DEPARTMENT OF STATISTICS
The Pennsylvania State University
University Park, PA 16802 U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 13-01: March 2013

Feature Selection for Varying Coefficient Models with Ultrahigh Dimensional Covariates

Jingyuan Liu[1], Runze Li[2] and Rongling Wu[3]

[1]Department of Statistics, The Pennsylvania State University
[2]Department of Statistics and The Methodology Center, The Pennsylvania State University
[3]Department of Public Health Science, Penn State Hershey College of Medicine

Feature Selection for Varying Coefficient Models With Ultrahigh Dimensional Covariates

Jingyuan Liu, Runze Li and Rongling Wu
The Pennsylvania State University

March 2013

Abstract

This paper is concerned with feature screening and variable selection for varying coefficient models with ultrahigh dimensional covariates. We propose a new feature screening procedure for these models based on conditional correlation coefficient. We systematically study the theoretical properties of the proposed procedure, and establish their sure screening property and the ranking consistency. To enhance the finite sample performance of the proposed procedure, we further develop an iterative feature screening procedure. Monte Carlo simulation studies were conducted to examine the performance of the proposed procedures. In practice, we advocate a two-stage approach for varying coefficient models. The two stage approach consists of (a) reducing the ultrahigh dimensionality by using the proposed procedure and (b) applying regularization methods for dimension-reduced varying coefficient models to make statistical inferences on the coefficient functions. We illustrate the proposed two-stage approach by a real data example.

Key Word: Feature selection, varying coefficient models, ranking consistency, sure screening property.

*Jingyuan Liu is a graduate student, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111. Email: jul221@psu.edu. Runze Li is Distinguished Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111. Email: rzli@psu.edu. His research was supported by National Institute on Drug Abuse (NIDA) grant P50-DA10075. Rongling Wu is Professor, Department of Public Health Sciences, Penn State Hershey College of Medicine, Hershey, PA 17033. Email: RWu@phs.psu.edu. His research was supported by a NSF grant IOS-0923975 and a NIH grant, UL1RR0330184. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH or NIDA.

1. Introduction

Varying coefficient models with ultrahigh dimensional covariates (*ultrahigh dimensional varying coefficient models* for short) could be very useful for analyzing genetic study data to examine varying gene effects. This study was motivated by an empirical analysis of a subset of Framingham Heart Study (FHS) data. See Section 4 for more details. Of interest in this empirical analysis is to identify genes strongly associated with body mass index (BMI). Some initial exploratory analysis on this data subset indicates that the effects of genes on the BMI are age-dependent. Thus, it is natural to apply varying coefficient model for this analysis. There are thousands of single-nucleotide polymorphisms available in the FHS database. This leads us to consider ultrahigh dimensional varying coefficient models. It is typical that only hundreds of samples are available in genetic study data. Thus, feature screening and variable selection become indispensable for estimation of ultrahigh dimensional varying coefficient models.

Some variable selection methods have been developed for varying coefficient models with low dimensional covariates in literature. Li and Liang (2008) proposed a generalized likelihood ratio test to select significant covariates with varying effects. Wang, Li and Huang (2008) developed a regularized estimation procedure based on the basis function approximations and the SCAD penalty (Fan and Li, 2001) to simultaneously select significant variables and estimate the nonzero smooth coefficient functions. Wang and Xia (2009) proposed a shrinkage method integrated local polynomial regression techniques (Fan and Gijbels, 1996) and the LASSO regression (Tibshirani, 1996). Nevertheless, these variable selection procedures were developed for the varying coefficient models with fixed dimensional covariates. As a result, these procedures cannot be directly applied to the ultrahigh dimensional varying coefficient models.

