers the cardinality of a set. The choice of C_d, which defines the RV-SIS procedure, is discussed below.

3.2.2 Thresholding Rule

We adopt the idea of the soft thresholding rule by Zhu *et al.* (2011) as a method for choosing the threshold parameter C_d. This method consists of randomly generating a vector $\mathbf{Z} = (X_{p+1}, \ldots, X_{p+d})$ of d auxiliary random variables from the uniform distribution between (0, 1), $X_{p+i} \sim \text{Unif}(0,\,1)$ for $i = 1, \ldots, d$, that are independent of both $\mathbf{X}$ and $\mathbf{Y}$. By design, the auxiliary variables are inactive predictors. The soft thresholding rule chooses the threshold parameter as

$$C_d = \max_{j \in \mathcal{B}} \tilde{S}^2_{m_j}, \tag{3.2.7}$$

where $\mathcal{B} = \{p+1, \ldots, p+d\}$ denotes the set of indices of the d auxiliary variables.

Theorem 3.2.1 provides an upper bound on the probability of selecting inactive predictors from using the proposed soft thresholding rule provided the following *exchangeability condition* holds.

Exchangeability Condition: Let $k \in \mathcal{D}^c$ and $j \in \mathcal{B}$. Then, the probability that $\widetilde{S}^2_{m_k}$ is greater than $\widetilde{S}^2_{m_j}$ is equal to the probability that $\widetilde{S}^2_{m_k}$ is less than $\widetilde{S}^2_{m_j}$.

Theorem 3.2.1. Under the exchangeability condition, for any integer $r \in (0, p)$ we have

$$P(|\widehat{\mathcal{D}} \cap \mathcal{D}^c| \geq r) \leq \left(1 - \frac{r}{p+d}\right)^d. \tag{3.2.8}$$

3.2.3 Sure Screening Properties

In this section, we show the RV-SIS possesses the sure independent screening property. The following conditions are required for technical proofs:

(C1) There exists positive constants t, C_1 and C_2 such that,

(a) $\max\limits_{1 \le k \le p} \mathrm{E}\{\exp(t|Y_j - m_k(X_{kj})|)\} < C_1 < \infty$,

(b) $\max\limits_{1 \le k \le p} \mathrm{E}(\exp(t(X_{ki} - X_{kj})^2)) < C_2 < \infty$

(C2) The kernel $K(\cdot)$ has bounded support, is symmetric, and is Lipschitz continuous, i.e, it satisfies, for some $\Lambda_1 < \infty$ and for all $u, u' \in \mathbb{R}$,

$$|K(u) - K(u')| \le \Lambda_1 |u - u'|.$$

(C3) If $f_k(x)$ denotes the marginal density of the kth predictor, we have

$$\sup_x |x|^s E(|Y||X_k = x) f_k(x) \le B < \infty \text{ for some } s \ge 1$$

$\sup_x f_k(x) < \infty$, $\inf\limits_x f_k(x) > 0$, and $f_k(x)$ is uniformly continuous, for all $k = 1, \ldots, p$.

(C4) The conditional expected value $m_k(\cdot)$ is a Lipschitz continuous for all $k = 1, \ldots, p$, that is for some $\Lambda_2 < \infty$ and for all $u, u' \in \mathbb{R}$,

$$|m_k(u) - m_k(u')| \le \Lambda_2 |u - u'|.$$

(C5) For some constants $c > 0$ and $0 < \kappa < 2/5$,

$$\min_{k \in \mathcal{D}} \sigma^2_{m_k} \ge c n^{-\kappa} + C_d,$$

where C_d is defined in (3.2.7).

In words, Condition (C1) requires that the moment generating functions of the absolute value of the error terms of the marginal regressions and the square difference between two covariates, is finite at least for some $t > 0$. Conditions (C2) and (C3) are standard conditions for establishing uniform convergence rates of needed for the kernel density estimator. Condition (C5) sets a lower bound on the variance of the marginal regression functions of the active predictors.

Theorem 3.2.2. Let $\sigma^2_{m_k}$, $\widetilde{S}^2_{m_k}$, $\mathcal{D}$, $\widehat{\mathcal{D}}$, be defined in (3.2.4), (3.2.5), (3.2.3) and (3.2.6), respectively.

1. Under condition (C1) $\sim$ (C4) for any $0 < \kappa < 2/5$ and $0 < \gamma < 2/5 - \kappa$, there exists positive constants c, c_1, and c_2 such that,

   $$P(\max_{1 \le k \le p} |\widetilde{S}^2_{m_k} - \sigma^2_{m_k}| \ge cn^{-\kappa}) \le O(p[n \exp(-c_1 n^{4/5-2(\gamma+\kappa)}) + n^2 \exp(-c_2 n^{\gamma})])$$

2. Under condition (C1) $\sim$ (C5), $c, c_1,\ c_2, \gamma$ and κ as in part 1,

   $$P(\mathcal{D} \subseteq \widehat{\mathcal{D}}) \ge 1 - O(|\mathcal{D}|[n \exp(-c_1 n^{4/5-2(\gamma+\kappa)}) + n^2 \exp(-c_2 n^{\gamma})]),$$

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

The second part of Theorem 3.2.2 shows that the screened submodel includes all active predictors with the probability approaches to 1 with an exponential rate.

3.3 Numerical Studies

3.3.1 Simulation Studies

Here we present the result of several simulation studies comparing performance of the SIS, DC-SIS, NIS, RRCS and RV-SIS methods. In all cases, $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ comes from a multivariate normal with mean zero and covariance $\Sigma = (\sigma_{ij})_{p\times p}$, and $\epsilon \sim N(0,1)$. We use three different covariance matrices: (i) $\sigma_{ij} = 0.5^{|i-j|}$, (ii) $\sigma_{ij} = 0.8^{|i-j|}$, and (iii) $\sigma_{ij} = 0.5$. We set the dimension of covariates p to be 2000 and the sample size n to be 200. We replicate the experiment 500 times and base the comparisons on the following three criteria

R1: The $5\%, 25\%, 50\%, 75\%, 95\%$ quantiles of the minimum model size that includes all active covariates.

R2: The proportion of times each individual active covariate is selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$.

R3: The proportion of times all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$.

We consider the following four models:

3.(a) $Y = 2X_1 + 0.5X_2 + 3 \cdot \mathbb{1}\{X_{12} < 0\} + 2X_{22} + \epsilon$

3.(b) $Y = 1.5X_1 \cdot X_2 + 3 \cdot \mathbb{1}\{X_{12} < 0\} + 2X_{22} + \epsilon$

3.(c) $Y = 2\cos(2\pi X_1) + 0.5X_2^2 + 3 \cdot \mathbb{1}\{X_{12} < 0\} + 2X_{22} + \epsilon$

3.(d) $Y = 2\cos(2\pi X_1)X_2^2 + 3X_{12} + 2\exp(\mathbb{1}\{X_{22} < 0\}) + \epsilon$

All models include an indicator variable. Model 3.(a) is linear, Model 3.(b) includes an interaction term, Model 3.(c) is additive but nonlinear, and Model 3.(d) is nonlinear with an interaction term.

Tables 3.1, 3.2 and 3.3 present the simulation results for R1 using each of the above models with $\sigma_{ij} = 0.5^{|i-j|}, \sigma_{ij} = 0.8^{|i-j|}$, and $\sigma_{ij} = 0.5$, respectively. Tables 3.4, 3.5 and 3.6 presents the simulation results for R2 and R3 with $\sigma_{ij} = 0.5^{|i-j|}, \sigma_{ij} = 0.8^{|i-j|}$, and $\sigma_{ij} = 0.5$, respectively.

These results show that the comparisons in term of the three criteria are similar. All procedures perform worse when we use the equal covariance matrix, $\sigma_{ij} = 0.5$. SIS and RRCS perform rather poorly except in Model 3.(a) where all methods have similar performance. For Model 3.(b), NIS performs slightly better than RV-SIS, while RV-SIS performs somewhat better than DC-SIS when $\sigma_{ij} = 0.8^{|i-j|}$, considerably better when $\sigma_{ij} = 0.5^{|i-j|}$, and significantly better when $\sigma_{ij} = 0.5$. In Models 3.(c) and 3.(d) DC-SIS and NIS have similar performance but RV-SIS performs considerably better than either of them.

Finally, Table 3.7 presents the execution time, in seconds, of the DC-SIS, NIS and RV-SIS for Model 3.(d). The RV-SIS procedure takes significantly less time than the DC-SIS and slightly less time than the NIS.

3.3.2 Thresholding Simulation

In this section we use simulations to compare the soft thresholding rule to the hard thresholding approaches for selecting the submodel. We consider following three models relating the response Y to covariates $X_1, X_2, \ldots, X_p$, where $p = 2,000$:

3.(e) $Y = c_1X_1 + \ldots + c_{25}X_{25} + \epsilon$

3.(f) $Y = c_1X_1 + \ldots + c_{10}X_{10} + \epsilon$

3.(g) $Y = c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4 + c_5X_5 + \epsilon,$

where the covariate vector has the p-variate normal distribution with mean zero and covariance $\Sigma = (0.5^{|i-j|})_{p\times p}$, $\epsilon \sim N(0,1)$, and the coefficients $c_1, \ldots, c_{25}$ were randomly generated from the uniform distribution between (1, 2.5), and kept fixed throughout the simulation. From each of these models, we generated 500 data sets of size $n = 200$.

For the soft thresholding approach, we randomly generate the auxiliary variable $\mathbf{Z} = (X_{2001}, \ldots,\ X_{3000})$, where the X_{p+i} are independent Unif(0, 1). For the hard thresholding we consider three model sizes: $d_1 = [n/\log n] = 37, d_2 = [2n/\log n] = 75, d_3 = [3n/\log n] = 113$. The two approaches are compared in terms of the proportion of each active covariate is selected. We also record the 5%, 25%, 50%, 75% and 95% quantiles of the submodel size using the soft thresholding rule.

The 5%, 25%, 50%, 75% and 95% quantiles of the submodel size using the soft thresholding rule for Models 3.(e), 3.(f) and 3.(g), are presented in Table 3.9. The proportion that each of the active covariates is selected with the different approaches for Models 3.(e), 3.(f) and 3.(g) are shown in Tables 3.6, 3.7 and 3.8, respectively.

From Table 3.9 it is seen that all percentiles decrease as the number of active covariates decreases; this is a nice feature of the soft thresholding approach. Also, for all models, the median submodel size falls between d_1 and d_2, but is always closer to d_1. Regarding the proportion that each active predictor is included in the submodel, Tables 3.6 and 3.7 show that soft thresholding outperforms hard thresholding with d_1 in Model(e), but does slightly worse in Model(f); hard thresholding with d_2 and d_3 outperform soft thresholding. Finally, Table 3.8 shows that all active predictors were selected 100% of the time by all approaches.

3.3.3 A Real Data Example

Here we apply the DC-SIS, NIS and RV-SIS methods to identify the most influential genes for over-expression of a G protein-coupled receptor (Ro1) in mice in the Cardiomyopathy microarray dataset (Segal, Dahlquist, and Conklin, 2003). In this data set, which has also been used in Hall and Miller (2009), and Li *et al.* (2012b), $n = 40$ and $p = 6,319$, with the covariates corresponding to expression levels of different genes. Figure 3.1 shows the scatterplots of the expression levels of two genes versus Ro1, with fitted cubic spline curves. Because these curves, which are typical for most genes, suggest nonlinear effects, we did not apply SIS to this data.

