Published in final edited form as:

Sci China Math. 2017 March ; 60(3): 551–568. doi:10.1007/s11425-016-0186-8.

Model-Free Conditional Independence Feature Screening For Ultrahigh Dimensional Data

Luheng Wang¹, Jingyuan Liu², Yong Li³, and Runze Li⁴

¹School of Mathematics, Beijing Normal University, Beijing 100875, P.R. China

²Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China

³School of Statistics, Beijing Normal University, Beijing 100875, P.R. China

⁴Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA 16802-2111, USA

Abstract

Feature screening plays an important role in ultrahigh dimensional data analysis. This paper is concerned with conditional feature screening when one is interested in detecting the association between the response and ultrahigh dimensional predictors (e.g., genetic makers) given a low-dimensional exposure variable (such as clinical variables or environmental variables). To this end, we first propose a new index to measure conditional independence, and further develop a conditional screening procedure based on the newly proposed index. We systematically study the theoretical property of the proposed procedure and establish the sure screening and ranking consistency properties under some very mild conditions. The newly proposed screening procedure enjoys some appealing properties. (a) It is model-free in that its implementation does not require a specification on the model structure; (b) it is robust to heavy-tailed distributions or outliers in both directions of response and predictors; and (c) it can deal with both feature screening and the conditional screening in a unified way. We study the finite sample performance of the proposed procedure by Monte Carlo simulations and further illustrate the proposed method through two real data examples.

Keywords

Feature screening; Conditional feature screening; High dimensional data; Variable selection

1 Introduction

With the remarkable improvements in computing power and storage technology, high-dimensional data are frequently collected in a variety of research areas such as genome-wide association studies, biomedical imaging, proteomics studies, finance and tumor classification. For high dimensional data, the number of variables p can be much larger than

the sample size n . This poses many challenges and provides opportunities for statistical analyses (Donoho, 2000; Fan and Li, 2006). The sparsity assumption, which implies that only a small subset of predictors contribute to the response, is frequently adopted and deemed useful in high-dimensional data analysis. Thus, variable selection and feature screening play critical roles in the analysis of high dimensional data. Regularization methods that automatically result in sparse models have been developed in the literature. Popular regularization methods include Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006) and Dantzig selector (Candes and Tao, 2007). These methods have been successfully applied in high-dimensional data analysis, but they may not perform well for analyzing ultrahigh dimensional data. Their computational expediency, statistical accuracy, and algorithmic stability become major concerns (Fan, Samworth and Wu, 2009). Feature screening with low cost computation is appeal for analysis of ultrahigh dimensional data.

Various feature screening procedures have been proposed to reduce the dimensionality p from ultrahigh scale (e.g., $p = \exp(O(n^a))$ for some $a > 0$) to a large scale d (e.g., $d = \lfloor n/\log(n) \rfloor$ or $d = n-1$). A selective review for feature screening procedures can be found at Liu, Zhong and Li (2015). Fan and Lv (2008) proposed the sure independent screening (SIS) and iterative SIS (ISIS) procedure based on marginal Pearson correlation. Hall and Miller (2009) proposed a feature screening procedure using generalized correlation which can characterize both linear and nonlinear relationships between the predictor and response variables. Li, Peng, Zhang and Zhu (2012) investigated the Kendall τ correlation to identify the important variables, which is robust to heavy-tailed distributions. Fan and Song (2010) extended the SIS to generalized linear models using the maximum marginal likelihood estimates. Fan, Feng and Song (2011) proposed a nonparametric independence screening for additive models based on B-spline method, and Fan, Ma and Dai (2014) and Liu, Li and Wu (2014) further extended the nonparametric independence screening method to varying coefficient models. The performance of the aforementioned methods are effective when the model is correctly specified but may be poor when the model is mis-specified. This issue motivates researchers to develop model-free feature screening methods. Zhu, Li, Li and Zhu (2011) developed a screening procedure under a very general semiparametric model framework, which includes many commonly-used models as special cases thereof. Li, Zhong and Zhu (2012) developed a model-free distance correlation based screening procedure, which can handle multivariate response and grouped predictors but is not robust to heavy-tailed distributions of predictors and the presence of potential outliers. Mai and Zou (2012) proposed a sure screening procedure based on Kolmogorov distance, which is only suitable for binary classification problems. Cui, Li and Zhong (2014) developed a model-free sure independence procedure for ultrahigh dimensional discriminant analysis. All aforementioned procedures cannot be used for conditional feature screening given a low-dimensional variables. This paper aims to fill the gap by developing a model-free conditional feature screening for ultrahigh dimensional data.

This work was motivated by empirical analyses of genetic data consisting of a huge number of genetic markers and a low dimensional environmental factors. Of interest is to examine the potential genetic effect on a phenotype given a certain environmental factor. Without further specifying a model structure, we formulate such a problem as detecting conditional

dependence between the phenotype and genetic markers given the environmental factors and propose a new index to detect the conditional association. Aforementioned SIS procedures may fail easily to detect conditional association. To see this, let us consider a toy example. Suppose that $P(Z = \pm 1) = P(X = \pm 1) = 0.5$, X and Z are independent. Let $Y = XZ$, it is easy to check that Y is independent of X . Thus, the marginal independent screening procedure cannot detect the association between X and Y without using the information of Z . This toy example indicates that it is more desirable to consider conditional feature screening. We will further illustrate the importance of conditional feature screening in our numerical examples in Section 3.

In this paper we first propose a new index to measure conditional independence. The new index equals zero if and only if two random variables are conditionally independent. The new index is invariant under strictly monotone transformation and can be easily evaluated without involving any numerical optimization algorithm. Based on this new index, we further develop a model free conditional feature screening procedure. We investigate the theoretical properties of the proposed procedure, and establish the sure screening property and ranking consistency property under mild conditions. The proposed conditional feature screening procedure enjoys several appealing properties. (a) The proposed conditional feature screening procedure becomes a feature screening procedure in the absence of conditional variable. Thus, the proposed procedure can be used for both feature screening and conditional feature screening. (b) It is model-free in that its implementation does not require to specify a particular model structure and (c) it allows heavy-tailed error distributions or outliers in both response and predictors. We conduct Monte Carlo simulation to examine the finite sample performance of the proposed procedure and compare with existing feature screening methods. Our numerical results show that our proposed procedure outperforms commonly-used methods. We further illustrate the proposed procedure by two real data examples.

The rest of this paper is organized as follows. In section 2, we propose a new index to measure conditional independence, and develop a conditional feature screening procedure based on the new index. We systematically study the theoretical properties of the proposed index and procedure. In section 3, we conduct various Monte Carlo simulations to investigate the finite sample performance of the proposed procedure, and illustrate the new procedure for two real data examples. Concluding remarks are given in section 4 and the technical proofs are relegated to the Appendix.

2 A New feature screening procedure

In this section, we first propose a new index to detect conditional independence, and then further develop a conditional feature screening based on the new index.

2.1 A New Index for Conditional Independence

Suppose that X , Y and Z are random variables with supports $\mathbb{R}_Y, \mathbb{R}_X, \mathbb{R}_Z$ subsets of the real line $\mathbb{R}$, respectively. Define $F_Z = P(Z = z)$. Let $\{\tilde{X}, \tilde{Y}, \tilde{Z}\}$ be an independent copy of $\{X, Y, Z\}$. Denote $F_{\tilde{Z}} = E\{I(Z = \tilde{Z})|\tilde{Z}\}$, where $I(\cdot)$ is the indicator function. With slight abuse of notation, we let $P(Z = \tilde{Z})$ stand for $E\{I(Z = \tilde{Z})|\tilde{Z}\}$. Thus, $F_{\tilde{Z}} = P(Z = \tilde{Z})$. Similarly, denote

$F_{\tilde{Y}, \tilde{Z}} = \mathbb{P}(Y \leq \tilde{Y}, Z \leq \tilde{Z}), F_{\tilde{Y}|\tilde{Z}} = \mathbb{P}(Y \leq \tilde{Y} | Z \leq \tilde{Z}), F_{\tilde{Y}, \tilde{X}, \tilde{Z}} = \mathbb{P}(Y \leq \tilde{Y}, X \leq \tilde{X}, Z \leq \tilde{Z}), F_{\tilde{Y}, \tilde{X}|\tilde{Z}} = \mathbb{P}(Y \leq \tilde{Y}, X \leq \tilde{X} | Z \leq \tilde{Z})$. Similarly, we define the notation $F_{\tilde{X}, \tilde{Z}}$ and $F_{\tilde{X}|\tilde{Z}}$. Thus, Y and X are independent given Z if and only if $F_{\tilde{Y}, \tilde{X}|\tilde{Z}} = F_{\tilde{Y}|\tilde{Z}} F_{\tilde{X}|\tilde{Z}}$ for any $\tilde{X} \in \mathbb{R}_X, \tilde{Y} \in \mathbb{R}_Y, \tilde{Z} \in \mathbb{R}_Z$ which is equivalent with

$$F_{\tilde{Y}, \tilde{X}, \tilde{Z}} F_{\tilde{Z}} = F_{\tilde{Y}, \tilde{Z}} F_{\tilde{X}, \tilde{Z}}. \quad (2.1)$$

It is natural to define the following index to measure conditional independence based on Cramer-von Mises statistic for conditional independence

$$MC(Y, X|Z) = E \left[F_{\tilde{Y}, \tilde{X}, \tilde{Z}} F_{\tilde{Z}} - F_{\tilde{Y}, \tilde{Z}} F_{\tilde{X}, \tilde{Z}} \right]^2. \quad (2.2)$$

Here ‘M’ and ‘C’ are picked up from the first letter of Mises and conditional, respectively. By its definition, $MC(Y, X|Z) = MC(X, Y|Z)$, and $MC(Y, X|Z) = 0$ if and only if X and Y are conditionally independent given Z . This property implies that $MC(Y, X|Z)$ can be served as a marginal utility for model free conditional feature screening proposed in this paper.