To deal with the ultrahigh dimensionality, one appealing method is the two-stage approach. First, a computationally efficient screening procedure is applied to reduce the ul-

trahigh dimensionality to a moderate scale under sample size, and then the final sparse model is recovered from the screened submodel by a regularization method. Several screening techniques for the first stage have been developed for various models. Fan and Lv (2008) showed that the sure independence screening (SIS) possesses sure screening property in the linear model setting. Hall and Miller (2009) extended the methodology from linear models to nonlinear models using generalized empirical correlation learning, but it is not trivial to choose the optimal transformation function. Fan and Song (2010) extended SIS to the generalized linear model by ranking the maximum marginal likelihood estimates. Fan, Feng and Song (2011) explored the feature screening technique for ultrahigh dimensional additive models, by ranking the magnitude of spline approximations of the nonparametric components. Zhu, Li, Li and Zhu (2011) proposed a sure independence ranking and screening procedure to select important predictors under the multi-index model setting. Li, Peng, Zhang and Zhu (2012) proposed rank correlation feature screening for a class of semiparametric models, such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation, even when there are nonparametric functions in the models. Model-free screening procedures have been advocated in the literature. Li, Zhong and Zhu (2012) developed a model free feature screening procedure based on a distance correlation, which are directly applicable for multiple response and grouped predictors. He, Wang and Hong (2013) proposed a quantile-adaptive model-free feature screening procedure for heterogeneous data. This paper aims to develop kernel-regression based feature screening method for ultrahigh dimensional varying coefficient models to reduce dimensionality.

Suppose that the varying-coefficients in the varying coefficient models are a function of covariate u. Thus, conditioning on u, the varying coefficient models are linear models. Therefore, it is natural to employ the conditional Pearson correlation coefficient as a measure for the strength of association between a predictor and the response. In this paper, we propose using kernel regression techniques to estimate the conditional correlation coefficients,

and further develop a marginal utility for feature screening based on the kernel regression estimate. We investigate the finite sample performance of the proposed procedure via Monte Carlo simulation study and illustrate the proposed methodology by an empirical analysis of a subset of FHS data. This paper makes the following contribution to the literature. We first establish the concentration inequality for the kernel regression estimate of the conditional Pearson correlation coefficient. Based on the concentration inequality, we further establish several desirable theoretical properties for the proposed procedure. We show that the proposed procedure possesses the consistency in ranking property (Zhu, *et al.*, 2011). By consistency in ranking, it means with probability tending to 1, the important predictors rank before the unimportant ones. We also show that the proposed procedure enjoys the sure screening property (Fan and Lv, 2008) under the setting of ultrahigh dimensional varying coefficient models. The sure screening property guarantees the probability that the model chosen by our screening procedure includes the true model tends to 1 in an exponential rate of the sample size.

The rest of the paper is organized as follows. In section 2, we propose a new feature screening procedure for ultrahigh dimensional varying coefficient models. In this section, we also study the theoretical property of the proposed procedure. In section 3, Monte Carlo simulations are conducted to assess the finite performance of the proposed procedure. In section 4, we propose a two-stage approach for ultrahigh dimensional varying coefficient models, and illustrate the approach by examining the age-specific SNP effects on BMI using the FHS data. Conclusion remark is given in section 5, and the technical proofs are given in the appendices.

2. A New Feature Screening Procedure

Let y be the response, and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ be the p-dimensional predictor.

Consider a varying coefficient model

$$y = \beta_0(u) + \mathbf{x}^T\boldsymbol{\beta}(u) + \varepsilon, \tag{2.1}$$

where $E(\varepsilon|\mathbf{x}, u) = 0$, $\beta_0(u)$ is the intercept function and $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_p(u))^{\mathrm{T}}$ consists of p unknown smooth functions $\beta_j(u)$, $j = 1, \ldots, p$, of u.

Note that given u, the varying coefficient model becomes a linear regression model. Fan and Lv (2008) proposed a sure independence screening procedure for linear regression model based on Pearson correlation coefficient. Thus, it is natural to consider conditional Pearson correlation coefficient for feature screening. Specifically, given u, the conditional correlation between the response y and each predictor x_j, $j = 1, \ldots, p$, is defined as

$$\rho(x_j, y|u) = \frac{\mathrm{cov}(x_j, y|u)}{\sqrt{\mathrm{cov}(x_j, x_j|u)\mathrm{cov}(y, y|u)}}, \tag{2.2}$$

which is a function of u. Intuitively define the marginal utility for feature screening as

$$\rho^*_{j0} = E\{\rho^2(x_j, y|u)\}.$$

To estimate ρ^*_{j0}, let us proceed with estimation of $\rho(x_j, y|u)$, which essentially requires estimation of five conditional means $E(y|u)$, $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$ and $E(x_j y|u)$. Throughout this paper, it is assumed that these five conditional means are nonparametric smoothing functions of u. Therefore, the conditional correlation in (2.2) can be estimated through nonparametric mean estimation techniques. We will use the kernel smoothing method (Fan and Gijbels, 1996) to estimate these conditional means.