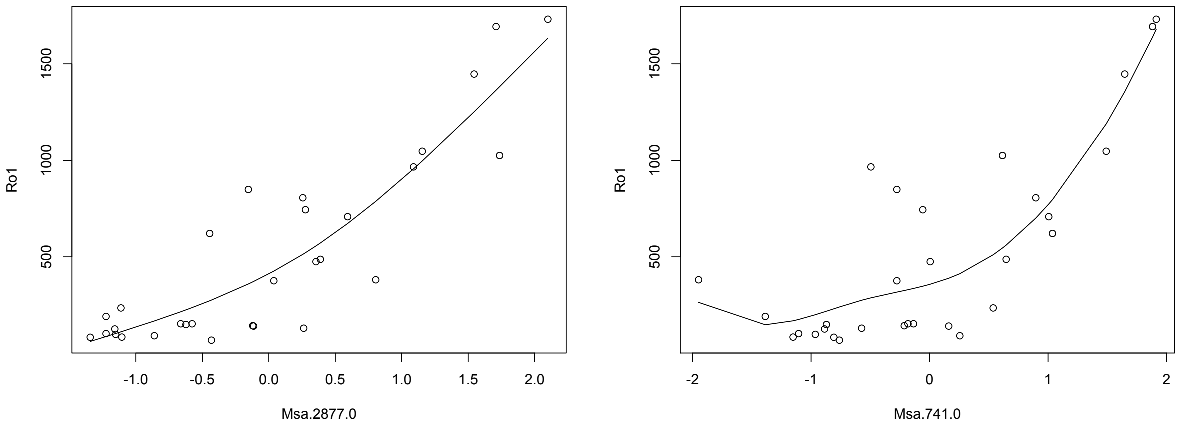

Figure 3.1. The spline curve of Msa.2877.0 and Msa.741.0

The top two most influential genes identified by RV-SIS, DC-SIS and NIS are (Msa.2877.0, Msa.741.0), (Msa.2134.0, Msa. 2877.0) and (Msa.2877.0, Msa.1166.0), respectively. To compare the models chosen by the three methods, we fit a semiparametric single index model (SIM)

$$Y = g_k(\beta_1 X_{k1} + \beta_2 X_{k2}) + \epsilon \qquad \text{for k= 1, 2, 3,}$$

where $(X_{k1}, X_{k2}),\ k = 1, 2, 3$, are the top two variables chosen by RV-SIS, DC-SIS and NIS, respectively, and use the nonparametric coefficient of determination, R^2; see Doksum and Samarov (1995). The R^2-value achieved by RV-SIS, DC-SIS and NIS are 0.927, 0.976 and 0.844, respectively.

The top four most influential genes identified by RV-SIS, DC-SIS and NIS are (Msa.2877.0, Msa.741.0, Msa.1166.0, Msa.26025.0), (Msa.2134.0, Msa. 2877.0, Msa.26025.0, Msa.5583.0) and (Msa.2877.0, Msa.1166.0, Msa.741.0, Msa.18571.0), respectively. Fitting again semiparametric SIMs we obtain R^2-values of 0.9995776, 0.9990484 and 0.9290883 for RV-SIS, DC-SIS and NIS, respectively.

It is seen that, though the selected sets of variables are not identical, RV-SIS, DC-SIS have similar behavior in terms of the nonparametric R^2 criterion, while NIS does somewhat worse.

3.3.4 A Real Data Example II

Kim *et al.* (2014) analyzed the ovarian cancer data from The Cancer Genome Atlas (TCGA) to identify the important genes for predicting the ovarian cancer. This data consists of 258 subject and 12,042 gene expressions. We apply RV-SIS, NIS, and DC-SIS procedures to identify the most influential gene expression for predicting ovarian cancer.

The top ten influential genes identified by RV-SIS, DC-SIS, and NIS are (NDRG3, FSTL1 SCRN1, FAM89B, AP1G2, SF3B1, C16orf45, C9orf95, C9orf61, GRK5), (NDRG3, RFX3, GALNT10, MARCH6, ABHD6, NEBL, CPNE1, CLEC3B, C20orf3, GNS) and (NDRG3, PADI2, ANXA4, C9orf61, SCRN1, FKBP4, C9orf95, CA7, STIL, GALNT10) respectively. We use top ten influential genes because RV-SIS with soft thresholding contains 10 covariates in the submodel. To compare the models chosen by the three methods, we fit a Klein and Spady's binary choice estimator,

$$Y_i = 1(\mathbf{X}_i^T \beta \geq e_i) \text{ for i= 1, 2, 3}$$

where $\mathbf{X}_i = (X_{i_1}, \ldots, X_{i_{10}}),\ i = 1, 2, 3$ are the top ten variables chosen by RV-SIS, DC-SIS, and NIS, respectively, and use the overall correct classification ratio to compare the performance. The overall correct classification ratio by RV-SIS, DC-SIS, and NIS are: 0.8023256, 0.7364341, and 0.7596899, respectively. We also use the performance measure suggested by McFadden *et al.* (1977) for performance comparison. This measure is achieved by RV-SIS, DC-SIS, and NIS are: 0.7824199, 0.7002284, and 0.7288925, respectively.

This result show that the top ten influential genes identified by RV-SIS have a better classification rate than the genes identified by DC-SIS and NIS.

3.4 Theoretical Properties

3.4.1 Some Lemmas

In all that follows, $f(x)$ is a generic notation for any of the marginal densities $f_k(x)$. Lemmas 1, 2, 3, and 4 are used to prove the Theorem 3.2.2.

Lemma 3.4.1. For any random variable X which has a moment generating function $\mathrm{E}\{\exp(tX)\}$ for $0 < t < t_0$,

$$P(X - E(X) \geq \epsilon) \leq \exp(-t\epsilon)\mathrm{E}\{\exp(t(X - \mathrm{E}(X)))\},\ t > 0$$

If $P(|X| \leq M) = 1$, then,

$$\mathrm{E}\{\exp(t(X - \mathrm{E}(X)))\} \leq \exp\left(\frac{1}{2}t^2M^2\right),\ t > 0$$

Proof. It follows directly from Theorem 5.6.1.A of Serfling (2009, pp 201). □

Lemma 3.4.2. Suppose $\widehat{f}(x)$ be the kernel density estimator of $f(x)$. Under conditions (C2) and (C3), and $h = O(1)$, we have

$$\sup_{x \in \mathbb{R}} |\widehat{f}(x) - f(x)| = O\left(\left(\frac{\log(n)}{nh}\right)^{1/2} + h^2\right)$$

almost surely.

Proof. It follows by writing $|\widehat{f}(x) - f(x)| \leq |\widehat{f}(x) - E\widehat{f}(x)| + |E\widehat{f}(x) - f(x)|$, using Theorem 5 of Hansen (2008) with $Y \equiv 1$ to get $\sup_x |\widehat{f}(x) - E\widehat{f}(x)| = O((\log(n)/nh)^{1/2})$, and $|E\widehat{f}(x) - f(x)| = O(h^2)$, which follows by a direct calculation. □

Lemma 3.4.3 Let $W_j(x) = K(\frac{x-X_j}{h})/\sum_{i=1}^n K(\frac{x-X_i}{h})$ be the weight function of the Nadaraya-Watson estimator. Then, under the same assumptions as in Lemma 3.4.2, we have

$$\sum_{j=1}^n W_j^2(x) = O\left(\frac{1}{nh}\right), \quad \text{almost surely.}$$

Proof. Noting that $K^2(\cdot)/\int K^2(u)du$ is a symmetric kernel function, by Lemma 3.4.2 it is easily seen that

$$\sum_{j=1}^n W_j^2(x) = \frac{1}{nh}\frac{\left(\frac{1}{nh}\sum_{j=1}^n K^2(\frac{x-X_j}{h})\right)}{\left(\frac{1}{nh}\sum_{i=1}^n K(\frac{x-X_i}{h})\right)^2} = O\left(\frac{1}{nh}\right), \quad \text{almost surely.} \ \square$$

Lemma 3.4.4. Under condition (C1)-(a), (C2), (C3), and (C4) and any $0 < \gamma < 2/5$ there exists positive constants c_1, and c_2 such that,

$$P(\max_i |\widehat{m}(X_i) - m(X_i)| > \epsilon) \leq O(n \exp(-c_1 \epsilon^2 n^{4/5 - 2\gamma}) + n^2 \exp(-c_2 n^\gamma))$$

Proof. By adding and subtracting $\sum_j m(X_j)W_j(X_i)$ we have the inequality

$$\begin{aligned}
& P(|\widehat{m}(X_i) - m(X_i)| > \epsilon) \\
\leq\ & P(|\sum_{j=1}^{n}(y_j - m(X_j))W_j(X_i)| > \frac{\epsilon}{2}) + P(|\sum_{j=1}^{n}(m(X_j) - m(X_i))W_j(X_i)| > \frac{\epsilon}{2}) \\
\equiv\ & A + B.
\end{aligned}$$

Note that the dependence of A and B on i is suppressed for convenience. Consider first A. Letting $I_{1j} = I\{|y_j - m(X_j)| \leq M\}$, where M will be allowed to tend to ∞ with n, and $I_{2j} = 1 - I_{1j}$, and noting that $\mathrm{E}[\sum_{j=1}^{n}(y_j - m(X_j))W_j(X_i)] = 0$, we have the following inequality

$$\begin{aligned}
A \ \leq\ & P(|\sum_{j=1}^{n}(y_j - m(X_j))W_j(X_i)I_{1j} - \sum_{j=1}^{n}\mathrm{E}((y_j - m(X_j))W_j(X_i)I_{1j})| > \frac{\epsilon}{4}) \\
& + P(|\sum_{j=1}^{n}(y_j - m(X_j))W_j(X_i)I_{2j} - \sum_{j=1}^{n}\mathrm{E}((y_j - m(X_j))W_j(X_i)I_{2j})| > \frac{\epsilon}{4}) \\
\equiv\ & A_1 + A_2.
\end{aligned}$$

Arguing conditionally on $(X_1, \ldots, X_n)$, and using Markov's inequality and Lemma 3.4.1,

$$P(\sum_{j=1}^{n}(y_j - m(X_j))W_j(X_i)I_{1j} - \sum_{j=1}^{n}\mathrm{E}((y_j - m(X_j))W_j(X_i)I_{1j}) > \frac{\epsilon}{4})$$

$$
\begin{aligned}
&\leq \ \exp(-t_1\frac{\epsilon}{4})\prod_{j=1}^{n}\mathrm{E}(\exp(t_1[(y_j-m(X_j))W_j(X_i)I_{1j}-\mathrm{E}((y_j-m(X_j))W_j(X_i)I_{1j})])) \\
&\leq \ \exp(-t_1\frac{\epsilon}{4})\prod_{j=1}^{n}\exp(\frac{1}{2}t_1^2W_j^2(X_i)M^2) \ = \ \exp(-t_1\frac{\epsilon}{4})\exp(\frac{1}{2}t_1^2M^2\sum_{j=1}^{n}W_j^2(X_i)) \\
&= \ \exp(-\frac{1}{32}\frac{\epsilon^2}{\sum_j W_j^2(X_i)M^2}), \text{ by choosing } t_1=\frac{\epsilon}{4\sum\limits_{j=1}^{n}W_j^2(X_i)M^2} \\
&\simeq \ \exp(-\frac{1}{32}\frac{\epsilon^2 nh}{M^2}), \text{ by Lemma 3.4.3}
\end{aligned}
$$

Similarly,

$$
P(\sum_{j=1}^{n}(y_j-m(X_j))W_j(X_i)I_{1j}-\sum_{j=1}^{n}\mathrm{E}((y_j-m(X_j))W_j(X_i)I_{1j})<-\frac{\epsilon}{4})\leq\exp(-\frac{1}{32}\frac{\epsilon^2 nh}{M^2})
$$

Thus, also unconditionally, we have that for each i,

$$
A_1 \ \leq \ 2\exp\left(-\frac{1}{32}\frac{\epsilon^2 nh}{M^2}\right)
$$

For the A_2 part,

$$
A_2\leq P(|\sum_{j=1}^{n}(y_j-m(X_j))W_j(X_i)I_{2j}|+\sum_{j=1}^{n}|\mathrm{E}((y_j-m(X_j))W_j(X_i)I_{2j})|>\frac{\epsilon}{4})
$$