Let $\{(X_i, Y_i, Z_i)\}, i = 1, \dots, n$, be a random sample from population (X, Y, Z) . Natural estimators for $F_{\tilde{Z}}, F_{\tilde{Y}, \tilde{Z}}, F_{\tilde{X}, \tilde{Z}}, F_{\tilde{Y}, \tilde{X}, \tilde{Z}}$ are their sample counterparts:

$$\begin{aligned} \hat{F}_{\tilde{Z}} &= \hat{\mathbb{P}}(Z \leq \tilde{Z}) = \frac{1}{n} \sum_{i=1}^n I(Z_i \leq \tilde{Z}), \\ \hat{F}_{\tilde{Y}, \tilde{Z}} &= \hat{\mathbb{P}}(Y \leq \tilde{Y}, Z \leq \tilde{Z}) = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq \tilde{Y}, Z_i \leq \tilde{Z}), \\ \hat{F}_{\tilde{X}, \tilde{Z}} &= \hat{\mathbb{P}}(X \leq \tilde{X}, Z \leq \tilde{Z}) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq \tilde{X}, Z_i \leq \tilde{Z}), \\ \hat{F}_{\tilde{Y}, \tilde{X}, \tilde{Z}} &= \hat{\mathbb{P}}(Y \leq \tilde{Y}, X \leq \tilde{X}, Z \leq \tilde{Z}) = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq \tilde{Y}, X_i \leq \tilde{X}, Z_i \leq \tilde{Z}), \end{aligned} \quad (2.3)$$

where $I(\cdot)$ is the indicator function. Then, we estimate the index $MC(Y, X|Z)$ by the following estimator

$$\widehat{MC}(Y, X|Z) = \frac{1}{n} \sum_{i=1}^n \left(\hat{F}_{Y_i, X_i, Z_i} \hat{F}_{Z_i} - \hat{F}_{Y_i, Z_i} \hat{F}_{X_i, Z_i} \right)^2 \quad (2.4)$$

From the definition above, it is easy to check that MC and its estimator are invariant with respect to strictly monotone transformations. That is, for any arbitrary strictly monotone increasing function $f(\cdot), g(\cdot)$ and $h(\cdot)$, it follows that

$$MC(Y, X|Z) = MC(f(Y), g(X)|h(Z)), \text{ and } \widehat{MC}(X, Y|Z) = \widehat{MC}(f(Y), g(X)|h(Z)).$$

This property implies that the screening procedure based on MC may be more effective than the marginal Pearson correlation in the presence of nonlinear relationship between the important variables and the response. One may define $MC(Y, X|Z)$ and $\widehat{MC}(Y, X|Z)$ for multidimensional random vectors X , Y and Z . However, $\widehat{MC}(Y, X|Z)$ may require more samples in order to have good estimate as the dimensions of X , Y and Z increase. Thus, we would focus on finite fixed dimensional cases.

2.2 Sure Conditional Independence Screening

This section is devoted to developing a conditional feature screening procedure using the index MC. Let Y be the one-dimensional response variable, $\mathbf{x} = (X_1, \dots, X_p)^T$ be the predictor vector, and Z be the one-dimensional predictor variable. Denote $\{Y_i, \mathbf{x}_i, Z_i, i = 1, \dots, n\}$ be the random sample from $\{Y, \mathbf{x}, Z\}$. Following the convention in the literature (Li, Zhong and Zhu, 2012; Zhu, Li, Li and Zhu, 2011), we define the active set of Y given Z without specifying any regression model:

$$\mathcal{D} = \{k: F_{Y|Z, \mathbf{x}} \text{ functionally depends on } X_k \text{ for some } y \in \mathbb{R}_{Y|Z}\},$$

where $\mathbb{R}_{Y|Z}$ is the support of Y given Z , and denote by $I = \{1, \dots, p\} \setminus \mathcal{D}$ be the inactive predictor subset of $Y|Z$. Define $w_k = MC(Y, X_k|Z)$ to measure conditional independence between Y and X_k given Z . In other words, w_k can be used to be a marginal utility to measure importance of X_k on Y given Z . Let $\hat{w}_k = \widehat{MC}(Y, X_k|Z)$. Intuitively, the larger value $\hat{w}_k$ is, the more importance X_k is. Thus, for a pre-specified threshold d , we would consider

$$\hat{\mathcal{D}} = \{k: \hat{w}_k \text{ is among the top } d \text{ largest of all}\}, \quad (2.6)$$

as an estimator of $\mathcal{D}$.

The proposed procedure does not impose assumption on the model structure, and hence it is model free. Since the estimated marginal utility is based on the empirical cumulative distribution functions (CDF), it is robust to outlier or is invariant under strictly monotone increasing transformation on Y , X_k or Z .

We next study the theoretical properties of the proposed conditional feature screening procedure based on w_k s. We first show that the proposed conditional screening procedure enjoys the sure screening property under the condition below.

There exists positive constant $c > 0$ and $0 \leq \tau \leq 1/2$ such that $\min_{k \in \mathcal{A}} w_k \geq 2cn^{-\tau}$ (C1)

Condition (C1) assumes that the minimum signal cannot be too small, and it is a typical assumption in the literature of feature screening (see, e.g., Condition 3 in Fan and Lv (2008), Condition E in Fan and Song (2010), Condition T2 in Xu and Chen (2014), Condition 2 in Li, Zhong and Zhu (2012), Condition 2 in Cui, Li and Zhong (2014)).

Theorem 2.1. [Sure Screening Property] (a) Under Condition (C1), there exists a positive constants b such that

$$\left(\max_{1 \leq k \leq p} |\hat{w}_k - w_k| > cn^{-\tau} \right) \leq O(p \exp\{-bn^{1-2\tau} + \log(8n+2)\}) \quad (2.7)$$

(b) Define

$$\hat{\mathcal{D}}_* = \{k: \hat{w}_k \geq cn^{-\tau}, \text{ for } 1 \leq k \leq p\}, \quad (2.8)$$

Under Condition (C1), it follows that

$$(\mathcal{D} \subseteq \hat{\mathcal{D}}_*) \geq 1 - O(s_n \exp\{-bn^{1-2\tau} + \log(8n+2)\}) \quad (2.9)$$

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

The property in (2.9) is referred to as the sure screening property of feature screening procedure (Fan and Lv, 2008). Thus, we refer the proposed feature screening procedure as MC based sure conditional independent screening (MC-SCIS for short). We next show that MC-SCIS enjoys the ranking consistency property (Zhu, Li, Li and Zhu, 2011). To establish the ranking consistency property, we impose the following condition for MC-SCIS.

$$\liminf_{p \rightarrow \infty} \left\{ \min_{k \in \mathcal{A}} w_k - \min_{k \in I} w_k \right\} \geq c_1. \quad (C2)$$

This condition is imposed in Cui, Li and Zhong (2014) and is weaker than the partial orthogonality condition (Huang, Horowitz and Ma, 2008). Under this condition, we show that the MC-SCIS enjoys the ranking consistency property.

Theorem 2.2. [Ranking Consistency Property] Under Condition (C2), if $p = o(\exp(an))$ with $a < b(c_1/4c)^2$, c and c_1 given in Conditions (C1) and (C2), respectively, and b given in Theorem 2.1(a), it follows that

$$\liminf_{n \rightarrow \infty} (\min_{k \in \mathcal{A}} \hat{w}_k - \max_{k \in I} \hat{w}_k) > 0, a.s. \quad (2.10)$$

It should be noted that the assumption here is not the weakest. Theorem 2.2 implies that the MC-SCIS can rank important variables above unimportant one with a high probability.