Suppose $\{(u_i, \mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is a random sample from (2.1). Let $K(t)$ be a kernel function, and h be a bandwidth. Then the kernel regression estimates for $E(y|u)$ is given

$$\widehat{E}(y|u) = \sum_{i=1}^{n} \frac{K_h(u_i - u)y_i}{\sum_{i=1}^{n} K_h(u_i - u)}, \tag{2.3}$$

where $K_h(t) = h^{-1}K(t/h)$. Similarly, we may define kernel regression estimate $\widehat{E}(y^2|u)$, $\widehat{E}(x_j|u)$, $\widehat{E}(x_j^2|u)$ and $\widehat{E}(x_jy|u)$ for $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$ and $E(x_jy|u)$, respectively. The conditional covariance $\mathrm{cov}(x_j, y|u)$ can be estimated by $\widehat{\mathrm{cov}}(x_j, y|u) = \widehat{E}(x_jy|u) - \widehat{E}(x_j|u)\widehat{E}(y|u)$, and the conditional correlation is naturally estimated by

$$\widehat{\rho}(x_j, y|u) = \frac{\widehat{\mathrm{cov}}(x_j, y|u)}{\sqrt{\widehat{\mathrm{cov}}(x_j, x_j|u)\widehat{\mathrm{cov}}(y, y|u)}}. \tag{2.4}$$

Remark. We employ the kernel regression rather than local linear regression because local linear regression estimate cannot guarantee $\widehat{\mathrm{cov}}(y, y|u) \geq 0$ and $\widehat{\mathrm{cov}}(x_j, x_j|u) \geq 0$. Furthermore, it is required to set the bandwidth h the same for all the five conditional means in order to guarantee that $|\hat{\rho}(x_j, y|u)| \leq 1$. In our simulations and real data example, we first select an optimal bandwidth for $E(x_jy|u)$ by using plug-in method (Ruppert, Sheather and Wang, 1995), and then use this bandwidth for other four conditional means.

The plug-in estimate of ρ_{j0}^* is

$$\widehat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2(x_j, y|u_i). \tag{2.5}$$

Based on $\widehat{\rho}_j^*$, we propose a screening procedure for ultrahigh dimensional varying coefficient models as follows: sort $\widehat{\rho}_j^*,\ j = 1, \ldots, p$ in the decreasing order, and defines the screened submodel as

$$\widehat{\mathcal{M}} = \{j : 1 \leq j \leq p,\ \widehat{\rho}_j^* \text{ ranks among the first } d\}, \tag{2.6}$$

where the submodel size d is taken to be smaller than the sample size n. Thus, the ultrahigh dimensionality p is reduced to the moderate scale d. Fan and Lv (2008) suggested setting $d = [n/\log(n)]$, where $[a]$ refers to the integer part of a. In the kernel regression setting, it is known that the effective sample size is nh rather than n, and the optimal rate of the bandwidth $h = O(n^{-1/5})$ (Fan and Gijbels, 1996). Thus we may set $d = [n^{4/5}/\log(n^{4/5})]$ for ultrahigh dimensional varying coefficient models. We will examine the impact of the

choice of d in our simulation by considering $d = \nu[n^{4/5}/\log(n^{4/5})]$ with different values for ν. This proposed procedure is referred to as conditional correlation sure independence screening (CC-SIS for short).

We next study the theoretical properties of the newly proposed screening procedure CC-SIS. Let us introduce some notation first. The support of u is assumed to be bounded and denoted by $\mathbb{U} = [a, b]$ with finite constants a and b. Define the true model index set and its complement by

$$\begin{aligned} \mathcal{M}_* &= \{1 \le j \le p : \ \beta_j(u) \neq 0 \text{ for some } u \in \mathbb{U}\}, \\ \mathcal{M}_*^c &= \{1 \le j \le p : \ \beta_j(u) \equiv 0 \text{ for any } u \in \mathbb{U}\}. \end{aligned}$$

The following regularity conditions are imposed in order to establish the ranking consistency property and sure screening property of the proposed screening procedure.