We first show that $\sum_{j=1}^{n}|\mathrm{E}((y_j-m(X_j))W_j(X_i)I_{2j})|$ is bounded by $\epsilon/8$ for n large enough. By the Cauchy-Schwartz and Markov inequalities, we have

$$
\begin{aligned}
&|E[(y_j-m(X_j))W_j(X_i)I_{2j}]|\leq\sqrt{\mathrm{E}[\{(y_j-m(X_j))W_j(X_i)\}^2]P(|y_j-m(X_j)|>M)} \\
&\leq \ \sqrt{\mathrm{E}[\{y_j-m(X_j)\}^2W_j^2(X_i)]\exp(-tM)\mathrm{E}\{\exp(t|y_j-m(X_j)|)\}}
\end{aligned}
$$

By condition (C1)-(a), there exists a constant t such that $E\{\exp(t|y_j - m(X_j)|)\} < C_1$. Also, by Lemma 3.4.2, $\mathrm{E}[\{y_j - m(X_j)\}^2 W_j^2(X_i)] = O(1/(nh)^2)$, uniformly in i. Then, by choosing $M = n^\gamma$, some $\gamma > 0$, we have $\sum_{j=1}^n |\mathrm{E}((y_j - m(X_j))W_j(X_i)I_{2j})| < \epsilon/8$, for n large enough. Hence, for n large enough,

$$A_2 \leq P(|\sum_{j=1}^n (y_j - m(X_j))W_j(X_i)I_{2j}| > \frac{\epsilon}{8}).$$

To bound this, note first that

$$\left\{|\sum_{j=1}^n (y_j - m(X_j))W_j(X_i)I_{2j}| > \epsilon/8\right\} \subset \bigcup_{j=1}^n \{|y_j - m(X_j)| > M\}.$$

Indeed, if the event on the left hand side holds it must be that $|y_j - m(X_j)| > M$ for at least one j since, otherwise $|(y_j - m(X_j))W_j(X_i)I_2| = 0$ for all j which contradicts $|\sum_{j=1}^n (y_j - m(X_j))W_j(X_i)I_2| > \epsilon/8$. Thus, by condition (C1)-(a), it follows that

$$\begin{aligned} A_2 &\leq P(\bigcup_j \{|y_j - m(X_j)| > M\}) \leq nP(|y_j - m(X_j)| > M) \\ &\leq n\exp(-tM)\mathrm{E}[\exp(t|y_j - m(X_j)|)] = nC_1\exp(-tM) \end{aligned}$$

Then by choosing $M = n^\gamma$, $0 < \gamma < 2/5$, we have

$$\begin{aligned} A &\leq 2\exp(-\frac{1}{32}\frac{\epsilon^2 nh}{M^2}) + nC_1\exp(-tM) \\ &= 2\exp(-\frac{1}{32}\epsilon^2 n^{1-2\gamma}h) + nC_1\exp(-tn^\gamma) \qquad (3.4.1) \end{aligned}$$

Consider now part B. By condition (C4) and for n large enough, we have

$$\begin{aligned} B &\leq P(\sum_{j=1}^{n} |(m(X_j) - m(X_i))W_j(X_i)| > \frac{\epsilon}{2}) \\ &\leq P(\sum_{j=1}^{n} |\Lambda_2(X_j - X_i)W_j(X_i)| > \frac{\epsilon}{2}) \\ &\leq P(\Lambda_2 h \sum_{j=1}^{n} W_j(X_i) > \frac{\epsilon}{2}) = P(\Lambda_2 h > \frac{\epsilon}{2}) = 0. \end{aligned} \tag{3.4.2}$$

Therefore, by (3.4.1) and (3.4.2), we have that for all n large enough

$$P(|\widehat{m}(X_i) - m(X_i)| > \epsilon) \leq 2\exp(-\frac{1}{32}\epsilon^2 n^{1-2\gamma} h) + nC_1 \exp(-tn^{\gamma}).$$

It follows that under condition (C1)-(a), (C2), (C3), and (C4), and any $0 < \gamma < 2/5$, there exists positive constants c_1 and c_2 such that,

$$\begin{aligned} P(\max_i |\widehat{m}(X_i) - m(X_i)| > \epsilon) &\leq O(n\exp(-c_1\epsilon^2 n^{1-2\gamma} h) + n^2 \exp(-c_2 n^{\gamma})) \\ &\leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2 \exp(-c_2 n^{\gamma})), \end{aligned} \tag{3.4.3}$$

by substituting $n^{-1/5}$ for h. □

3.4.2 Proof of Theorem 3.2.2

For part 1 write

$$P(|\tilde{S}^2_{m_k} - \sigma^2_{m_k}| \geq \epsilon) \leq P(|\tilde{S}^2_{m_k} - S^2_{m_k}| \geq \epsilon/2) + P(|S^2_{m_k} - \sigma^2_{m_k}| \geq \epsilon/2) \equiv T_1 + T_2, \tag{3.4.4}$$

where $S_{m_k}^2 = \frac{1}{n}\sum_{i=1}^n [m_k(X_i) - (\frac{1}{n}\sum_{l=1}^n m_k(X_i))]^2$. For convenience in notation, we will omit the subscript k from m_k and X_{kj}, $j = 1, \ldots, n$, for the rest of this proof. For T_1 we have

$$
\begin{aligned}
T_1 &= P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}^2(x_i) - m^2(X_i)) - ((\frac{1}{n}\sum_{l=1}^n \widehat{m}(X_i))^2 - (\frac{1}{n}\sum_{l=1}^n m(X_i))^2| > \frac{\epsilon}{2})) \\
&\leq P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))(\widehat{m}(X_i) + m(X_i))| > \frac{\epsilon}{4}) \\
&\quad + P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))(\frac{1}{n}\sum_{l=1}^n(\widehat{m}(X_i) + m(X_i))| > \frac{\epsilon}{4})) \\
&\leq P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))^2| > \frac{\epsilon}{8}) + P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))(2m(X_i))| > \frac{\epsilon}{8}) \\
&\quad + P([\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))]^2 > \frac{\epsilon}{8}) + P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))\frac{1}{n}\sum_{i=1}^n 2m(X_i)| > \frac{\epsilon}{8}) \\
&\equiv A_1 + A_2 + A_3 + A_4.
\end{aligned}
$$

The following inequalities all follow by Lemma 3.4.4 (so that $0 < \gamma < 2/5$):

$$
\begin{aligned}
A_1 &\leq P(\max_i(\widehat{m}(X_i) - m(X_i))^2 > \frac{\epsilon}{8}) \\
&\leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^\gamma)), \\
A_2 &\leq P(\max_i|(\widehat{m}(X_i) - m(X_i))| > \frac{\epsilon}{16\sup_x|m(x)|}) \\
&\leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^\gamma)), \\
A_3 &\leq P(\max_i|(\widehat{m}(X_i) - m(X_i))| > \epsilon) \\
&\leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^\gamma)), \\
A_4 &\leq P(|\frac{1}{n}\sum_{i=1}^n(\widehat{m}(X_i) - m(X_i))| > \frac{\epsilon}{16\sup_x|m(x)|}) \\
&\leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^\gamma)).
\end{aligned}
$$

Combining the above we have

$$T_1 \leq O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^\gamma)). \tag{3.4.5}$$

Consider now T_2, and let $h(X_i, X_j)$ be the kernel of the U-statistic $U_m = [n/(n-1)]S_m^2$. For a constant M, we decompose U_m as $U_m = U_{1m} + U_{2m}$, where

$$U_{1m} = \frac{1}{n(n-1)}\sum_{i\neq j} h(X_i, X_j)I\{h(X_i, X_j) \leq M\} \text{ and } U_{2m} = U_m - U_{1m}.$$

Similarly, we decompose $\sigma_m^2 = \mathrm{E}(U_m)$ as $\sigma_m^2 = \sigma_{1m}^2 + \sigma_{1m}^2$, where

$$\sigma_{1m}^2 = \mathrm{E}(h(X_i, X_j)I\{h(X_i, X_j) \leq M\}) \text{ and } \sigma_{2m}^2 = \sigma_m^2 - \sigma_{1m}^2.$$

Then we have following inequality

$$\begin{aligned} T_2 &= P\left(|\frac{n-1}{n}U_m - \sigma_m^2| \geq \epsilon/2\right) \\ &\leq P\left(|\frac{n-1}{n}(U_{1m} - \sigma_{1m}^2)| \geq \epsilon/4\right) + P\left(|\frac{n-1}{n}(U_{2m} - \sigma_{2m}^2) - \frac{1}{n}\sigma_m^2| \geq \epsilon/4\right) \\ &\equiv C_1 + C_2 \end{aligned} \tag{3.4.6}$$

By Lemma 3.4.1 we have that for any $t > 0$,

$$P\left(\frac{n-1}{n}(U_{1m} - \sigma_{1m}^2) \geq \epsilon/4\right) \leq \exp\left(-\frac{t\epsilon n}{4(n-1)}\right)\exp(-t\sigma_{1m}^2)\mathrm{E}(\exp(tU_{1m})). \tag{3.4.7}$$

Next, using the representation $U_{m1} = \frac{1}{n!}\sum_{n!} W(X_{i_1}, \ldots, X_{i_n})$, where $W(X_1, \ldots, X_n) = \frac{1}{m}\sum_{i=1}^m h(X_{2i-1}, X_{2i})I\{h(X_{2i-1}, X_{2i}) \leq M\}$ is an average of $m = [n/2]$ i.i.d random variables, and $\sum_{n!}$ denotes the summation over all possible permutations of $(1, \ldots, n)$

(cf. Serfling, 1981, pp. 180-181), we have

$$\begin{aligned} \mathrm{E}(\exp(tU_{1m})) &= \mathrm{E}\left(\exp\left(\frac{t}{n!}\sum_{n!} W(X_{i_1},\ldots,X_{i_n})\right)\right) \\ &\le \frac{1}{n!}\sum_{n!} \mathrm{E}[\exp\{tW(X_{i_1},\ldots,X_{i_n})\}] \\ &= \mathrm{E}\left(\exp\left(\sum_{i=1}^{m} \frac{t}{m} h(X_{2i-1},X_{2i}) I\{h(X_{2i-1},X_{2i}) \le M\}\right)\right) \\ &= \mathrm{E}^m\left(\exp\left(\frac{t}{m} h(X_{2i-1},X_{2i}) I\{h(X_{2i-1},X_{2i}) \le M\}\right)\right), \end{aligned}$$

where Jensen's inequality was also used. Substituting this in (3.4.7) we have,

$$\begin{aligned} & P\left(\frac{n-1}{n}(U_{1m}-\sigma_{1m}^2) \ge \epsilon/4\right) \\ \le\ & \exp\left(-\frac{tn\epsilon}{4(n-1)}\right) \mathrm{E}^m(\exp(\frac{t}{m}(h(X_{2i-1},X_{2i}) I\{h(X_{2i-1},X_{2i}) \le M\} - \sigma_{m1}^2))) \\ \le\ & \exp\left(-\frac{t\epsilon n}{4(n-1)} + \frac{t^2M^2}{2m}\right) \text{ by Lemma 3.4.1} \\ \le\ & \exp\left(-\frac{n^2\epsilon^2 m}{32M^2(n-1)^2}\right) \text{ by choosing } t = \frac{n\epsilon m}{4M^2(n-1)} \end{aligned}$$

Therefore, for C_1 given in (4.4.2) we have

$$C_1 \le 2\exp\left(-\frac{n^2\epsilon^2 m}{32M^2(n-1)^2}\right) \le 2\exp\left(-\frac{\epsilon^2 n}{64M^2}\right) \tag{3.4.8}$$

Consider now C_2 given in (4.4.2). Note first that $\sigma_m^2/n < \epsilon/16$ for all n sufficient large. Also, by the Cauch-Schwartz and Markov inequalities, we have

$$\begin{aligned} \sigma_{2m}^2 &\le \sqrt{E(h^2(X_i,X_j))P(h(X_i,X_j) > M)} \\ &\le \sqrt{E(h^2(X_i,X_j))exp(-tM)E(exp(th(X_i,X_j))} \end{aligned}$$

so that, by choosing $M = n^{\gamma}$, $\gamma > 0$, condition (C1)-(b) yields $(n-1)\sigma_{2m}^2/n < \epsilon/16$ for n sufficient large. Thus, for n large enough,

$$C_2 \leq P\left(|\frac{n-1}{n}U_{2m}| > \epsilon/8\right).$$

To bound this, observe that $\{|\frac{n-1}{n}U_{2m}| \geq \epsilon/8\} \subseteq \bigcup_{i\neq j}\{|h(X_i, X_j)| \geq M\}$. Thus, by Markov's inequality and condition (C1)-(b), it follows that

$$\begin{aligned} C_2 \leq P\left(\bigcup_{i\neq j}|h(X_i,X_j)| \geq M\right) &\leq n^2\exp(-tM)\mathrm{E}(\exp(t|h(X_i,X_j)|)) && (3.4.9) \\ &\leq n^2 C_3 \exp(-tn^{\gamma}) && (3.4.10) \end{aligned}$$

Combining (4.4.2), (3.4.8) with $M = n^{\gamma}$, for $\gamma < 1/2$, and (3.4.9), we have

$$T_2 \leq 2\exp\left(-\frac{\epsilon^2 n^{1-2\gamma}}{64}\right) + n^2\exp\left(-tn^{\gamma}\right) = O(\exp(-c_3\epsilon^2 n^{1-2\gamma}) + n^2\exp(-c_4 n^{\gamma})), \quad (3.4.11)$$

for some positive constants c_3 and c_4.