Remark: The proposed conditional screening procedure can be directly applicable for the feature screening in the absence of Z variable. In this situation, the index definition of (2.2) can be modified by $MC(Y, X) = E[F_{\tilde{Y}, \tilde{X}} - F_{\tilde{Y}} F_{\tilde{X}}]^2$. Using this unconditional MC index, the MC-SCIS can be used for feature screening. Results in Theorems 2.1 and 2.2 are also valid under similar conditions in (C1) and (C2). Thus, we refer to this unconditional feature screening procedure as MC-SIS. We will compare the performance of MC-SIS with some commonly-used model free feature screening procedures in Section 3.1.

2.3 Some extensions

Here we consider some extensions of MC-SCIS. From the definition of MC in section 2.1, the MC-SCIS is directly applicable for situations in which the response and conditional variable are multivariate. Let $\mathbf{y} = \{Y_1, \dots, Y_l\}$ and $\mathbf{z} = \{Z_1, \dots, Z_m\}$. Define the multivariate comparison of $\mathbf{y}$ and $\mathbf{z}$ by $I(\mathbf{y}, \tilde{\mathbf{y}}) = I(Y_1, \tilde{Y}_1, \dots, Y_l, \tilde{Y}_l)$ and $I(\mathbf{z}, \tilde{\mathbf{z}}) = I(Z_1, \tilde{Z}_1, \dots, Z_m, \tilde{Z}_m)$, respectively. Then denote $F_{\tilde{\mathbf{z}}} = P(\mathbf{z} = \tilde{\mathbf{z}})$, $F_{\tilde{\mathbf{y}}, \tilde{\mathbf{z}}} = P(\mathbf{y} = \tilde{\mathbf{y}}, \mathbf{z} = \tilde{\mathbf{z}})$, $F_{\tilde{\mathbf{y}}|\tilde{\mathbf{z}}} = P(\mathbf{y} = \tilde{\mathbf{y}}|\mathbf{z} = \tilde{\mathbf{z}})$, $F_{\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{z}}} = P(\mathbf{y} = \tilde{\mathbf{y}}, X = \tilde{X}, \mathbf{z} = \tilde{\mathbf{z}})$, $F_{\tilde{\mathbf{y}}, \tilde{\mathbf{X}}|\tilde{\mathbf{z}}} = P(\mathbf{y} = \tilde{\mathbf{y}}, X = \tilde{X}|\mathbf{z} = \tilde{\mathbf{z}})$. Similarly, we define the notation $F_{\tilde{\mathbf{X}}|\tilde{\mathbf{z}}}$ and $F_{\tilde{\mathbf{X}}, \tilde{\mathbf{z}}}$. Here $\{\tilde{X}, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}\}$ is an independent copy of $\{X, \mathbf{y}, \mathbf{z}\}$. Then the corresponding multivariate index for conditional independence can be defined as

$$MC(\mathbf{y}, X|\mathbf{z}) = E[F_{\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{z}}} F_{\tilde{\mathbf{z}}} - F_{\tilde{\mathbf{y}}, \tilde{\mathbf{z}}} F_{\tilde{\mathbf{X}}, \tilde{\mathbf{z}}}]^2. \quad (2.11)$$

The definition in (2.11) is for any dimensional $\mathbf{y}$ and $\mathbf{z}$. In the practical implementation, both l and m should be small because the quantity involved in the definition of $MC(\mathbf{y}, X|\mathbf{z})$ cannot be estimated accurately without a large number of sample size.

The proposed MC-SCIS can be extended for situation when the response or the predictors are categorical. Here we only consider the case when the response variable Y is one dimension categorical since the other cases are similar. Let Y be a categorical response with classes $\{R_1, R_2, \dots, R_c\}$, and naturally define $F_{\tilde{Y}, \tilde{Z}} = P(Y = \tilde{Y}, Z = \tilde{Z})$, $F_{\tilde{Y}|\tilde{Z}} = P(Y = \tilde{Y}|Z = \tilde{Z})$, $F_{\tilde{Y}, \tilde{X}, \tilde{Z}} = P(Y = \tilde{Y}, X = \tilde{X}, Z = \tilde{Z})$, $F_{\tilde{Y}, \tilde{X}|\tilde{Z}} = P(Y = \tilde{Y}, X = \tilde{X}|Z = \tilde{Z})$. Thus, we can further define the corresponding $MC(X, Y|Z)$ and $\widehat{MC}(X, Y|Z)$.

The sure screening property and ranking consistency property for the extensions discussed above can be established along the same line as those for the univariate case, so we omit the details to save space.

3 Numerical Examples

In this section, we assess the finite sample performance of MC-SCIS by Monte Carlo simulations, and compare MC-SIS with other feature screening procedures under various models. We adopt the following criteria to evaluate the performance of feature screening procedures:

1. P_j : the proportion that an individual active variable X_j is selected for a given screening size d .
2. P_a : the proportion that all active variables are selected for a given screening size d .
3. MMS: The minimum model size to include all active variables.

In our numerical examples, we set d to be $[n/\log(n)]$, the integer part of $n/\log(n)$.

3.1 Comparison the performance of MC-SIS with other feature screening procedures

In the absence of Z -variable, the MC-SCIS becomes MC-SIS. Thus, we first conduct numerical comparison between existing methods and MC-SIS. The dimension p is 1000 and for each scenario, we conduct 500 simulations.

Example 1. This example is designed to compare the performance of MC-SIS with SIS (Fan and Lv, 2008), SIRS (Zhu, Li, Li and Zhu, 2011), DC-SIS (Li, Zhong and Zhu, 2012) when data are generated from a linear model:

$$Y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon. \quad (3.1)$$

Here we set $\boldsymbol{\beta} = (\beta_1, \dots, \beta_4, 0, \dots, 0)^T$ and $\beta_j = c_j(-1)^{U(a+|Z|)}$ for $j = 1, 2, 3$ and 4 , where $U \sim \text{Bernoulli}(0.4)$, $a = \log(n)/n$ and $Z \sim N(0, 0.5^2)$, $c = (c_1, \dots, c_4) = (1, 0.5, 1, 0.5)$. ε is independent with the predictors and is generated from the four kinds of distributions which stated below. $\mathbf{x} = (X_1, \dots, X_p)$ is generated from $N(0, \Sigma)$ with $\Sigma = (\sigma_{ij})$. Here we consider AR covariance matrix structures $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$ and 0.8 . To compare the robustness to heavy tail and outliers for SIS procedure, we consider four different distributions for ε : Case 1: $N(0,1)$, Case 2: $0.9N(0,1) + 0.1\text{Cauchy}(0,10)$, where $\text{Cauchy}(0,10)$ stands for the Cauchy distribution with location parameter being 0 and scatter parameter being 10. Case 3: $t(3)$, the t -distribution with 3 degrees of freedom, and Case 4: the Cauchy distribution with location parameter being 0 and scatter parameter being 1. For ease of presentation, we refer Case 2 to as “mixtures” in Tables in this section since it is a mixture of normal and Cauchy distributions.

Table 1 displays P_j , P_a and the average and standard deviation of MMS over 500 simulations with the sample size $n = 100$. The performance of SIS, SIRS, DC-SIS and MC-SIS are quite

similar when the error is normal, or t distribution. When the error follows a Cauchy distribution or in the presence with outliers (i.e. Case 2), MC-SIS and SIRS behave quite similar and outperform the other ones. Simulation results in this example clearly demonstrate that MC-SIS is more robust to different error distribution than other methods.

Example 2. In this example, we investigate the finite sample performance of MC-SIS when the response are generated from nonlinear models:

$$Y = 2\beta_1 X_1^2 + 2\beta_2 X_2^2 + 2\beta_3 X_3^2 + 2\beta_4 X_4^2 + \varepsilon \quad (\text{a1})$$

$$Y = 2\beta_1 X_1^2 + 2\beta_2 X_2 X_3 + 2\beta_3 I(X_4 > 0) + \varepsilon \quad (\text{a2})$$

$$Y = \exp\{\beta_1 \sin(X_1) + \beta_2 \exp(X_2) + \beta_3 I(X_3 > 0) + \log(|X_4|)\} + \varepsilon \quad (\text{a3})$$

$$Y = \exp\{\beta_1 X_1 + \beta_2 X_2\} + \exp\{\beta_3 X_3 + \beta_4 X_4\} + \varepsilon \quad (\text{a4})$$

The $\mathbf{x}$, β and the error distribution are set to the same as those in Example 1. Simulation results for $\rho = 0.5$ are similar to those for $\rho = 0.8$. Thus, we opt not to report results for $\rho = 0.5$ to save space. Tables 2 and 3 summarize the simulation results with the sample size $n = 200$. These two tables imply that the MC-SIS outperforms the other methods across all the cases. In particular, the MC-SIS is very robust to the heavy tailed error distribution and outliers.