(C1) Denote the density function of u by $f(u)$. Assume that $f(u)$ has continuous second order derivative on $\mathbb{U}$. Further assume that $\sup_{u\in\mathbb{U}} f(u) \le M_1$, $\sup_{u\in\mathbb{U}} |f'(u)| \le M_2$ and $\sup_{u\in\mathbb{U}} |f''(u)| \le M_3$ for some finite constants M_1, M_2 and M_3.

(C2) The kernel function $K(\cdot)$ is bounded uniformly such that $\sup_{u\in\mathbb{U}} |K(u)| \le M_4 < \infty$. And $\mu_2(K) = \int t^2 K(t)dt < \infty$.

(C3) The random variables x_j and y satisfy the sub-exponential tail probability uniformly in p. That is, there exists $s_0 > 0$, such that for $0 \le s < s_0$,

$$\begin{aligned} &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E\{\exp(2sx_j^2|u)\} < \infty, \quad \sup_{u\in\mathbb{U}} E\{\exp(2sy^2|u)\} < \infty, \\ &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E\{\exp(2sx_jy|u)\} < \infty. \end{aligned}$$

(C4) All conditional means $E(y|u)$, $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$ and $E(x_jy|u)$, their first and

second order derivatives are finite uniformly in $u \in \mathbb{U}$. Further assume that

$$\inf_{u\in\mathbb{U}} \min_{1\le j\le p} \mathrm{var}(x_j|u) > 0, \quad \inf_{u\in\mathbb{U}} \mathrm{var}(y|u) > 0.$$

Condition (C1) and (C2) are mild conditions on the density function $f(u)$ and the kernel function $K(\cdot)$, which can be guaranteed by most commonly used distributions and kernels. Moreover, (C2) implies that $K(\cdot)$ has every finite moment, i.e. $E(|K(u)|^r) < \infty, \quad \forall r > 0$. Condition (C3) is relatively strong and only used to facilitate the technical proofs. Condition (C4) require the mean-related quantities bounded and the variances positive, in order to guarantee that the conditional correlation is well defined. We first establish the ranking consistency property of CC-SIS.

Theorem 1. *(Ranking Consistency Property) Suppose that*

$$\liminf_{n\to\infty} \{ \min_{j\in\mathcal{M}_*} \rho^*_{j0} - \max_{j\in\mathcal{M}_*^c} \rho^*_{j0} \} > 0, \tag{2.7}$$

and the bandwidth $h \to 0$ but $nh^3 \to \infty$ as $n \to \infty$. Under conditions (C1)-(C4), for $p = o\{\exp(an)\}$ with some $a > 0$, we have

$$\liminf_{n\to\infty} \{ \min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j \} > 0 \quad \textit{in probability.}$$

Condition (2.7) provides a clear separation between the important and unimportant predictors in terms of the population level unconditioned-squared correlation ρ^*_{j0}. This condition rules out the situation when certain unimportant predictors have large ρ^*_{j0}'s and are selected only because they are highly correlated with the true ones. Theorem 1 states that with an overwhelming probability, the truly important predictors have larger $\widehat{\rho}^*$'s than the unimportant ones, and hence all the true predictors are ranked in the top by the proposed screening procedure. We next develop the sure screening property of CC-SIS.

Theorem 2. *(Sure Screening Property) Under conditions (C1)-(C4), suppose the bandwidth*

$h = O(n^{-\gamma})$ where $0 < \gamma < 1/3$, then we have

$$P\left(\max_{1\le j\le p} |\widehat{\rho}_j^* - \rho_{j0}^*| > c_3 \cdot n^{-\kappa}\right) \le O\{np\exp(-n^{\frac{1}{5}-\kappa}/\xi)\},$$

and if we further assume that there exist some $c_3 > 0$ and $0 \le \kappa < \gamma$, such that

$$\min_{j\in\mathcal{M}_*} \rho_{j0}^* \ge 2c_3 n^{-\kappa}. \tag{2.8}$$

then

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \ge 1 - O\{ns_n\exp(-n^{\gamma-\kappa}/\xi)\},$$

where ξ is some positive constant determined by c_3, and s_n is the cardinality of $\mathcal{M}_*$, which is sparse and may vary with n.