By (3.4.4), (3.4.5) and (3.4.11), for $0 < \gamma < 2/5$ we have

$$P(|\widetilde{S}^2 - \sigma_m^2| \geq \epsilon) = O(n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^{\gamma}))$$

It follows that for $0 < \gamma < 2/5$

$$\begin{aligned} P(\max_k |\widetilde{S}_k^2 - \sigma_{m_k}^2| \geq \epsilon) &\leq O(p[n\exp(-c_1\epsilon^2 n^{4/5-2\gamma}) + n^2\exp(-c_2 n^{\gamma})]) \\ &= O(p[n\exp(-c_1 n^{4/5-2(\gamma+\kappa)}) + n^2\exp(-c_2 n^{\gamma})]) \end{aligned}$$

The last equality holds by choosing $\epsilon = cn^{-\kappa}$ for a constant $c > 0$, $0 < \kappa < 2/5$ and

$0 < \gamma < 2/5 - \kappa$.

For part 2 of Theorem 3.2.2, if $\mathcal{D} \not\subseteq \widehat{\mathcal{D}}$, then there must exists some $k \in \mathcal{D}$ such that $k \notin \widehat{\mathcal{D}}$. Thus $\widetilde{S}_k^2 > C_d$, and by condition (C5) it follows that $\sigma_{m_k}^2 - \widetilde{S}_k^2 > cn^{-\kappa}$. Thus, $\{\mathcal{D} \not\subseteq \widehat{\mathcal{D}}\} \subseteq \{\max\limits_{k \in \mathcal{D}}\{|\widetilde{S}_k^2 - \sigma_{mk}^2| > cn^{-\kappa}\}$.

Using part 1 of this theorem we have

$$\begin{aligned} P(\mathcal{D} \subseteq \widehat{\mathcal{D}}) &\geq 1 - P(\min_{k \in \mathcal{D}} |\widetilde{S}_k^2 - \sigma_{mk}^2| > cn^{-\kappa}) \\ &= 1 - |\mathcal{D}| P(|\widetilde{S}_k^2 - \sigma_{mk}^2| > cn^{-\kappa}) \\ &\geq 1 - O(|\mathcal{D}|[\exp(-n \exp(-c_1 n^{4/5 - 2(\gamma+\kappa)}) + n^2 \exp(-c_2 n^{\gamma})]). \ \square \end{aligned}$$

Table 3.1. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

Model 3.(a)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
4.00	4.00	4.00	5.00	7.00	4.00	4.00	4.00	5.00	6.00
NIS					RRCS				
4.00	4.00	4.00	5.00	7.05	4.00	4.00	4.00	5.00	6.00
RV-SIS									
4.00	4.00	4.00	5.00	9.05					
Model 3.(b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS					DC-SIS				
84.60	526.75	1179.00	1655.00	1923.35	9.00	26.00	68.50	169.25	516.50
NIS					RRCS				
4.00	4.00	6.00	14.00	100.20	214.85	786.50	1355.50	1708.75	1931.10
RV-SIS									
4.00	4.00	7.00	22.00	273.20					
Model 3.(c)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
232.00	853.50	1363.50	1689.75	1933.00	103.95	316.25	565.00	860.00	1420.50
NIS					RRCS				
55.00	312.25	749.00	1264.25	1786.15	255.65	929.00	1384.50	1732.25	1943.10
RV-SIS									
5.00	15.00	62.50	277.00	1208.10					
Model 3.(d)									
SIS					DC-SIS				
106.90	583.75	1149.50	1628.75	1930.00	102.90	326.25	654.50	1069.00	1583.70
NIS					RRCS				
33.50	389.00	882.00	1463.25	1915.00	231.55	832.25	1337.00	1678.25	1944.05
RV-SIS									
6.00	20.00	89.00	327.25	1144.55					

Table 3.2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$.

Model 3.(a)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
8.00	11.00	17.00	37.25	249.55	6.00	9.00	12.00	17.00	76.05
NIS					RRCS				
6.00	9.00	13.00	26.00	153.40	6.00	9.00	13.00	22.00	141.35
RV-SIS									
5.00	8.00	11.00	26.25	146.60					
Model 3.(b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS					DC-SIS				
29.90	256.50	924.00	1544.75	1935.05	8.00	10.00	13.00	18.00	40.00
NIS					RRCS				
4.00	6.00	8.00	10.00	22.00	111.60	502.50	1133.00	1636.00	1938.10
RV-SIS									
4.00	6.00	7.00	10.00	32.05					
Model 3.(c)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
93.90	520.25	1122.50	1647.25	1925.15	40.95	148.00	334.00	629.00	1149.85
NIS					RRCS				
16.00	74.00	239.50	625.00	1454.60	145.85	595.50	1207.00	1585.25	1930.10
RV-SIS									
9.00	17.00	55.50	244.00	978.30					
Model 3.(d)									
SIS					DC-SIS				
34.80	183.75	701.50	1449.25	1899.40	31.90	142.00	344.00	675.25	1322.10
NIS					RRCS				
18.00	106.00	418.00	1111.00	1815.20	83.90	373.50	979.50	1534.25	1930.05
RV-SIS									
9.00	20.00	45.00	171.50	893.40					

Table 3.3. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5$.

Model (a)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
36.00	47.00	55.00	67.00	91.00	36.95	47.00	55.00	67.00	95.00
NIS					RRCS				
36.00	47.00	55.00	67.00	90.10	36.00	47.00	56.00	68.00	94.00
RV-SIS									
37.95	47.00	55.00	67.00	94.00					
Model (b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS					DC-SIS				
145.95	232.00	411.00	910.75	1979.70	122.00	196.75	322.00	651.25	1716.40
NIS					RRCS				
78.00	119.75	180.00	267.75	919.40	150.95	273.00	449.00	1089.00	2000.00
RV-SIS									
77.00	119.75	180.50	314.00	1164.50					
Model (c)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
151.90	245.75	417.00	806.50	1999.05	132.90	227.00	367.50	732.25	1902.45
NIS					RRCS				
115.95	192.75	310.50	593.25	1713.60	148.95	251.75	451.50	1013.00	2000.00
RV-SIS									
69.00	99.00	142.00	208.00	456.05					
Model (d)									
SIS					DC-SIS				
29.00	41.00	54.00	84.00	160.05	27.95	39.00	53.00	77.00	169.10
NIS					RRCS				
28.00	39.00	52.00	78.00	164.40	27.95	39.00	53.00	82.00	181.35
RV-SIS									
24.00	31.00	40.00	54.25	101.15					

Table 3.4. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

Model 3.(a)															
	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
	SIS					DC-SIS					RRCS				
d1	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d2	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	1.00
d3	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	1.00
	NIS					RV-SIS									
d1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99					
d2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00					
d3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00					
Model 3.(b)															
	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
	SIS					DC-SIS					RRCS				
d1	0.08	0.08	1.00	1.00	0.03	0.51	0.51	1.00	1.00	0.33	0.03	0.04	1.00	1.00	0.01
d2	0.12	0.14	1.00	1.00	0.05	0.68	0.68	1.00	1.00	0.52	0.07	0.07	1.00	1.00	0.02
d3	0.16	0.17	1.00	1.00	0.06	0.76	0.78	1.00	1.00	0.65	0.09	0.10	1.00	1.00	0.02
	NIS					RV-SIS									
d1	0.95	0.93	1.00	1.00	0.89	0.89	0.88	1.00	1.00	0.80					
d2	0.97	0.96	1.00	1.00	0.94	0.94	0.93	1.00	1.00	0.88					
d3	0.98	0.97	1.00	1.00	0.96	0.95	0.94	1.00	1.00	0.90					
Model 3.(c)															
	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
	SIS					DC-SIS					RRCS				
d1	0.01	0.03	1.00	1.00	0.00	0.04	0.13	1.00	1.00	0.01	0.01	0.02	1.00	1.00	0.00
d2	0.03	0.05	1.00	1.00	0.00	0.08	0.26	1.00	1.00	0.04	0.04	0.03	1.00	1.00	0.00
d3	0.06	0.07	1.00	1.00	0.01	0.12	0.37	1.00	1.00	0.06	0.06	0.05	1.00	1.00	0.00
	NIS					RV-SIS									
d1	0.04	0.59	1.00	1.00	0.03	0.94	0.42	1.00	1.00	0.40					
d2	0.09	0.68	1.00	1.00	0.07	0.97	0.54	1.00	1.00	0.53					
d3	0.12	0.74	1.00	1.00	0.10	0.99	0.60	1.00	1.00	0.59					
Model 3.(d)															
	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
	SIS					DC-SIS					RRCS				
d1	0.05	0.10	1.00	0.99	0.02	0.04	0.11	1.00	1.00	0.01	0.04	0.04	1.00	1.00	0.01
d2	0.08	0.17	1.00	1.00	0.04	0.08	0.24	1.00	1.00	0.03	0.06	0.07	1.00	1.00	0.01
d3	0.11	0.21	1.00	1.00	0.06	0.11	0.35	1.00	1.00	0.06	0.08	0.10	1.00	1.00	0.02
	NIS					RV-SIS									
d1	0.09	0.35	1.00	0.99	0.05	0.58	0.61	1.00	0.97	0.35					
d2	0.14	0.41	1.00	1.00	0.09	0.71	0.69	1.00	0.98	0.49					
d3	0.19	0.46	1.00	1.00	0.12	0.79	0.72	1.00	0.99	0.57					

Table 3.5. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$.