Example 3. As illustrated in Section 2.3, the MC-SIS can be directly applied for multivariate response. It seems that only DC-SIS (Li, Zhong and Zhu, 2012) has been applied for multivariate response. This example is designed to compare the finite sample performance of MC-SIS with DC-SIS. We adopt the same multivariate model structure of Li, Zhong and Zhu (2012), but further consider multivariate Cauchy distribution to examine the robust property of MC-SIS, in addition to multinormal distribution. Specifically, the bivariate response $\mathbf{y} = (Y_1, Y_2)^T$ were generated from $\mathcal{N}((0, 0)^T, \Sigma_{\mathbf{y}|\mathbf{x}})$, or $\text{Cauchy}((0, 0)^T, \Sigma_{\mathbf{y}|\mathbf{x}})$ with the scatter matrix $\Sigma_{\mathbf{y}|\mathbf{x}} = (\sigma_{x,ij})_{2 \times 2}$, where $\sigma_{\mathbf{x},11} = \sigma_{\mathbf{x},22} = 1$ and $\sigma_{\mathbf{x},12} = \sigma_{\mathbf{x},21} = \sigma(\mathbf{x})$. We consider the following two models for $\sigma(\mathbf{x})$:

$$\sigma(\mathbf{x}) = \sin(\beta_1^T \mathbf{x}), \text{ where } \beta_1 = (0.8, 0.6, 0, \dots, 0)^T. \quad (\text{b1})$$

$$\sigma(\mathbf{x}) = \{\exp(\beta_2^T \mathbf{x}) - 1\} \prod \{\exp(\beta_2^T \mathbf{x}) + 1\}, \quad (\text{b2})$$

Where $\beta_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, \dots, 0)$ and U_j 's are independent and identically distributed with uniform distribution Uniform[0,1].

The summary of simulation results for multivariate response with $n = 200$ is depicted in Table 4, for which it can be seen that for multinormal response, both MC-SIS and DC-SIS perform very well, although the DC-SIS performs slightly better than the MC-SIS. However, for multivariate Cauchy response, the MC-SIS method performs much better than the DC-SIS method. In particular, the DC-SIS cannot detect the active predictors with a reasonable probability. Table 4 also indicates that the selected proportion of active predictors by MC-SIS varies very little for two different error distributions. This implies the MC-SIS method is more robust than the DC-SIS.

3.2 Performance of MC-SCIS

This section is designed to examine the finite sample performance of our proposed conditional feature screening procedure MC-SCIS. We take $p = 1000$ in Examples 5 and 6. Four error distributions for ε are set to be the same as those in Example 1. For each scenario, we conduct 500 simulations.

Example 4. This example is to mimic the setting of genome-wide association studies, in which the SNP variables are codes as binary variables, and environmental factor is also a binary variable. The environmental factor and some SNPs have interaction effects on the response. To this end, we generate the response from the following model, $Y = X_1 Z + X_2 Z + \varepsilon$, where $X_j, j = 1, \dots, p$, and Z are independent and identically distributed with $P(X_j = \pm 1) = 0.5$ and $P(Z = \pm 1) = 0.5$. It is easy to verify that Y is independent of both X_1 and X_2 , SIS procedures cannot detect the signal of X_1 and X_2 . While MC-SCIS can be effective to detect genetic effect conditioning on the environmental factor Z . In our simulation we set $n = 200$. Table 5 summarizes the simulation results. It can be seen from Table 5 that all the other feature screening procedures fail, but the MC-SCIS can identify both X_1 and X_2 perfectly. This is what we expect.

Example 5. This example is designated to illustrate the performance of the proposed method when the conditional variable Z is a random vector instead of a random variable. To this end, we set $P(Z_1 = \pm 1) = 0.5$, $P(Z_2 = \pm 1) = 0.5$ and let $Y = \mathbf{x}^T \beta \cdot Z_1 Z_2 + \varepsilon$, where β and ε are the same as those in Example 1, and $\mathbf{z} = (Z_1, Z_2)^T$ is independent with $\mathbf{x}$. In this example we conduct 500 simulations with $n = 100$. The simulation results are summarized in Table 6, from which it can be seen that the MC-SCIS outperforms other methods included in this simulation study. In particular, the other methods has very low P_a across all four cases, while the MC-SCIS has 100% P_a 's for error distributions with $N(0,1)$ and $t(3)$, and 99% P_a for Case 2, and 92% P_a for Cauchy error.

Example 6. In this example, the response was generated from a varying-coefficient model:

$$Y = \beta_1(U)X_1 + \beta_2(U)X_2 + \beta_3(U)X_3 + \beta_4(U)X_4 + \varepsilon, \quad (3.2)$$

where $\mathbf{x}$ and the four error distributions for ε are taken to be the same as those in Example 1, U is generated from the uniform distribution over $[0, 1]$. We set $\beta_1(u) = 2$, $\beta_2(u) = 3u$, $\beta_3(u) = (u + 1)^2$ and $\beta_4(u) = 4\sin(2\pi u)/\{2 - \sin(2\pi u)\}$. This model was used in the simulation study of Fan, Ma and Dai (2014). In this example, we set $n = 400$.

Given U , the varying-coefficient model is a linear regression model. Thus, the varying coefficient model is a natural extension of linear model. We compare the performance of MC-SIS and MC-SCIS with $Z = U$ to get insight into the price without using conditional feature screening procedure. Simulation results are summarized in Table 7, from which we can see that (a) the MC-SIS performs better than other feature screening procedures without using information of variable U , and (b) the MC-SCIS further improves the performance of MC-SIS significantly.

3.3 Real data examples

In this section, we illustrate the proposed methodology by two real data examples.

Example 6. This example is used to illustrate the MC-SIS by an empirical of a rat eye expression dataset. The dataset was previously analyzed by Scheetz et al. (2006) and Huang, Ma and Zhang (2008). In this dataset, 120 twelve-week-old male rats were selected for tissue harvesting from the eyes and microarray analysis. There involves 31,042 probe sets on these rats. Many of these probes have very small variation. In our empirical analysis, we use only 3000 genes with largest variances. Researchers are interested in finding the genes that associate with gene TRIM32 (probe 1389163_at), which is considered to have a association with the disease of Bardet-Biedl syndrome (Chiang et al., 2006). Thus, our analysis is based on the predictors of 3000 genes and the response variable of gene TRIM32.

We first apply screening procedures to select the most important variables with a model size $[2n/\log(n)]$, then we use the adaptive lasso (ALasso) and SCAD procedure with the BIC tuning parameter selection criterion to select a sparse linear regression model. To assess the performance of different methods, we randomly partition the dataset into a training dataset with 100 observations and a testing dataset with 20 observations. We fit a linear model with the selected predictors and calculate the predicted mean squared error (MSE) and the correlation between the predicted response with the observed response. Table 8 displays the results of the SIS, SIRS, DC-SIS, MC-SIS with the ALasso and SCAD methods. In Table 8, we use SIS-ALasso and SIS-SCAD for the two-stage procedure using SIS for screening and ALasso and SCAD for variable selection, respectively. The same notation applies for SIRS, DS-SIS and MC-SIS. For sake of comparison, Table 8 also lists the result of Huang, Ma and Zhang (2008) in the HMZ row. Table 8 clearly shows that the model size of the two stage procedure using SIS, SIRS, DC-SIS and MC-SIS is much smaller than the HMZ's method. Moreover, the MC-SIS has smaller MSE than the other SIS procedures, and the MC-SIS has larger correlation between the predicted value and the observed value than the other SIS procedures. The MSE value of MC-SIS is quite close to the HMZ's method and the correlation of the HMZ's method is larger, but the HMZ's method yields a much larger model size. In this example, our empirical analysis using linear model after screening is merely for the purpose of illustration, partly also because SIS was proposed based on linear models.

Example 7. In this example, we illustrate the proposed MC-SCIS procedure by an empirical analysis of a subset of Framingham Heart Study data. This data subset consists of 349,985 nonrare SNPs, but only 976 subjects are available after omitting one subject whose majority part being unobserved. It is of interest to detect the important SNPs that are associated with BMI. The empirical analysis in Liu, Li and Wu (2014) implies that the effect of SNPs on BMI may change with age. It is appropriate for us to apply the MC-SCIS with age as the conditional variable. In this example, we take the BMI as the response Y , the SNPs as the predictor vector, and age to be the conditional variable. Furthermore, each SNP is further coded by the dominant effect and the additive effect. Thus, there are 699,970 x -variables for screening.