Condition (2.8) guarantees the true unconditioned-squared correlations between the important x_j's and y to be bounded away from 0. However, the lower bound depends on n, thus ρ_{j0}^*'s are allowed to go to 0 in the asymptotic sense. This condition rules out the situation where the predictors are marginally uncorrelated with y but jointly correlated. Theorem 2 ensures that the probability of the true model being selected into the screened submodel by CC-SIS tends to 1 with an exponential rate.

3. Numerical Examples and Extensions

In this section, we first conduct Monte Carlo simulation study to illustrate the ranking consistency and the sure screening property of the proposed procedure empirically, and compare its finite sample performance with some other screening procedures under different model settings. We further consider a two-stage approach for analyzing ultrahigh dimensional data using varying coefficient models in section 3.2. We study an iterative sure screening procedure to enhance finite sample performance of CC-SIS in section 3.3.

For each simulation example (i.e. Examples 1 and 3 below), the covariate u and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ are generated as follows: First draw u^* and $\mathbf{x}$ from $(u^*, \mathbf{x}) \sim N(\mathbf{0}, \Sigma)$, where

Σ is a $(p+1)\times(p+1)$ covariance matrix with element $\sigma_{ij}=\rho^{|i-j|}$, $\quad i,j=1,\cdots,p+1$. We consider $\rho=0.8$ and 0.4 for a high correlation and a low correlation, respectively. Then take $u=\Phi(u^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus, u follows a uniform distribution $U(0,1)$ and is correlated with $\mathbf{x}$, and all the predictors $x_1,\ldots,x_p$ are correlated with each other. The random noise ε is drawn from $N(0,1)$. The model dimension p is taken to be 1000, and the sample size n is 200. This leads to $[n^{4/5}/\log(n^{4/5})]=16$. In our simulation we consider $d=\nu[n^{4/5}/\log(n^{4/5})]$ with $\nu=1$, 2 and 3. All the simulation results are based on 1000 replications.

The following criteria are used to assess the performance of CC-SIS:

- R_j: The average of the ranks of x_j in terms of the screening criterion based on 1000 replications. For SIS, R_j is the average rank of the Pearson correlation between x_j and y in the decreasing order; for CC-SIS, R_j refers to the average rank of $\widehat{\rho}_j^*$.

- M: the minimum size of the submodel that contains all the true predictors. In other words, M is the largest rank of the true predictors: $M=\max_{j\in\mathcal{M}_*}R_j$, where $\mathcal{M}_*$ is the true model. We report the 5%, 25%, 50%, 75% and 95% quantiles of M from 1000 repetitions.

- p_a: The proportion of all truly important predictors being selected into $\widehat{\mathcal{M}}$ with size d over 1000 simulations.

- p_j: The proportion of x_j being selected into the submodel $\widehat{\mathcal{M}}$ with size d over 1000 simulations.

The criteria are used to empirically verify the theoretical properties in Theorems 1 and 2. The ranking consistency of a screening procedure refers to the property that the screening scores of true predictors rank above the unimportant ones, hence a reasonable screening procedure is expected to guarantee that R_j's of the true predictors are small, and so is the minimum submodel size M. The sure screening property claims an overwhelming probability

of all true predictors being selected into $\widehat{\mathcal{M}}$, thus it can be verified if the overall selection rate p_a and the individual selection rates p_j's of the important x_j's are close to one. In addition, M being smaller than d also implies that all important predictors are included in the submodel with size d.

3.1. Monte Carlo simulation

In this section, we conduct Monte Carlo simulations to examine the finite sample performance of CC-SIS, and compare its performance with that of SIS (Fan and Lv, 2008), SIRS (Zhu, *et al.*, 2011) and DC-SIS (Li, Zhong and Zhu, 2012).

Example 1. The true model index set in this example is taken to be $\mathcal{M}_* = \{2, 100, 400, 600, 1000\}$. To make a fair comparison, we consider the following two cases for coefficient functions. In Case I, the nonzero coefficient functions are truly varying over u, while in Case II, the nonzero coefficient functions are constants, therefore the true model indeed is a linear model. Specifically, the coefficient functions are given below.