Model 3.(a)															
	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
	SIS					DC-SIS					RRCS				
d1	1.00	1.00	0.75	1.00	0.75	1.00	1.00	0.90	1.00	0.90	1.00	1.00	0.84	1.00	0.84
d2	1.00	1.00	0.85	1.00	0.85	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.91	1.00	0.92
d3	1.00	1.00	0.89	1.00	0.89	1.00	1.00	0.97	1.00	0.97	1.00	1.00	0.94	1.00	0.94
	NIS					RV-SIS									
d1	1.00	1.00	0.82	1.00	0.82	1.00	1.00	0.81	1.00	0.81					
d2	1.00	1.00	0.91	1.00	0.92	1.00	1.00	0.90	1.00	0.90					
d3	1.00	1.00	0.94	1.00	0.94	1.00	1.00	0.92	1.00	0.93					
Model 3.(b)															
	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
	SIS					DC-SIS					RRCS				
d1	0.10	0.11	0.98	1.00	0.07	0.97	0.96	1.00	1.00	0.94	0.02	0.04	1.00	1.00	0.01
d2	0.16	0.18	0.99	1.00	0.11	0.99	0.99	1.00	1.00	0.99	0.07	0.08	1.00	1.00	0.03
d3	0.20	0.23	1.00	1.00	0.15	0.99	1.00	1.00	1.00	0.99	0.11	0.12	1.00	1.00	0.05
	NIS					RV-SIS									
d1	1.00	1.00	0.99	1.00	0.98	1.00	1.00	0.96	1.00	0.96					
d2	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.98					
d3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99					
Model 3.(c)															
	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
	SIS					DC-SIS					RRCS				
d1	0.03	0.05	0.99	1.00	0.02	0.09	0.12	1.00	1.00	0.05	0.02	0.03	0.99	1.00	0.01
d2	0.07	0.09	1.00	1.00	0.05	0.17	0.24	1.00	1.00	0.11	0.05	0.07	1.00	1.00	0.03
d3	0.09	0.11	1.00	1.00	0.06	0.27	0.33	1.00	1.00	0.18	0.07	0.09	1.00	1.00	0.04
	NIS					RV-SIS									
d1	0.16	0.55	1.00	1.00	0.14	0.99	0.43	1.00	1.00	0.43					
d2	0.29	0.68	1.00	1.00	0.26	1.00	0.54	1.00	1.00	0.55					
d3	0.38	0.74	1.00	1.00	0.35	1.00	0.62	1.00	1.00	0.62					
Model 3.(d)															
	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
	SIS					DC-SIS					RRCS				
d1	0.09	0.20	1.00	0.86	0.06	0.07	0.20	1.00	0.99	0.06	0.03	0.09	1.00	0.96	0.02
d2	0.17	0.27	1.00	0.94	0.15	0.17	0.34	1.00	0.99	0.15	0.07	0.14	1.00	0.98	0.05
d3	0.23	0.31	1.00	0.96	0.19	0.24	0.44	1.00	1.00	0.21	0.10	0.19	1.00	0.99	0.08
	NIS					RV-SIS									
d1	0.20	0.38	1.00	0.88	0.13	0.88	0.60	1.00	0.83	0.45					
d2	0.28	0.46	1.00	0.95	0.21	0.94	0.70	1.00	0.92	0.61					
d3	0.34	0.50	1.00	0.96	0.26	0.96	0.76	1.00	0.94	0.69					

Table 3.6. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n], d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.5$.

Model (a)															
	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
	SIS					DC-SIS					RRCS				
d1	0.09	0.09	0.64	1.00	0.06	0.13	0.12	0.43	1.00	0.06	0.13	0.12	0.40	1.00	0.05
d2	0.85	0.85	1.00	1.00	0.86	0.86	0.86	0.99	1.00	0.86	0.87	0.86	0.98	1.00	0.86
d3	0.99	0.99	1.00	1.00	0.99	0.98	0.98	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99
	NIS					RV-SIS									
d1	0.09	0.09	0.71	1.00	0.07	0.10	0.09	0.73	1.00	0.07					
d2	0.87	0.86	1.00	1.00	0.87	0.85	0.85	1.00	1.00	0.85					
d3	0.98	0.98	1.00	1.00	0.98	0.98	0.98	1.00	1.00	0.98					
Model (b)															
	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
	SIS					DC-SIS					RRCS				
d1	0.00	0.00	0.13	1.00	0.00	0.00	0.00	0.08	1.00	0.00	0.00	0.00	0.07	1.00	0.00
d2	0.00	0.00	0.92	1.00	0.00	0.00	0.00	0.86	1.00	0.00	0.00	0.00	0.84	1.00	0.00
d3	0.01	0.01	1.00	1.00	0.01	0.03	0.03	0.99	1.00	0.03	0.01	0.01	0.99	1.00	0.01
	NIS					RV-SIS									
d1	0.00	0.00	0.27	1.00	0.00	0.00	0.00	0.31	1.00	0.00					
d2	0.04	0.04	0.96	1.00	0.04	0.04	0.04	0.97	1.00	0.04					
d3	0.22	0.21	1.00	1.00	0.23	0.21	0.21	1.00	1.00	0.22					
Model (c)															
	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
	SIS					DC-SIS					RRCS				
d1	0.00	0.00	0.15	1.00	0.00	0.00	0.00	0.07	1.00	0.00	0.00	0.00	0.07	1.00	0.00
d2	0.00	0.00	0.91	1.00	0.00	0.00	0.00	0.83	1.00	0.00	0.00	0.00	0.81	1.00	0.00
d3	0.01	0.01	1.00	1.00	0.01	0.02	0.03	0.97	1.00	0.03	0.02	0.02	0.96	1.00	0.02
	NIS					RV-SIS									
d1	0.00	0.00	0.27	1.00	0.00	0.00	0.00	0.32	1.00	0.00					
d2	0.00	0.00	0.95	1.00	0.00	0.09	0.09	0.96	1.00	0.09					
d3	0.05	0.05	1.00	1.00	0.05	0.32	0.32	1.00	1.00	0.34					
Model (d)															
	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
	SIS					DC-SIS					RRCS				
d1	0.19	0.19	1.00	1.00	0.18	0.22	0.24	1.00	1.00	0.22	0.24	0.23	1.00	1.00	0.23
d2	0.71	0.71	1.00	1.00	0.72	0.74	0.74	1.00	1.00	0.74	0.72	0.72	1.00	1.00	0.72
d3	0.88	0.88	1.00	1.00	0.88	0.87	0.87	1.00	1.00	0.88	0.86	0.86	1.00	1.00	0.86
	NIS					RV-SIS									
d1	0.22	0.21	1.00	1.00	0.21	0.45	0.45	1.00	1.00	0.44					
d2	0.73	0.74	1.00	1.00	0.74	0.87	0.87	1.00	1.00	0.87					
d3	0.88	0.88	1.00	1.00	0.88	0.97	0.97	1.00	1.00	0.97					

Table 3.7. The comparison of execution time of DC-SIS and RV-SIS in seconds for Model 3.(d) when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

	DC-SIS					NIS					RV-SIS				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
(d)	18.92	19.17	19.30	19.45	19.86	2.32	2.35	2.36	2.38	2.46	1.81	1.82	1.82	1.83	1.90

Table 3.8. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(e)

Model 3.(e)									
Hard Threshold with model size d_1									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.718	0.786	0.806	0.880	0.878	0.910	0.582	0.512	0.704	0.918
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.862	0.744	0.934	0.892	0.882	0.944	0.770	0.730	0.688	0.930
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.960	0.948	0.974	0.888	0.388					
Hard Threshold with model size d_2									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.924	0.868	0.974	0.950	0.950	0.986	0.840	0.834	0.816	0.970
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.884	0.858	0.922	0.894	0.890	0.918	0.796	0.792	0.754	0.920
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.986	0.978	0.988	0.952	0.554					
Hard Threshold with model size d_3									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.884	0.920	0.928	0.958	0.952	0.968	0.776	0.750	0.862	0.970
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.946	0.902	0.982	0.966	0.964	0.990	0.904	0.876	0.878	0.984
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.992	0.984	0.988	0.960	0.634					
Soft Threshold									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.746	0.816	0.844	0.898	0.894	0.920	0.650	0.602	0.742	0.928
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.892	0.800	0.942	0.918	0.900	0.954	0.788	0.770	0.750	0.942
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.966	0.952	0.970	0.918	0.462					

Table 3.9. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(f)

Model 3.(f)									
Hard Threshold with model size d_1									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.976	0.986	0.988
Hard Threshold with model size d_2									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.992	0.994
Hard Threshold with model size d_3									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.996	1.000
Soft Threshold									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.968	0.978	0.986

Table 3.10. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model(g)

Model 3.(g)				
Hard Threshold with model size d_1				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000
Hard Threshold with model size d_2				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000
Hard Threshold with model size d_3				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000
Soft Threshold				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000

Table 3.11. The 5%, 25%, 50%, 75%, and 95% quantiles of submodel size using soft thresholding rule for Models 3.(e), 3.(f), and 3.(g).

Model 3.(e)				
5%	25%	50%	75%	95%
20.00	37.00	53.00	75.00	116.00
Model 3.(f)				
5%	25%	50%	75%	95%
13.00	26.00	43.00	66.25	109.00
Model 3.(g)				
5%	25%	50%	75%	95%
8.00	19.75	38.00	62.00	100.15

Chapter 4

Hypothesis testing of a constant regression function and Test-based Screening

4.1 Introduction

In this chapter, we first develop the asymptotic theory for the variance of the regression function and use it to introduce a new test procedure for testing the significance of a predictor. Using the set of p-values we introduce a variable screening procedure using multiple testing ideas.

This chapter is organized as follows. In section 4.2, we introduce a test procedure for the hypothesis of a constant regression function, and also a test-based variable screening procedure. In section 4.3, we present the result of simulation studies. In section 4.4, we present the technical proofs for the asymptotic theory of the test statistics and a method for estimating the percentiles of the limiting distribution, which is used in the simulation in section 4.3.

4.2 Hypothesis testings of a constant regression function for variable screening and test-based screening

4.2.1 Preliminaries

Consider a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Consider a nonparametric regression function

$$m(x) = \mathrm{E}(Y_i | X_i = x). \tag{4.2.1}$$

We propose a hypothesis test of a constant regression function:

$$H_0 : m(x) = c \text{ for all } x \tag{4.2.2}$$

for some unknown constant c. Under the null hypothesis, the variance of the regression function $\sigma_m^2 = \mathrm{var}(m(x)) = 0$. Suppose $\hat{m}(x)$ is a Nadaraya Watson estimator of $m(x)$, then sample variance of estimators $\tilde{S}_m^2$ estimates the variance of the regression function σ_m^2,

$$\tilde{S}_m^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{m}^2(X_i) - (\frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i))^2$$

The bandwidth of a Nadaraya Watson estimator is set to be of the order $h = C_0 n^{-1/5}$.

Let $Z_i = (X_i, Y_i)$ and $\psi(Z_{i_1}, Z_{i_2}, Z_{i_3})$ and $\psi_2(Z_1, Z_2)$ are defined as following

$$\begin{aligned}\psi(Z_{i_1}, Z_{i_2}, Z_{i_3}) &= \frac{1}{3h}(Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})} + Y_{i_1}Y_{i_3}K_{i_1i_2}K_{i_3i_2}\frac{1}{f^2(X_{i_2})} \\ &\quad +Y_{i_1}Y_{i_2}K_{i_1i_3}K_{i_2i_3}\frac{1}{f^2(X_{i_3})}) \qquad (4.2.3) \\ \psi_2(Z_1, Z_2) &= E(\psi(Z_1, Z_2, Z_3)|Z_1 = z_1, Z_2 = z_2) \qquad (4.2.4)\end{aligned}$$

where $K_{i_2i_1} = K(\frac{X_{i_2}-X_{i_1}}{h})$.

4.2.2 Asymptotic Properties

The following conditions are required for the technical proofs:

(C1) The kernel $K(\cdot)$ has bounded support, is symmetric, and is Lipschitz continuous, i.e, it satisfies, for some $\Lambda_1 < \infty$ and for all $u, u' \in \mathbb{R}$,

$$|K(u) - K(u')| \leq \Lambda_1|u - u'|.$$

(C2) If $f_k(x)$ denotes the marginal density of the kth predictor, we have

$$\sup_x |x|^s E(|Y||X_k = x)f_k(x) \leq B < \infty \text{ for some } s \geq 1$$

$\sup_x f_k(x) < \infty$, $\inf\limits_x f_k(x) > 0$, and $f_k(x)$ is uniformly continuous, for all $k = 1, \ldots, p$.