To compare our method with CC-SIS (Liu, Li and Wu, 2014), we follow Liu, Li and Wu (2014) and choose $\lceil n^{4/5} / \log n^{4/5} \rceil = 45$ SNPs at the screening stage. To examine the genetic effect on the BMI at different age period, we partition the data into five subsets according to age and we fit a linear model on each age interval. The R-square and MSE of MC-SCIS and CC-SIS are presented in Table 9, from which it can be seen that the MC-SCIS achieved the larger R square and smaller MSE than CC-SIS over most age periods. We further examine the estimated coefficients for the five different age periods, we find that there are at least 20 SNPs having age-dependent effect on BMI. The estimated coefficients of these 20 SNPs along with their confidence interval are plotted in Figure 1, which clearly shows that interaction effects between SNPs and age on the BMI.

4 Summary

In this paper, we propose a model-free conditional feature screening method based on MC-SCIS, and establish the sure screening property and ranking consistency property under certain conditions. The new conditional feature screening procedure is robust to heavy-tailed error distributions or outliers. Similar to SIS, one further considers iterated procedures for MC-SCIS (MC-ISCIS). Based on limited experience from our simulation, the iterated procedure might improve the performance of MC-SCIS. Systematical study of such iterative procedures is a good research topic for future research.

Acknowledgments

Luheng Wang's research was supported by NSF grant P50 DA039838 and the program of China Scholarships Council (CSC No. 201506040130) as a visiting student at the Pennsylvania State University. Jingyuan Liu's research was supported by National Natural Science Foundation of China (NNSFC) grant 11401497 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. Yong Li's research was supported by the National Key Basic Research Development Program of China (NKBRDPC) grant 2010CB950703 and the Fundamental Research Funds for the Central Universities (FRFCU). Runze Li's research was supported by NIDA, NIH grants P50 DA036107, P50 DA039838 and NSF grant DMS 1512422. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDA, NIH, NSF, NKBRDPC, FRFCU, CSC or NNSFC.

5 Appendix

To prove Theorem 1, we need the following lemmas.

Lemma 5.1. [Hoeffding's Inequality] (Hoeffding, 1963) Let $X_1, X_2, \dots, X_n$ be independent random variables with $\mathbb{P}(X_i \in [a_i, b_i]) = 1, 1 \leq i \leq n$. Then for any $t > 0$,

$$(|\bar{X} - E(\bar{X})| \geq t) \leq 2 \exp \left\{ -\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right\} \quad (5.1)$$

Lemma 5.2. [Bernstein's Inequality] (Van Der Vaart and Wellner, 1996, Lemma 2.2.9) Let $X_1, X_2, \dots, X_n$ be independent random variables with bounded $[-M, M]$ and zero means, then for any $t > 0$,

$$(|X_1 + \dots + X_n| > t) \leq 2 \exp \left\{ -\frac{t^2}{2(v + Mt/3)} \right\} \quad (5.2)$$

for $v = \text{Var}(X_1 + \dots + X_n)$.

Denote

$$\begin{aligned} f_z^{(i)} &= I\{Z_i \leq z\} \\ f_{z,x}^{(i)} &= I\{Z_i \leq z, X_{k,i} \leq x\} \\ f_{y,z}^{(i)} &= I\{Y_i \leq y, Z_i \leq z\} \\ f_{y,x,z}^{(i)} &= I\{Y_i \leq y, X_{k,i} \leq x, Z_i \leq z\} \end{aligned}$$

here $x \in \mathbb{R}_X, y \in \mathbb{R}_Y, z \in \mathbb{R}_Z$, where $\mathbb{R}_X, \mathbb{R}_Y, \mathbb{R}_Z$ is the support of X, Y, Z , respectively.

Lemma 5.3. For any $\varepsilon \in (0, 1)$, and $x \in \mathbb{R}_X, y \in \mathbb{R}_Y, z \in \mathbb{R}_Z$, the following inequalities hold

$$\begin{aligned} \left\{ \left| \hat{F}_Z - F_Z \right| \geq \varepsilon \right\} &\geq 2 \exp \left\{ -\frac{n\varepsilon^2}{2(p_1 + \varepsilon/3)} \right\} \\ \left\{ \left| \hat{F}_{z,x} - F_{z,x} \right| \geq \varepsilon \right\} &\leq 2 \exp \left\{ -\frac{n\varepsilon^2}{2(p_2 + \varepsilon/3)} \right\} \\ \left\{ \left| \hat{F}_{y,z} - F_{y,z} \right| \geq \varepsilon \right\} &\leq 2 \exp \left\{ -\frac{n\varepsilon^2}{2(p_3 + \varepsilon/3)} \right\} \\ \left\{ \left| \hat{F}_{y,x,z} - F_{y,x,z} \right| \geq \varepsilon \right\} &\leq 2 \exp \left\{ -\frac{n\varepsilon^2}{2(p_4 + \varepsilon/3)} \right\} \end{aligned} \quad (5.3)$$

where $p_1 = E(f_z^{(i)})$, $p_2 = E(f_{z,x}^{(i)})$, $p_3 = E(f_{y,z}^{(i)})$ and $p_4 = E(f_{y,x,z}^{(i)})$, respectively.

Proof. The proof of the first four inequalities are similar, we only deal with the first one.

Since $f_z^{(i)} = I\{Z_i \leq z\}$, then $f_z^{(i)} \sim \text{Bernoulli}(p_1)$, where $p_1 = E(f_z^{(i)}) = F_z$. Then

$\text{Var} \left(\sum_{i=1}^n (f_z^{(i)} - F_z) \right) = np_1(1 - p_1) \leq np_1$ and $|f_z^{(i)} - F_z| \leq 1$. Then by the Bernstein's inequality,

$$\begin{aligned} \left\{ \left| F_z - \hat{F}_z \right| \geq \varepsilon \right\} &= \left\{ \left| \sum_{i=1}^n (f_z^{(i)} - F_z) \right| \geq n\varepsilon \right\} \\ &\leq 2\exp \left\{ -\frac{n^2\varepsilon^2}{2(np_1 + n\varepsilon/3)} \right\} = 2\exp \left\{ -\frac{n\varepsilon^2}{2(p_1 + \varepsilon/3)} \right\} \end{aligned}$$

Lemma 5.4. Under Condition (C1), for any $\varepsilon \in (0, 1/2)$, we have

$$(|\hat{w}_k - w_k| \geq \varepsilon) \leq (8n+2)\exp\{-b_1 n\varepsilon^2\} \quad (5.4)$$

$$\text{where } b_1 = \min \left\{ \frac{2}{81}, \frac{(1/9^2)}{2(1+\varepsilon/9)} \right\}.$$

Proof. Denote $\hat{E}f(X) = 1/n \sum_{i=1}^n f(X_i)$, which is the sample mean of $f(X)$. Thus

$$\begin{aligned} |\hat{w}_k - w_k| &= |\hat{E}[(\hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} \hat{F}_{\hat{Z}} - \hat{F}_{\hat{Y}, \hat{Z}} \hat{F}_{\hat{X}_k, \hat{Z}})] - E[(F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}} - F_{\hat{Y}, \hat{Z}} F_{\hat{X}_k, \hat{Z}})^2]| \leq |\hat{E}[(\hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} \hat{F}_{\hat{Z}} - \hat{F}_{\hat{Y}, \hat{Z}} \hat{F}_{\hat{X}_k, \hat{Z}})^2] - (F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}} - \\ &\quad + |(\hat{E} - E)[(F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}} - F_{\hat{Y}, \hat{Z}} F_{\hat{X}_k, \hat{Z}})^2]| = I_1 + I_2 \end{aligned} \quad (5.5)$$

We first deal with I_2 , since $(F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}} - F_{\hat{Y}, \hat{Z}} F_{\hat{X}_k, \hat{Z}})^2 \leq 1$, we can apply Hoeffding's inequality

$$\{|I_2| \geq \varepsilon\} \leq 2\exp\{-2n\varepsilon^2\} \quad (5.6)$$

Next we will deal with I_1 ,

$$\begin{aligned} |I_1| &\leq 2 \hat{E} \left[\left| (\hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} \hat{F}_{\hat{Z}} - F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}}) - (\hat{F}_{\hat{Y}, \hat{Z}} \hat{F}_{\hat{X}_k, \hat{Z}} - F_{\hat{Y}, \hat{Z}} F_{\hat{X}_k, \hat{Z}}) \right| \right] \\ &\leq 2 \hat{E} \left[\left| (\hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} - F_{\hat{Y}, \hat{X}_k, \hat{Z}}) \hat{F}_{\hat{Z}} + (\hat{F}_{\hat{Z}} - F_{\hat{Z}}) F_{\hat{Y}, \hat{X}_k, \hat{Z}} + (\hat{F}_{\hat{Y}, \hat{Z}} - F_{\hat{Y}, \hat{Z}}) \hat{F}_{\hat{X}_k, \hat{Z}} + (\hat{F}_{\hat{X}_k, \hat{Z}} - F_{\hat{X}_k, \hat{Z}}) F_{\hat{Y}, \hat{Z}} \right| \right] \\ &\leq 2 \hat{E} \left[\left| \hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} - F_{\hat{Y}, \hat{X}_k, \hat{Z}} \right| + \left| \hat{F}_{\hat{Z}} - F_{\hat{Z}} \right| + \left| \hat{F}_{\hat{Y}, \hat{Z}} - F_{\hat{Y}, \hat{Z}} \right| + \left| \hat{F}_{\hat{X}_k, \hat{Z}} - F_{\hat{X}_k, \hat{Z}} \right| \right] \\ &\quad = 2(L_1 + L_2 + L_3 + L_4) \end{aligned} \quad (5.7)$$

where the first inequality holds by $(\hat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} \hat{F}_{\hat{Z}} - F_{\hat{Y}, \hat{X}_k, \hat{Z}} F_{\hat{Z}}) + (\hat{F}_{\hat{Y}, \hat{Z}} \hat{F}_{\hat{X}_k, \hat{Z}} - \hat{F}_{\hat{Y}, \hat{Z}} F_{\hat{X}_k, \hat{Z}}) \leq 2$.