Lemma 4.2.1. Suppose $\widehat{f}(x)$ be the kernel density estimator of $f(x)$. Under conditions (C1) and (C2), and $h = O(1)$, we have

$$\sup_{x\in\mathbb{R}} |\widehat{f}(x) - f(x)| = O\left(\left(\frac{\log(n)}{nh}\right)^{1/2} + h^2\right)$$

almost surely.

Theorem 4.2.2 Under conditions (C1) and (C2), and H_0 defined in Equation 4.2.6, as $n \to \infty$

$$nh\tilde{S}_m^2 \to 3\sum_{\nu=1}^{\infty} \lambda_\nu(Z_\nu^2 - 1) + 3\theta_2$$

where the Z_ν are i.i.d $N(0,1)$ and λ_ν are the eigenvalues of the integral equation

$$\int_{-\infty}^{\infty} \psi_2(z_1, z_2) f(z_2) dF(z_2) = \lambda f(z_1)$$

and $\theta_2 = E\psi(Z_1, Z_1, Z_2)$.

4.2.3 Estimating the Asymptotic Distribution of Theorem 4.4.2

Let U_n be degenerate U-statistics with kernel ψ_2,

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} \psi_2(Z_i, Z_j).$$

The asymptotic distribution of nU_n is $\sum_{\nu=1}^{\infty} \lambda_\nu(Z_\nu^2 - 1)$, where λ_ν and Z_ν are as in Theorem 4.4.2. As shown in Koltchinskii and Giné (2000), the asymptotic distribution of U-Statistics with degenerate Hilbert-Schmidt kernel ψ_2 can be approximated by the finite spectrum of the empirical matrix version of the operator, Ψ. Let $\tilde{\Psi}$ be the n by n matrix with $\tilde{\Psi}_{i,j} = \frac{1}{n}\psi_2(Z_i, Z_j)$, and let $\bar{\Psi}$ be obtained by by deleting the diagonal in the matrix $\tilde{\Psi}_n$. According to Koltchinskii and Giné (2000), $\bar{\Psi}$ is the empirical version of Ψ.

Since by Theorem 4.2.2, $nh\tilde{S}_m^2/3$ has an asymptotic distribution of $\sum_{\nu=1}^{\infty} \lambda_\nu(Z_\nu^2 - 1) + \theta_2$, we propose the following method for computing the p-value. The asymptotic distribution of $\sum_{\nu=1}^{\infty} \lambda_\nu(Z_\nu^2 - 1)$ is approximated by $\sum_{\nu=1}^{n} \hat{\lambda}_\nu(Z_\nu^2 - 1)$, which can be obtained by simulation. First, we obtain n eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ from the empirical matrix $\bar{\Psi}$ of the operator Ψ. Then, we randomly generate N sets of n standard normal random variables, $Z_{l,1}, \ldots, Z_{l,n}, l = 1, \ldots, N$, and then calculate N realizations of $\sum_{i=1}^{n} \hat{\lambda}_i(Z_{l,i}^2 - 1) + \hat{\theta}_2$ for $l = 1, \ldots, N$, where $\hat{\theta}_2$ is defined (4.4.4). Finally, the p-value is calculate as $1/N \sum_{l=1}^{N} I(nh\tilde{S}_m^2/3 < \sum_{i=1}^{n} \hat{\lambda}_i(Z_{l,i}^2 - 1) + \hat{\theta}_2)$.

4.2.4 Test-based Screening

We can use this test procedure to propose a new test-based variable screening procedure. Consider a random sample $(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)$, where $\mathbf{X_i} = (X_{1i}, \ldots, X_{pi})^T$ is a p-dimensional vector for $i = 1, \ldots, n$. Consider p marginal nonparametric regression functions

$$m_k(x) = \mathrm{E}(Y_i | X_{ki} = x) \quad \text{for } k = 1, \ldots, p \tag{4.2.5}$$

of Y on each variable X_k, and the corresponding p hypotheses testing problem.

$$H_k : m_k(x) = c \quad \text{for } k = 1, \ldots, p \tag{4.2.6}$$

Let $P_1, \ldots, P_p$ be the corresponding p-values. First, we order p-values, $P_{(1)} \leq \ldots \leq P_{(p)}$, and define $H_{(i)}$ to be the corresponding hypothesis with p-value $P_{(i)}$. Let c be the largest index j for which $P_{(j)} \leq \frac{j}{p}q$ and q be the desirable false discovery rate. The proposed procedure defines the set of active predictors to consist of the predictors corresponding to $H_{(i)}, i = 1, \ldots, c$. This method is introduced by Simes (1986) and shown by Benjamini and Hochberg (1995) to control the false discovery rate. Thus, this test-based screening procedure controls the false discovery rate.

4.3 Numerical Studies

4.3.1 Simulation Studies

We present the simulation results on the performance of the test-baed screening procedure. The covariate matrix $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is generate from a multivariate normal with mean zero, covariance matrices $\Sigma = (\sigma_{ij})_{p \times p}$, and $\epsilon \sim N(0, 1)$. We use three different covariance matrices: $\sigma_{ij} = 0.3^{|i-j|}, 0.5^{|i-j|}$, and $0.7^{|i-j|}$.We set the dimension of covariates p to be 2000 and the sample size n to be 200. We repeat the experiment 100 times. We consider the following three models:

4.(a) $Y = 2X_1 + 1.5X_5 + X_{20} + 2X_{21} + \epsilon$

4.(b) $Y = 1.5X_1 + 2\cos(\pi X_5) + 2X_{20} + 1.5\sin(\pi/2X_{21}) + \epsilon$

4.(c) $Y = 1.5X_1 + X_5^2 + 1.5X_{20} + 2\log(|X_{21}|) + \epsilon$

We recorded following three outcomes

R1: The minimum, 25%, 50%, 75% quantiles, and maximum submodel size.

R2: The proportion of times each individual active predictor is selected in a submodel.

R3: The proportion of times all active predictors are selected in a submodel.

We compare the result of test based screening procedure to the RV-SIS. For test based screening, we use two multiple testing procedure: Benjamini and Hochberg (BH) approach and Bonferroni correction. We use the false discovery rate (FDR) to be 0.01 and 0.05 for BH approach and $\alpha = 0.05$ for Bonferroni correction. For the RV-SIS, we randomly generate the auxiliary variable of size $p/2$ from uniform distribution, $U(0, 1)$.

Model 4.(a) is linear, and Models 4.(b) and 4.(c) includes additive nonlinear. All models have four active predictors (X_1, X_5, X_{20}, and X_{21}).

Tables 4.1, 4.2, and 4.3 present the simulation result for R1. Tables 4.4, 4.5 and 4.6 present the result for R2 and R3. Tables 4.1, 4.2 and 4.3 show that the submodel size of RV-SIS is smaller than test-based screening procedure for all Models and all covariance. Tables 4.4, 4.5, and 4.6 show that the test-based screening procedure with BH (FDR=0.1) approach and RV-SIS have similar performance of capturing the active predictors, but test-based procedure with Bonferroni correction performs slightly worse.

4.3.2 A Real Data Example

Hall and Miller (2009) and Li *et al.* (2012b) use screening procedures to identify the most influential genes for over-expression of a G protein-coupled receptor (Ro1) in mice in the Cardiomyopathy microarray dataset (Segal, Dahlquist, and Conklin, 2003). In this data set, the number of subjects $n = 40$ and the number of gene expressions $p = 6,319$. We apply the test-based screening and RV-SIS procedure to identify the influential genes.

For the RV-SIS, we randomly generate the auxiliary variable of size $\lfloor p/2 \rfloor$ from uniform distribution, $U(0,1)$. Th RV-SIS identifies that there are only three influential genes, (Msa.1166.0, Msa.2877.0, Msa.741.0). However, the test-based screening procedure with FDR=0.01 identifies that there are 41 influential genes, (Msa.1024.0, Msa.1088.0, Msa.1194.0, Msa.14089.0, Msa.14686.0, Msa.15442.0, Msa.15467.0, Msa.1590.0, Msa.1637.0, Msa.19368.0, Msa.2000.0, Msa.2134.0, Msa.21996.0, Msa.2228.0, Msa.22488.0, Msa.25655.0, Msa.26025.0, Msa.2668.0, Msa.27429.0, Msa.28021.0, Msa.2888.0, Msa.3097.0, Msa.3214.0, Msa.3248.0, Msa.3269.0, Msa.32975.0, Msa.33332.0, Msa.451.0, Msa.4636.0, Msa.5595.0, Msa.5727.0, Msa.657.0, Msa.668.0, Msa.736.0, Msa.7780.0, Msa.84.0, Msa.8671.0, Msa.8673.0, Msa.916.0, Msa.925.0, Msa.978.0).

None of influential genes that are identified by RV-SIS is included in the set of influential genes that is identified by the test-based screening procedure.

We fit the nonparametric regression and use the nonparametric coefficient of determination (Doksum and Samarov, 1995) to compare the performance. We obtain R^2-values of 0.9384 and 1 for RV-SIS and the test-based screening, respectively.

4.4 Theoretical Properties

4.4.1 Proof of Theorem 4.2.2

Let $K_{i_2 i_1}$ denotes $K(\frac{X_{i_2}-X_{i_1}}{h})$. Let $Y_i^* = Y_i - c$ be the centered Y_i under the null hypothesis. Then under H_0, the sample variance of the regression function, $\tilde{S}_m^2$ can be expressed as following,

$$\begin{aligned}
\tilde{S}^2_m &= \frac{1}{n}\sum_i \widehat{m}^2(X_i) - (\frac{1}{n}\sum_i \widehat{m}(X_i))^2 \\
&= \frac{1}{n}\sum_i (\widehat{m}(X_i) - c)^2 + 2c\frac{1}{n}\sum_i \widehat{m}(X_i) - c^2 \\
&\quad -(\frac{1}{n}\sum_i \widehat{m}(X_i) - c)^2 - \frac{2c}{n}\sum_i (\widehat{m}(X_i) - c) - c^2 \\
&= \frac{1}{n}\sum_i (\widehat{m}(X_i) - c)^2 - (\frac{1}{n}\sum_i \widehat{m}(X_i) - c)^2 \\
&= \frac{1}{n}\sum_{i_1} (\frac{1}{nh}\sum_{i_2} (Y_{i_2} - c)K_{i_2i_1}\frac{1}{\hat{f}(X_{i_1})})^2 - (\frac{1}{n^2h}\sum_{i_1}\sum_{i_2} (Y_{i_2} - c)K_{i_2i_1}\frac{1}{\hat{f}(X_{i_1})})^2 \\
&= \frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y^*_{i_2}Y^*_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{\hat{f}^2(X_{i_1})} - (\frac{1}{n^2h}\sum_{i_1}\sum_{i_2} Y^*_{i_2}K_{i_2i_1}\frac{1}{\hat{f}(X_{i_1})})^2 \\
&= \frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y^*_{i_2}Y^*_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})} - (\frac{1}{n^2h}\sum_{i_1}\sum_{i_2} Y^*_{i_2}K_{i_2i_1}\frac{1}{\hat{f}(X_{i_1})})^2 \\
&\quad +\frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y^*_{i_2}Y^*_{i_3}K_{i_2i_1}KX_{i_3i_1}(\frac{1}{\hat{f}^2(X_{i_1})} - \frac{1}{f^2(X_{i_1})}) \\
&\equiv T_1 + T_2 + T_3
\end{aligned}$$

For convenience in notation, we omit * from the centered Y^*_i for the rest of the proof. First, we consider T_1. Then the following representation of T_1 as a V-statistic and U-statistic is useful

$$\begin{aligned}
T_1 &= \frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})} \\
&= \frac{1}{n^3h}\sum_{i_1}\sum_{i_2}\sum_{i_3} \psi(Z_{i_1}, Z_{i_2}, Z_{i_3}) \\
&= \frac{6}{n^3h}\binom{n}{3}U^{(3)} + \frac{6}{n^3h}\binom{n}{2}U^{(2)} + \frac{1}{n^3h}\binom{n}{1}U^{(1)}
\end{aligned}$$

where $U^{(j)}$ is a U-statistic of degree j. Define $H_n^{(2)}$ be

$$H_n^{(2)} = \binom{n}{2}^{-1} \sum_{i_1 < i_2} \psi_2(Z_{i_1}, Z_{i_2}).$$

Since $E(\psi(Z_1, Z_2, Z_3)) = 0$ and $\psi_1(z_1) = E(\psi(Z_1, Z_2, Z_3)|Z_1 = z_1) = 0$, $nU_n^{(3)}$ and $3nH_n^{(2)}$ have the same asymptotic distribution. By Corollary 1 of Lee (1990, pp 83), we have

$$nh(\frac{6}{n^3h}\binom{n}{3}U^{(3)}) \to 3\sum_{\nu=1}^{\infty}\lambda_\nu(Z_\nu^2 - 1)$$

and

$$nh(\frac{6}{n^3h}\binom{n}{2}U^{(2)} + \frac{1}{n^3h}\binom{n}{1}U^{(1)}) \to 3\theta_2 \qquad \text{as } n \to \infty$$

Therefore,

$$nhT_1 \to 3\sum_{\nu=1}^{\infty}\lambda_\nu(Z_\nu^2 - 1) + 3\theta_2 \tag{4.4.1}$$

Consider now T_2. We show that $nhT_2 \to 0$ in probability. First, we show $\sqrt{nhT_2} \to 0$ in probability by showing $E(\sqrt{nhT_2})^2 \to 0$ as n approaches to infinity.