Here we only consider the inequality of L_1 here since the deduction of the other part is quite similar.

$$\begin{aligned}
 (L_1 \geq \varepsilon) &= \left(\widehat{E} \left[\left| \widehat{F}_{\hat{Y}, \hat{X}_k, \hat{Z}} - F_{\hat{Y}, \hat{X}_k, \hat{Z}} \right| \right] \geq \varepsilon \right) \\
 &= \left(\frac{1}{n} \sum_{i=1}^n \left| \widehat{F}_{Y_i, X_{k,i}, Z_i} - F_{Y_i, X_{k,i}, Z_i} \right| \geq \varepsilon \right) \\
 &\leq \sum_{i=1}^n \left(\left| \widehat{F}_{Y_i, X_{k,i}, Z_i} - F_{Y_i, X_{k,i}, Z_i} \right| \geq \varepsilon \right) \\
 &\leq 2n \exp \left\{ -\frac{n\varepsilon^2}{2(\max\{p_{4,i}\} + \varepsilon/3)} \right\} \\
 &\leq 2n \exp \left\{ -\frac{n\varepsilon^2}{2(1+\varepsilon/3)} \right\}
 \end{aligned} \tag{5.8}$$

where $p_{4,i} = F_{Y_i, X_{k,i}, Z_i}$. Thus, by inequalities (5.5) to (5.8), we can obtain that

$$\begin{aligned}
 P(|\hat{w}_k - w_k| \geq \varepsilon) &\leq (|I_1| \geq \frac{\varepsilon}{9} + \sum_{i=1}^4 (|L_i| \geq \frac{\varepsilon}{9})) \\
 &\leq 2 \exp \{-2n(\varepsilon/9)^2\} + 2n \sum_{i=1}^4 \exp \left\{ -\frac{n(\varepsilon/9)^2}{2(1+\varepsilon/9)} \right\} \\
 &\leq (8n+2) \exp \{-b_1 n \varepsilon^2\}
 \end{aligned} \tag{5.9}$$

where $b_1 = \min \left\{ \frac{2}{81}, \frac{(1/9^2)}{2(1+\varepsilon/9)} \right\}$. Thus we complete the proof of Lemma 5.4.

Proof of Theorem 2.1 By Lemma 5.4, we can get that

$$(|\hat{w}_k - w_k| > cn^{-\tau}) \leq (8n+2) \exp \{-bn^{1-2\tau}\} \tag{5.10}$$

where $b = b_1 c^2$.

For Theorem 2.1(a), by Lemma 5.4, we have

$$\begin{aligned}
 \left(\max_{1 \leq k \leq p} |\hat{w}_k - w_k| > cn^{-\tau} \right) &\leq O(p)(8n+2) \exp \{-bn^{1-2\tau}\} \\
 &\leq O(p) \exp \{-bn^{1-2\tau} + \log(8n+2)\}
 \end{aligned} \tag{5.11}$$

For Theorem 2.1(b), by Condition (C2), if $w_k \in \mathcal{D}$ but $\hat{w}_k \notin \hat{\mathcal{D}}^*$, then $|\hat{w}_k - w_k| > 2cn^{-\tau} - cn^{-\tau} = cn^{-\tau}$, which implies that $\{\mathcal{D} \subseteq \hat{\mathcal{D}}\} \subseteq \{|\hat{w}_k - w_k| > cn^{-\tau}, \text{ for some } k \in \mathcal{D}\}$, and hence $\mathcal{D}_n = \{\max_{k \in \mathcal{D}} |\hat{w}_k - w_k| > cn^{-\tau}\} \subseteq \{\mathcal{D} \subseteq \hat{\mathcal{D}}^*\}$. Therefore,

$$\begin{aligned}
(\mathcal{D} \subseteq \hat{\mathcal{D}}_*) &\geq \{\mathcal{D}_n\} = 1 - \{\mathcal{D}_n^c\} \\
&\geq 1 - \left\{ \min_{k \in \mathcal{D}} |\hat{w}_k - w_k| > cn^{-\tau} \right\} \\
&\geq 1 - s_n \{ |\hat{w}_k - w_k| > cn^{-\tau} \} \\
&\geq 1 - O(s_n \exp\{-bn^{1-2\tau} + \log(8n+2)\}) \quad (5.12)
\end{aligned}$$

Proof of Theorem 2.2

$$\begin{aligned}
&\left\{ \left(\min_{k \in \mathcal{A}} \hat{w}_k - \max_{k \in \mathcal{J}} \hat{w}_k \right) < c_1/2 \right\} \\
&\leq \left\{ \left(\min_{k \in \mathcal{A}} \hat{w}_k - \max_{k \in \mathcal{J}} \hat{w}_k \right) - \left(\min_{k \in \mathcal{A}} w_k - \max_{k \in \mathcal{J}} w_k \right) < -c_1/2 \right\} \\
&\leq \left\{ \left| \left(\min_{k \in \mathcal{A}} \hat{w}_k - \max_{k \in \mathcal{J}} \hat{w}_k \right) - \left(\min_{k \in \mathcal{A}} w_k - \max_{k \in \mathcal{J}} w_k \right) \right| > c_1/2 \right\} \\
&\leq \left\{ 2 \min_{1 \leq k \leq p} |\hat{w}_k - w_k| > c_1/2 \right\} \\
&\leq O(n) p \exp\{-b_2 n\} \quad (5.13)
\end{aligned}$$

where $b_2 = b(c_1/4c)^2$ and the second inequality holds because of condition C2. Then if n is large enough, we can get that $-b_2 n + \log(p) + O(\log n) < -b_3 n$ because $\log p = o(an)$ with $a < b_2$, where b_3 is a constant and $0 < b_3 < b_2$. Thus, we have for some n_0 ,

$$\sum_{n=n_0}^{+\infty} n p \exp\{-b_2 n\} \leq \sum_{n=n_0}^{+\infty} \exp\{-b_3 n\} < +\infty. \text{ Then by the Borel Contelli Lemma, we}$$

can obtain that $\liminf_{n \rightarrow \infty} (\min_{k \in \mathcal{A}} \hat{w}_k - \max_{k \in \mathcal{J}} \hat{w}_k) > 0, a.s.$

References

- Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*. 2007; 35(6):2313–2351.
- Chiang AP, Beck JS, Yen HJ, Tayeh MK, Scheetz TE, Swiderski RE, Nishimura DY, Braun TA, Kim KYA, Huang J, et al. Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet-biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*. 2006; 103(16):6287–6292.
- Cui H, Li R, Zhong W. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*. 2014; 110(510):630–641.
- Donoho DL. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*. 2000:1–32.
- Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*. 2011; 106(494):544–557. [PubMed: 22279246]
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*. 2001; 96(456):1348–1360.
- Fan J, Li R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*. 2006
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70(5):849–911. [PubMed: 19603084]

- Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*. 2014; 109(507):1270–1284. [PubMed: 25309009]
- Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*. 2009; 10:2013–2038. [PubMed: 21603590]
- Fan J, abd Song R. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*. 2010; 38(6):3567–3604.
- Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*. 2009; 18(3):533–550.
- Hoeffding W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*. 1963; 58(301):13–30.
- Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*. 2008; 36(2):587–613.
- Huang J, Ma S, Zhang CH. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*. 2008; 18(4):1603–1618.
- Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. *The Annals of Statistics*. 2012; 40(3):1846–1877.
- Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *Journal of the American Statistical Association*. 2012; 107(499):1129–1139. [PubMed: 25249709]
- Liu J, Li R, Wu R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*. 2014; 109(505):266–274. [PubMed: 24678135]
- Liu J, Zhong W, Li R. A selective overview of feature screening for ultrahigh dimensional data. *Science China: Mathematics*. 2015; 58(10):1–22.
- Mai Q, Zou H. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*. 2012; 100(1):229–234.
- Scheetz TE, Kim KYA, Swiderski RE, Philp AR, Braun TA, Knudtson KL, Dorrance AM, DiBona GF, Huang J, Casavant TL, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*. 2006; 103(39):14429–14434.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996; 58(1):267–288.
- Van Der Vaart, AW., Wellner, JA. *Weak Convergence and Empirical Processes*. Springer; 1996.
- Xu C, Chen J. The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*. 2014; 109(507):1257–1269. [PubMed: 25382886]
- Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*. 2011; 106(496):1464–1475. [PubMed: 22754050]
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*. 2006; 101(476):1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67(2):301–320.

Figure 1.
Estimated coefficients and their confidence interval of 20 SNPs over five different age period. The A and D represents the additive effect and dominant effect of the selected SNP respectively.

Table 1

Simulation Results for Example 1: Linear Model

Method	$\sigma_{ij} = 0.5^{ i-j }$					$\sigma_{ij} = 0.8^{ i-j }$						
	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS
Case 1: $e \sim N(0,1)$												
SIS	1.00	1.00	1.00	1.00	1.00	4.6(8.0)	1.00	1.00	1.00	1.00	1.00	4.0(0.1)
SIRS	1.00	1.00	1.00	1.00	1.00	5.3(16.7)	1.00	1.00	1.00	1.00	1.00	4.0(0.1)
DC-SIS	1.00	1.00	1.00	1.00	1.00	5.1(12.5)	1.00	1.00	1.00	1.00	1.00	4.0(0.1)
MC-SIS	1.00	1.00	1.00	1.00	1.00	7.2(35.1)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
Case 2: mixtures												
SIS	0.23	0.27	0.28	0.18	0.08	534.6(335.6)	0.40	0.40	0.41	0.34	0.27	399.5(358.2)
SIRS	0.99	1.00	1.00	0.93	0.93	9.1(18.9)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
DC-SIS	0.81	0.83	0.86	0.64	0.59	78.9(138.7)	0.91	0.92	0.93	0.91	0.89	30.3(98.9)
MC-SIS	0.99	1.00	1.00	0.92	0.91	13.5(39.3)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
Case 3: $e \sim t(3)$												
SIS	0.99	1.00	0.99	0.94	0.93	16.3(76.7)	1.00	1.00	1.00	1.00	1.00	4.8(16.5)
SIRS	1.00	1.00	1.00	0.96	0.96	7.6(18.2)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
DC-SIS	1.00	1.00	1.00	0.96	0.95	8.8(25.2)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
MC-SIS	1.00	1.00	1.00	0.94	0.93	10.1(28.8)	1.00	1.00	1.00	1.00	1.00	4.1(0.3)
Case4: $e \sim \text{Cauchy}(0,1)$												
SIS	0.23	0.25	0.26	0.14	0.05	561.6(318.2)	0.40	0.43	0.44	0.39	0.27	375.4(360.4)
SIRS	0.96	0.98	0.99	0.83	0.79	30.7(81.3)	1.00	1.00	1.00	1.00	1.00	4.3(1.7)
DC-SIS	0.70	0.75	0.80	0.50	0.41	128.3(190.2)	0.91	0.93	0.93	0.89	0.87	32.0(95.1)
MC-SIS	0.94	0.97	0.99	0.81	0.76	37.1(103.3)	1.00	1.00	1.00	1.00	1.00	4.3(1.8)

Table 2
Simulation Results for Example 2: Nonlinear Models (a1) and (a2)

Method	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS
Results for Model (a1)												
Case 1: $e \sim N(0,1)$												
SIS	0.26	0.31	0.35	0.30	0.15	466.5(357.9)	0.18	0.21	0.21	0.19	0.09	548.1(338.8)
SIRS	0.08	0.10	0.08	0.11	0.03	683.5(291.9)	0.08	0.10	0.09	0.07	0.02	678.3(296.2)
DC-SIS	1.00	1.00	1.00	1.00	1.00	4.0(0.2)	0.98	0.99	0.99	0.99	0.98	7.0(26.3)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.0(0.3)	1.00	1.00	1.00	1.00	1.00	4.3(2.4)
Case 3: $e \sim t(3)$												
SIS	0.26	0.31	0.34	0.29	0.15	468.2(358.7)	0.15	0.18	0.19	0.18	0.08	562.1(328.8)
SIRS	0.07	0.10	0.08	0.11	0.03	681.8(289.5)	0.08	0.08	0.08	0.09	0.02	683.8(284.4)
DC-SIS	1.00	1.00	1.00	1.00	1.00	4.0(0.2)	0.99	0.99	1.00	0.99	0.99	7.2(33.9)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.1(0.4)	1.00	1.00	1.00	1.00	1.00	4.2(1.3)
Case 2: mixtures												
Case 4: $e \sim \text{Cauchy}(0,1)$												
Results for Model (a2)												
Case 1: $e \sim N(0,1)$												
SIS	0.42	0.45	0.49	0.58	0.32	370.6(361.1)	0.16	0.17	0.23	0.27	0.09	563.3(333.0)
SIRS	0.20	0.34	0.54	0.91	0.16	422.7(341.0)	0.16	0.28	0.48	0.87	0.12	439.6(324.8)
DC-SIS	1.00	1.00	1.00	1.00	1.00	4.1(1.2)	0.98	0.98	0.98	0.95	0.95	14.4(55.1)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.1(0.4)	1.00	1.00	1.00	1.00	1.00	4.3(1.9)
Case 3: $e \sim t(3)$												
SIS	0.41	0.44	0.48	0.58	0.30	371.5(359.1)	0.19	0.21	0.26	0.30	0.12	533.7(342.1)
SIRS	0.19	0.32	0.53	0.89	0.16	427.7(335.3)	0.16	0.28	0.46	0.79	0.12	447.4(340.4)
DC-SIS	1.00	1.00	1.00	1.00	1.00	4.2(1.3)	0.97	0.97	0.97	0.96	0.95	15.9(59.4)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.2(0.9)	1.00	1.00	1.00	1.00	1.00	4.9(3.2)
Case 2: mixtures												
Case 4: $e \sim \text{Cauchy}(0,1)$												

Table 3
Simulation Results for Example 2: Nonlinear Models (a3) and (a4)

Method	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS
Results for Model (a3)												
Case 1: $\epsilon \sim N(0,1)$												
SIS	0.77	1.00	0.80	0.65	0.56	90.6(148.4)	0.52	0.75	0.53	0.39	0.23	259.8(291.9)
SIRS	0.99	1.00	0.99	0.98	0.98	8.2(32.1)	0.99	1.00	1.00	0.97	0.97	7.6(21.5)
DC-SIS	0.77	1.00	0.80	0.67	0.58	86.0(141.1)	0.75	0.96	0.75	0.61	0.48	110.1(156.9)
MC-SIS	1.00	1.00	1.00	1.00	1.00	5.0(15.9)	0.99	1.00	1.00	1.00	0.99	4.6(4.5)
Case 3: $\epsilon \sim t(3)$												
SIS	0.77	1.00	0.79	0.64	0.56	91.5(148.7)	0.44	0.69	0.48	0.34	0.21	300.4(313.0)
SIRS	1.00	1.00	0.99	0.98	0.98	8.4(31.8)	0.98	0.99	0.97	0.95	0.93	13.6(41.6)
DC-SIS	0.77	1.00	0.80	0.67	0.58	86.0(141.1)	0.69	0.91	0.72	0.58	0.45	112.4(154.7)
MC-SIS	1.00	1.00	1.00	0.99	0.99	5.5(11.8)	0.97	0.98	0.97	0.97	0.95	12.0(34.9)
Results for Model (a4)												
Case 1: $\epsilon \sim N(0,1)$												
SIS	0.90	0.96	0.95	0.90	0.78	42.5(93.8)	0.89	0.96	0.97	0.87	0.75	48.7(106.3)
SIRS	1.00	1.00	1.00	1.00	1.00	4.0(0.1)	1.00	1.00	1.00	1.00	1.00	4.0(0.1)
DC-SIS	0.96	0.99	0.99	0.96	0.91	19.5(50.3)	0.95	0.99	0.98	0.95	0.90	21.4(66.2)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.0(0.1)	1.00	1.00	1.00	1.00	1.00	4.0(0.1)
Case 3: $\epsilon \sim t(3)$												
SIS	0.87	0.96	0.96	0.90	0.75	48.6(107.7)	0.90	0.97	0.95	0.88	0.76	44.3(90.8)
SIRS	1.00	1.00	1.00	1.00	1.00	4.0(0.1)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
DC-SIS	0.95	0.99	0.99	0.95	0.90	20.2(52.8)	0.94	0.98	0.98	0.94	0.88	21.0(54.9)
MC-SIS	1.00	1.00	1.00	1.00	1.00	4.0(0.1)	1.00	1.00	1.00	1.00	1.00	4.0(0.2)
Case 4: $\epsilon \sim \text{Cauchy}(0,1)$												