Under the null hypothesis, we have

$$\begin{aligned} E(\sqrt{nhT_2})^2 &= E(\frac{1}{n}\sum_{i_1}\sum_{i_2} Y_{i_2}K(\frac{X_{i_2} - X_{i_1}}{h})\frac{1}{\hat{f}(X_{i_1})}) \\ &= E(\frac{1}{n}\sum_{i_1}\sum_{i_2} \epsilon_{i_2}K(\frac{X_{i_2} - X_{i_1}}{h})\frac{1}{\hat{f}(X_{i_1})}) \\ &\to 0 \quad \text{as } n \to \infty \end{aligned}$$

Thus we have $\sqrt{nhT_2} \to 0$ in probability. In addition, we have

$$nhT_2 \to 0 \quad \text{in probability.} \tag{4.4.2}$$

Now we consider T_3. For some constantC_1, we can express T_3 as following

$$\begin{aligned}
T_3 &= \frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}(\frac{1}{\hat{f}^2(X_{i_1})} - \frac{1}{f^2(X_{i_1})}) \\
&= \frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}(\frac{(f(X_i) - \hat{f}(X_i))(f(X_i) + \hat{f}(X_i))}{\hat{f}^2(X_{i_1})f^2(X_{i_1})}) \\
&\approx C_0(\sqrt{\frac{log(n)}{n}} + h^2)\frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}(\frac{f(X_i) + \hat{f}(X_i)}{\hat{f}^2(X_{i_1})f^2(X_{i_1})}) \text{ by Lemma 1} \\
&\leq C_1(\sqrt{\frac{log(n)}{n}} + h^2)(\frac{\sup_{X_i}(f(X_i) + \hat{f}(X_i))}{\inf_{X_i}\hat{f}^2(X_{i_1})})\frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})} \\
&= C_1(\sqrt{\frac{log(n)}{n}} + h^2)C_2\frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})}
\end{aligned}$$

where $C_2 = (\sup_{X_i}(f(X_i) + \hat{f}(X_i))/\inf_{X_i}\hat{f}^2(X_{i_1}))$.

Using this expression, we show that nhT_3 converges to zero in probability by showing that $E(nhT_3)^2$ approaches to zero as $n \to 0$.

$$\begin{aligned}
& E((nhC_1(\sqrt{\frac{log(n)}{n}} + h^2)C_2\frac{1}{n^3h^2}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})})^2) \\
= \;& O(\frac{log(n)}{n} + h^4 + \sqrt{\frac{log(n)}{n}}h^2)E((\frac{1}{n^2h}\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_2}Y_{i_3}K_{i_2i_1}K_{i_3i_1}\frac{1}{f^2(X_{i_1})})^2) \\
= \;& O(\frac{log(n)}{n} + h^4 + \sqrt{\frac{log(n)}{n}}h^2)\frac{1}{n^4h^2} \times \\
& E(\sum_{i_1}\sum_{i_2}\sum_{i_3}\sum_{i_4}\sum_{i_5}\sum_{i_6} Y_{i_3}Y_{i_4}Y_{i_5}Y_{i_6}K_{i_3i_1}K_{i_4i_1}K_{i_5i_2}K_{i_6i_2}\frac{1}{f^2(X_{i_1})f^2(X_{i_2})})
\end{aligned}$$

$$\begin{aligned}
&= O(\frac{log(n)}{n} + h^4 + \sqrt{\frac{log(n)}{n}}h^2)\frac{1}{n^4h^2} \times \{E(\sum_{i_1}\sum_{i_2}\sum_{i_3} Y_{i_3}^4 K_{i_3i_1}^2 K_{i_3i_2}^2 \frac{1}{f^2(X_{i_1})f^2(X_{i_2})}) \\
&\quad +E(\sum_{i_1}\sum_{i_2}\sum_{i_3\neq}\sum_{i_5} Y_{i_3}^2 Y_{i_5}^2 K_{i_3i_1}^2 K_{i_5i_2}^2 \frac{1}{f^2(X_{i_1})f^2(X_{i_2})}) \\
&\quad +2E(\sum_{i_1}\sum_{i_2}\sum_{i_3\neq}\sum_{i_4} Y_{i_3}^2 Y_{i_4}^2 K_{i_3i_1} K_{i_3i_2} K_{i_4i_1} K_{i_4i_2} \frac{1}{f^2(X_{i_1})f^2(X_{i_2})})\} \\
&\to 0 \quad \text{as } n \to 0
\end{aligned}$$

Thus, we have

$$nhT_3 \to 0 \quad \text{in probability} \tag{4.4.3}$$

Therefore, by 4.4.1, 4.4.2, and 4.4.3 we have

$$nS_m^2 \to 3\sum_{\nu=1}^{\infty} \lambda_\nu (Z_\nu^2 - 1) + 3\theta_2$$

4.4.2 Estimating the percentile of the limiting distribution

Following estimations are required to estimate the percentile of the limiting distribution in Section 4.2.3. Using the uniform kernel, we estimate $\psi_2(Z_1, Z_2)$ as following,

$$\begin{aligned}
&E(\psi(Z_1, Z_2, Z_3)|Z_1 = z_1, Z_2 = z_2) \\
&= E(\frac{1}{3h}(Y_2Y_3K_{21}K_{31}\frac{1}{f^2(X_1)} + Y_1Y_3K_{12}K_{32}\frac{1}{f^2(X_2)} + Y_1Y_2K_{13}K_{23}\frac{1}{f^2(X_3)}|Z_1 = z_1, Z_2 = z_2) \\
&= E(\frac{1}{3h}Y_1Y_2K_{13}K_{23}\frac{1}{f^2(X_3)}|Z_1 = z_1, Z_2 = z_2) \\
&= \int \frac{1}{3h}y_1y_2K_{13}K_{23}\frac{1}{f(X_3)}dX_3
\end{aligned}$$

$$
\begin{aligned}
= \quad & \frac{y_1y_2}{3h}\int K(\frac{x_1-X_3}{h})K(\frac{x_2-X_3}{h})\frac{1}{f(X_3)}I(0\le |x_2-x_1|<2h)dX_3 \\
& \frac{y_1y_2}{3}\int_{-1+\frac{x_2-x_1}{h}}^{1} K(u)K(\frac{x_2-x_1}{h}-u)\frac{1}{f(x_1-hu)}I(0\le |x_2-x_1|<2h)I(x_2>x_1)du \\
& +\frac{y_1y_2}{3}\int_{-1}^{1+\frac{x_2-x_1}{h}} K(u)K(\frac{x_2-x_1}{h}-u)\frac{1}{f(x_1-hu)}I(0\le |x_2-x_1|<2h)I(x_2<x_1)du \\
\approx \quad & \frac{y_1y_2}{3}\frac{1}{4}\frac{1}{f(x_1)}\int I(-1+\frac{x_2-x_1}{h}\le u\le 1)duI(0\le |x_2-x_1|<2h)I(x_2>x_1)du \\
& +\frac{y_1y_2}{3}\frac{1}{4}\frac{1}{f(x_1)}\int I(-1\le u\le 1+\frac{x_2-x_1}{h})duI(0\le |x_2-x_1|<2h)I(x_2<x_1)du \\
= \quad & \frac{y_1y_2}{12}\frac{1}{f(x_1)}(2-|\frac{x_2-x_2}{h}|)I(0\le |x_2-x_1|<2h)
\end{aligned}
$$

We can express θ_2 as following,

$$
\begin{aligned}
\theta_2 &= E\psi(Z_1,Z_1,Z_2) \\
&= E(\frac{1}{3h}(Y_1Y_2K_{11}K_{21}\frac{1}{f^2(X_1)}+Y_1Y_2K_{11}K_{21}\frac{1}{f^2(X_1)}+Y_1^2K_{21}^2\frac{1}{f^2(X_2)})) \\
&= E(\frac{1}{3h}Y_1^2K_{21}^2\frac{1}{f^2(X_2)}) \\
&= E(\frac{1}{3h}\sigma^2(X_1)K_{21}^2\frac{1}{f^2(X_2)}) \\
&= E(\frac{1}{3h}E(\sigma^2(X_1)K^2(\frac{X_2-X_1}{h})\frac{1}{f^2(X_2)}|X_2=x_2)) \\
&= E(\frac{1}{3h}\frac{1}{f^2(X_2)}\int \sigma^2(X_2-hu)K^2(u)f(X_2-hu)hdu \\
&= E(\frac{\int K^2(u)du}{3}\frac{1}{f^2(X_2)}(\sigma^2(X_2)f(X_2)+O(h^2))) \\
&\approx \frac{\int K^2(u)du}{3}E(\frac{\sigma^2(X_2)}{f(X_2)})
\end{aligned}
$$

Thus we can estimate θ by $\hat{\theta}$, where

$$\hat{\theta} \;=\; \frac{1}{6n}\sum_{i=1}^{n}(\frac{S_X^2}{\hat{f}(X_i)}) \qquad (4.4.4)$$

by using the uniform kernel.

Table 4.1. The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.3^{|i-j|}$

Model 4.(a)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	513.00	615.00	648.50	685.20	754.00
Test-based (BH, FDR = 0.01)	51.00	75.00	87.00	144.00	163.00
Test-based (Bonferroni, $\alpha = 0.05$)	51.00	71.75	79.00	87.00	103.00
RV-SIS	4.00	21.00	39.50	60.50	206.00
Model 4.(b)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	539.00	621.00	649.50	675.20	728.00
Test-based (BH, FDR = 0.01)	56.00	69.75	75.50	140.00	163.00
Test-based (Bonferroni, $\alpha = 0.05$)	56.00	69.00	74.00	80.25	104.00
RV-SIS	4.00	20.00	31.50	47.00	144.00
Model 4.(c)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	502.00	608.00	625.50	653.00	737.00
Test-based (BH, FDR = 0.01)	76.00	98.00	106.00	116.00	223.00
Test-based (Bonferroni, $\alpha = 0.05$)	76.00	98.00	106.00	115.20	156.00
RV-SIS	3.00	31.00	47.00	63.25	184.00

Table 4.2. The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.5^{|i-j|}$

Model 4.(a)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	512.00	625.00	653.50	694.20	784.00
Test-based (BH, FDR = 0.01)	53.00	74.75	136.00	127.00	181.00
Test-based (Bonferroni, $\alpha = 0.05$)	53.00	73.75	80.50	87.00	128.00
RV-SIS	5.00	24.75	39.50	62.25	140.00
Model 4.(b)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	517.00	626.00	661.50	698.20	758.00
Test-based (BH, FDR = 0.01)	56.00	72.00	81.00	138.00	170.00
Test-based (Bonferroni, $\alpha = 0.05$)	56.00	71.75	78.00	84.00	99.00
RV-SIS	5.00	20.75	37.50	54.75	105.00
Model 4.(c)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	451.00	608.00	644.00	672.00	761.00
Test-based (BH, FDR = 0.01)	71.00	100.00	113.00	125.20	225.00
Test-based (Bonferroni, $\alpha = 0.05$)	71.00	100.00	113.00	122.00	147.00
RV-SIS	4.00	22.75	38.50	68.25	136.00

Table 4.3. The minimum, 25%, 50%, 75% quantiles, and maximum selected submodel size when $\sigma_{ij} = 0.7^{|i-j|}$

Model 4.(a)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	563.00	638.80	667.50	700.50	791.00
Test-based (BH, FDR = 0.01)	59.00	86.50	141.00	150.20	191.00
Test-based (Bonferroni, $\alpha = 0.05$)	59.00	79.00	84.00	93.00	126.00
RV-SIS	10.00	28.00	42.50	71.00	199.00
Model 4.(b)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	554.00	620.00	652.50	698.00	781.00
Test-based (BH, FDR = 0.01)	55.00	74.00	134.00	142.20	165.00
Test-based (Bonferroni, $\alpha = 0.05$)	55.00	72.75	80.00	86.25	103.00
RV-SIS	7.00	23.00	35.00	64.00	125.00
Model 4.(c)					
	Min	25%	50%	75%	Max
Test-based (BH, FDR = 0.05)	460.00	610.80	644.50	687.00	782.00
Test-based (BH, FDR = 0.01)	66.00	102.00	113.50	127.20	223.00
Test-based (Bonferroni, $\alpha = 0.05$)	66.00	102.00	113.00	124.20	155.00
RV-SIS	8.00	31.75	47.00	67.25	162.00

Table 4.4. The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.3^{|i-j|}$

Model 4.(a)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	1.00	0.99	1.00	1.00	0.99
Test-based (Bonferrnoi, $\alpha = 0.05$)	1.00	0.99	0.99	1.00	0.99
RV-SIS	1.00	0.98	1.00	1.00	0.98
Model 4.(b)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	1.00	0.99	1.00	1.00	0.99
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.99	0.99	1.00	1.00	0.98
RV-SIS	1.00	0.99	1.00	1.00	0.99
Model 4.(c)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	0.98	0.90	0.99	1.00	0.88
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.98	0.90	0.99	1.00	0.88
RV-SIS	0.99	0.98	1.00	1.00	0.98

Table 4.5. The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.5^{|i-j|}$

Model 4.(a)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	1.00	0.99	1.00	1.00	0.99
Test-based (Bonferrnoi, $\alpha = 0.05$)	1.00	0.98	1.00	1.00	0.98
RV-SIS	1.00	0.98	1.00	1.00	0.98
Model 4.(b)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	0.99	0.98	1.00	1.00	0.97
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.99	0.96	1.00	1.00	0.95
RV-SIS	1.00	0.98	1.00	1.00	0.98
Model 4.(c)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	0.99	0.91	1.00	1.00	0.90
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.99	0.90	1.00	1.00	0.89
RV-SIS	0.99	0.97	0.99	1.00	0.96

Table 4.6. The proportion of times each individual active predictor is selected and all active predictors are selected in a submodel when $\sigma_{ij} = 0.7^{|i-j|}$

Model 4.(a)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	1.00	1.00	1.00	1.00	1.00
Test-based (Bonferrnoi, $\alpha = 0.05$)	1.00	1.00	1.00	1.00	1.00
RV-SIS	1.00	1.00	1.00	1.00	1.00
Model 4.(b)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	1.00	1.00	1.00	1.00
Test-based (BH, FDR = 0.01)	1.00	0.99	1.00	1.00	0.99
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.99	0.98	1.00	1.00	0.97
RV-SIS	0.98	0.97	1.00	1.00	0.96
Model 4.(c)					
	X_1	X_5	X_{20}	X_{21}	ALL
Test-based (BH, FDR = 0.05)	1.00	0.99	1.00	1.00	0.99
Test-based (BH, FDR = 0.01)	0.96	0.93	1.00	1.00	0.89
Test-based (Bonferrnoi, $\alpha = 0.05$)	0.96	0.92	0.99	1.00	0.87
RV-SIS	0.99	0.97	1.00	1.00	0.96

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, we reviewed the existing sure independence screening procedures for analyzing the ultrahigh dimensional data. We proposed a new sure independence screening procedure using the variance of the regression function (RV-SIS) in Chapter 3 and a hypothesis testing of a constant regression function and test-based screening in Chapter 4.

In Chapter 3, RV-SIS is introduced for variable screening in ultrahigh dimensional setting. The RV-SIS procedure is based on their predictive significance while other existing screening methods fail to discern between variables that have predictive significance from those that influence the variance function. We presented the theoretical properties of the RV-SIS and showed that the RV-SIS possesses the sure independence screening property. The simulation studies showed that the performance of RV-SIS is considerably better than other screening procedures.

In Chapter 4, we presented a hypothesis testing of a constant regression function.

Under the hypothesis of a constant regression function, the regression function has zero variance. The asymptotic distribution of the estimated variance of the constant regression function was presented. Using p-value from the hypothesis test, we introduced a variable screening procedure using multiple testing. The test-based screening procedure controls the false discovery rate by using the Benjamini and Hochberg (1995) approach. This approach does not require to set a thresholding value for the submodel. Numerical studies showed the performance of the proposed procedure.

5.2 Future Work

We are in the process of developing a goodness-of-fit test for testing the significant of the kth covariate based on the variance, $\sigma^2_{m_k}$, of the regression function $m_k(x) = \mathrm{E}(Y|X_k = x)$. Again the idea is that $m_k(x)$ is constant if and only if $\sigma^2_{m_k} = 0$. Define the selected covariate in the submodel as $\mathbf{X}_{\mathcal{A}}$ from the RV-SIS procedure and its complement as $\mathbf{X}_{\mathcal{A}^c}$. There can be some covariates that are jointly influential, but not marginally, and these covariates will not be selected using the RV-SIS. We develop a test statistic that provides a way for these covariates to be included in the submodel. The next step will be to apply selected covariates $\mathbf{X}_{\mathcal{A}}$ to get the residuals

$$e_i = Y_i - \widehat{m}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A},i})$$

For estimating $\widehat{m}_{\mathcal{A}}(\cdot)$, we will use the single index model. For each covariate X_k in $\mathbf{X}_{\mathcal{A}^c}$, we perform a goodness-of-fit test for each covariate using the variance $\sigma^{*2}_{m_k}$ of the regression function $m_k^*(x) = \mathrm{E}(e_i|X_{ki} = x)$. Specifically, we want to test

$$H_0 : \sigma_{m_k}^{*2} = 0 \tag{5.2.1}$$

for each X_k in $\mathbf{X}_{\mathcal{A}^c}$. Then, we include covariates from $\mathbf{X}_{\mathcal{A}^c}$ that are highly significant in the submodel.

Bibliography

Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Candes E, Tao T (2007). "The Dantzig selector: statistical estimation when p is much larger than n." *The Annals of Statistics*, pp. 2313–2351.

Doksum K, Samarov A (1995). "Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression." *The Annals of Statistics*, **23**(5), 1443–1473.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). "Least angle regression." *The Annals of statistics*, **32**(2), 407–499.

Fan J, Feng Y, Song R (2011). "Nonparametric independence screening in sparse ultra-high-dimensional additive models." *Journal of the American Statistical Association*, **106**(494).

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

Fan J, Lv J (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(5), 849–911.

Fan J, Samworth R, Wu Y (2009). "Ultrahigh dimensional feature selection: beyond the linear model." *The Journal of Machine Learning Research*, **10**, 2013–2038.

Fan J, Song R (2010). "Sure independence screening in generalized linear models with NP-dimensionality." *The Annals of Statistics*, **38**(6), 3567–3604.

Hall P, Miller H (2009). "Using generalized correlation to effect variable selection in very high dimensional problems." *Journal of Computational and Graphical Statistics*, **18**(3).

Hansen BE (2008). "Uniform convergence rates for kernel estimation with dependent data." *Econometric Theory*, **24**(03), 726–748.

Kim D, Li R, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD (2014). "Knowledge-driven genomic interactions: an application in ovarian cancer." *BioData mining*, **7**(1), 20.

Koltchinskii V, Giné E (2000). "Random matrix approximation of spectra of integral operators." *Bernoulli*, pp. 113–167.

Lee J (1990). *U-statistics: Theory and Practice*. CRC Press.

Li G, Peng H, Zhang J, Zhu L (2012a). "Robust rank correlation based screening." *The Annals of Statistics*, **40**(3), 1846–1877.

Li R, Zhong W, Zhu L (2012b). "Feature screening via distance correlation learning." *Journal of the American Statistical Association*, **107**(499), 1129–1139.

McFadden D, Puig C, Kirschner D (1977). "Determinants of the long-run demand for electricity." In *Proceedings of the American Statistical Association*, volume 1, pp. 109–19. Business and Economics Section.

Segal MR, Dahlquist KD, Conklin BR (2003). "Regression approaches for microarray data analysis." *Journal of Computational Biology*, **10**(6), 961–980.

Serfling RJ (2009). *Approximation theorems of mathematical statistics*, volume 162. Wiley. com.

Simes RJ (1986). "An improved Bonferroni procedure for multiple tests of significance." *Biometrika*, **73**(3), 751–754.

Székely GJ, Rizzo ML, *et al.* (2009). "Brownian distance covariance." *The Annals of Applied Statistics*, **3**(4), 1236–1265.

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

Wang L, Akritas MG, Van Keilegom I (2008). "An ANOVA-type nonparametric diagnostic test for heteroscedastic regression models." *Journal of Nonparametric Statistics*, **20**(5), 365–382.

Zhu LP, Li L, Li R, Zhu LX (2011). "Model-free feature screening for ultrahigh-dimensional data." *Journal of the American Statistical Association*, **106**(496).

Zou H (2006). "The adaptive lasso and its oracle properties." *Journal of the American statistical association*, **101**(476), 1418–1429.

Zou H, Hastie T (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

Vita

Won Chul Song

Education

The Pennsylvania State University, University Park, PA.
Ph.D. Candidate, Statistics (expected graduation date: 2015)

- Academic advisor: Dr. Michael G. Akritas
- Dissertation Topic: *Nonparametric Independence Screening and Test-based Screening via the Variance of the Regression Function*

Brigham Young University - Idaho, Rexburg, ID.
B.S., Applied Mathematics, April 2009

Research Interests

Nonparametric/Semiparametric Regression, Ultrahigh-demensional Analysis, Hypothesis Testing, Datamining

Teaching Experience

Instructor

STAT200, STAT200 (Online), STAT319, STAT401, STAT418

Teaching Assistant

STAT 414, STAT 462, STAT 480, STAT 503

Publications

Submitted (all peer-reviewed):

1. **Song, W.** and Akritas, M., Nonparametric Independence Screening via the Variance of the Regression Function.

In preparation:

1. **Song, W.** and Akritas, M., Hypothesis Testing of a Constant Regression Function and Test-based Screening.
2. **Song, W.**, Kim, D., Ritche, M., Identifying Genomic Interactions Associated with Stage in Ovarian Cancer Using Sure Independence Screening Approach.
3. Mastro, A., **Song, W.**, and Huang, Y., The Effect of Whey on Exercise Induced Changes in Circulating Immune Cell Populations.