Table 4

Simulation results of Example 3: Multivariate Response

$\sigma_{ij} = 0.5^{ i-j }$										$\sigma_{ij} = 0.8^{ i-j }$				
Method	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS		
Model (b1)														
$y \sim N((0,0)^T, \Sigma_{y X})$														
DC-SIS	0.99	0.86				0.85	19.2(24.7)	1.00	0.99		0.99	6.8(7.2)		
MC-SIS	0.86	0.72				0.64	55.3(89.1)	0.93	0.90		0.86	19.1(35.0)		
$y \sim Cauchy((0,0)^T, \Sigma_{y X})$														
DC-SIS	0.14	0.12				0.03	342.0(220.0)	0.18	0.14		0.09	278.2(211.3)		
MC-SIS	0.84	0.71				0.64	67.5(124.2)	0.91	0.89		0.83	25.4(56.0)		
Model (b2)														
$y \sim N((0,0)^T, \Sigma_{y X})$														
DC-SIS	0.87	0.99	0.99	0.85	0.71	35.9(45.8)	1.00	1.00	1.00	1.00	1.00	4.9(2.6)		
MC-SIS	0.69	0.91	0.90	0.73	0.47	96.4(132.8)	0.98	1.00	1.00	0.98	0.96	10.5(26.2)		
$y \sim Cauchy((0,0)^T, \Sigma_{y X})$														
DC-SIS	0.10	0.21	0.20	0.09	0.00	432.8(213.5)	0.29	0.37	0.40	0.25	0.12	232.9(195.0)		
MC-SIS	0.75	0.91	0.91	0.76	0.50	94.9(141.6)	0.97	1.00	1.00	0.98	0.96	10.2(19.1)		

Table 5

Simulation Results for Example 4

Method	P_1	P_2	P_a	MMS	P_1	P_2	P_a	MMS
Case 1: $e \sim N(0,1)$								
SIS	0.03	0.04	0.01	627.9(270.3)	0.04	0.03	0.00	666.8(240.0)
SIRS	0.04	0.04	0.00	631.3(258.8)	0.04	0.03	0.00	664.8(234.9)
DC-SIS	0.04	0.04	0.00	625.9(264.2)	0.04	0.04	0.00	654.7(243.0)
MC-SIS	0.03	0.04	0.00	627.6(259.9)	0.04	0.02	0.00	658.9(235.3)
MC-SCIS	1.00	1.00	1.00	2.0(0.0)	1.00	1.00	1.00	2.0(0.0)
Case 3: $e \sim t(3)$								
SIS	0.03	0.03	0.00	654.8(250.8)	0.04	0.04	0.00	670.9(240.1)
SIRS	0.04	0.04	0.00	673.9(241.4)	0.03	0.04	0.00	671.8(235.0)
DC-SIS	0.04	0.03	0.00	671.6(241.0)	0.03	0.04	0.00	665.3(226.8)
MC-SIS	0.03	0.03	0.00	670.4(241.9)	0.04	0.04	0.00	672.8(232.8)
MC-SCIS	1.00	1.00	1.00	2.0(0.0)	1.00	1.00	1.00	2.3(2.1)
Case 4: $e \sim \text{Cauchy}(0,1)$								

Table 6

Simulation Results for Example 5

Method	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS
Case 1: $e \sim N(0,1)$												
SIS	0.17	0.18	0.18	0.15	0.07	559.0(330.3)	0.04	0.04	0.04	0.03	0.00	705.0(253.4)
SIRS	0.10	0.11	0.10	0.08	0.02	611.0(301.1)	0.08	0.10	0.10	0.08	0.02	655.7(286.9)
DC-SIS	0.46	0.67	0.67	0.34	0.20	144.0(137.6)	0.15	0.23	0.23	0.10	0.02	386.4(241.9)
MC-SIS	0.22	0.32	0.34	0.14	0.06	298.6(197.5)	0.12	0.20	0.22	0.14	0.02	393.9(228.3)
MC-CSIS	1.00	1.00	1.00	1.00	1.00	4.2(1.0)	1.00	1.00	1.00	0.99	0.99	5.5(7.4)
Case 3: $e \sim t(3)$												
SIS	0.16	0.14	0.16	0.14	0.05	600.8(327.1)	0.05	0.05	0.04	0.04	0.00	711.7(261.2)
SIRS	0.09	0.10	0.11	0.08	0.02	625.5(290.3)	0.05	0.06	0.07	0.05	0.00	681.7(272.3)
DC-SIS	0.35	0.46	0.48	0.24	0.10	214.5(175.1)	0.10	0.09	0.10	0.08	0.01	534.1(263.7)
MC-SIS	0.18	0.19	0.28	0.14	0.05	369.8(220.5)	0.11	0.11	0.11	0.08	0.02	529.9(251.1)
MC-CSIS	1.00	1.00	1.00	1.00	1.00	4.5(2.5)	0.98	0.99	1.00	0.94	0.92	17.0(52.1)
Case 4: $e \sim \text{Cauchy}(0,1)$												

Table 7

Simulation Results for Example 6: Varying Coefficient Model

Method	P_1	P_2	P_3	P_4	P_a	MMS	P_1	P_2	P_3	P_4	P_a	MMS
	Case 1: $\varepsilon \sim N(0,1)$						Case 2: mixtures					
SIS	1.00	1.00	0.31	0.74	0.23	347.5(293.8)	0.46	0.38	0.10	0.13	0.01	696.6(237.3)
SIRS	1.00	1.00	0.14	0.84	0.13	456.9(309.6)	1.00	1.00	0.10	0.76	0.06	495.9(298.3)
DC-SIS	1.00	1.00	1.00	0.91	0.91	25.9(61.8)	0.98	0.96	0.97	0.61	0.60	114.6(165.5)
MC-SIS	1.00	1.00	1.00	0.93	0.93	25.0(69.4)	1.00	1.00	1.00	0.85	0.85	41.5(102.6)
MC-SCIS	1.00	1.00	1.00	1.00	1.00	4.4(2.4)	1.00	1.00	1.00	1.00	0.99	5.7(7.5)
	Case 3: $\varepsilon \sim t(3)$						Case 4: $\varepsilon \sim \text{Cauchy}(0,1)$					
SIS	1.00	1.00	0.30	0.72	0.22	355.9(300.0)	0.49	0.35	0.08	0.15	0.00	694.7(237.1)
SIRS	1.00	1.00	0.15	0.82	0.12	459.9(314.2)	1.00	1.00	0.12	0.69	0.08	482.1(291.1)
DC-SIS	1.00	1.00	1.00	0.87	0.87	33.9(79.1)	0.98	0.94	0.95	0.56	0.56	137.1(189.3)
MC-SIS	1.00	1.00	1.00	0.89	0.89	32.0(79.1)	1.00	1.00	1.00	0.76	0.76	67.8(132.6)
MC-SCIS	1.00	1.00	1.00	1.00	1.00	5.1(7.6)	1.00	0.99	1.00	0.99	0.98	9.4(16.6)

Table 8

The Model Size, MSE and correlation between the predicted value and observed value of Y . The results are based on 2000 random partitions.

Method	Model Size	MSE			corr($\hat{Y}$, Y)		
		median	mean	sd	50%	mean	sd
SIS-ALasso	8	0.385	0.429	0.185	0.695	0.686	0.170
SIRS-ALasso	7	0.385	0.433	0.191	0.696	0.688	0.165
DC-SIS-ALasso	6	0.400	0.460	0.226	0.694	0.679	0.170
MC-SIS-ALasso	8	0.378	0.410	0.166	0.709	0.698	0.159
SIS-SCAD	4	0.373	0.417	0.179	0.700	0.692	0.168
SIRS-SCAD	2	0.413	0.483	0.255	0.690	0.679	0.180
DC-SIS-SCAD	3	0.386	0.452	0.226	0.701	0.682	0.184
MC-SIS-SCAD	4	0.366	0.408	0.166	0.709	0.704	0.160
HMZ	19	0.375	0.416	0.187	0.718	0.713	0.149

Table 9
The R square and MSE of MC-SIS and CC-SIS of every part of the data that partitioned according to age

Age	Sample Size	R square		MSE	
		MC-SCIS	CC-SIS	MC-SCIS	CC-SIS
29 Age < 35	232	0.463	0.352	0.578	0.697
35 Age < 40	336	0.438	0.341	0.520	0.610
40 Age < 45	159	0.486	0.373	0.608	0.742
45 Age < 50	133	0.535	0.488	0.358	0.394
50 Age < 61	116	0.325	0.637	0.711	0.382
29 Age > 61	976	0.300	0.255	0.700	0.745