Case I. The nonzero coefficient functions are defined by

$$\beta_2(u) = 2I(u > 0.4), \quad \beta_{100}(u) = 1 + u, \quad \beta_{400}(u) = (2 - 3u)^2$$
$$\beta_{600}(u) = 2\sin(2\pi u), \quad \beta_{1000}(u) = \exp\{u/(u + 1)\}.$$

Case II. The nonzero coefficient functions are defined by

$$\beta_2(u) = 1, \quad \beta_{100}(u) = 0.8, \quad \beta_{400}(u) = 1.2, \quad \beta_{600}(u) = -0.8, \quad \beta_{1000}(u) = -1.2.$$

Table 1 reports R_j's of the active predictors. From the output, the ranking consistency of CC-SIS is demonstrated by the fact that $\widehat{\rho}_j^*$'s of the active predictors rank in the top for both $\rho = 0.4$ and $\rho = 0.8$. The Pearson correlation from SIS, however, ranks x_{600} behind and leaves it aliased with the unimportant x_j's. The reason is that $\beta_{600}(u) = 2\sin(2\pi u)$ has mean 0 if u is considered as a random variable from $U(0, 1)$. Therefore, when we mis-specify the

Table 1: R_j of the true predictors for Example 1.

Method	Case I: varying coefficient model					Case II: linear model				
	R_2	R_{100}	R_{400}	R_{600}	R_{1000}	R_2	R_{100}	R_{400}	R_{600}	R_{1000}
	$\rho = 0.4$									
SIS	3.516	1.542	6.480	461.341	2.209	3.074	5.367	1.744	5.392	1.729
SIRS	3.727	1.580	10.669	486.762	2.148	3.162	6.328	1.773	5.764	1.760
DC-SIS	3.110	1.649	10.137	350.455	2.241	3.190	6.351	1.790	6.849	1.821
CC-SIS	2.733	2.112	3.795	3.806	3.261	3.138	6.573	1.740	7.001	1.736
	$\rho = 0.8$									
SIS	7.940	1.761	13.981	468.062	3.126	5.235	12.633	1.896	12.850	2.160
SIRS	9.391	1.771	15.691	454.474	3.199	5.585	14.054	2.046	14.350	2.272
DC-SIS	6.312	1.884	12.894	341.803	3.439	5.621	13.576	2.128	15.015	2.134
CC-SIS	6.771	2.094	6.858	5.905	3.896	9.171	12.307	1.711	12.674	1.927

Table 2: The selecting rates p_a and p_j's for Example 1.

d	Method	Case I: varying coefficient model						Case II: linear model					
		p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
		$\rho = 0.4$											
16	SIS	0.989	1.000	0.951	0.031	0.997	0.029	1.000	0.980	1.000	0.977	1.000	0.957
	SIRS	0.986	1.000	0.897	0.011	1.000	0.010	0.998	0.964	1.000	0.968	1.000	0.931
	DC-SIS	0.994	1.000	0.917	0.017	0.997	0.015	0.997	0.959	1.000	0.960	0.999	0.917
	CC-SIS	0.999	1.000	0.999	0.995	0.996	0.989	1.000	0.957	1.000	0.951	1.000	0.910
32	SIS	0.997	1.000	0.979	0.068	1.000	0.066	1.000	0.996	1.000	0.992	1.000	0.988
	SIRS	0.993	1.000	0.944	0.027	1.000	0.025	1.000	0.981	1.000	0.987	1.000	0.968
	DC-SIS	1.000	1.000	0.954	0.050	0.999	0.048	0.999	0.984	1.000	0.983	1.000	0.966
	CC-SIS	0.999	1.000	1.000	0.999	0.999	0.997	1.000	0.979	1.000	0.978	1.000	0.957
48	SIS	0.998	1.000	0.990	0.088	1.000	0.087	1.000	0.998	1.000	0.998	1.000	0.996
	SIRS	0.998	1.000	0.959	0.036	1.000	0.034	1.000	0.990	1.000	0.994	1.000	0.984
	DC-SIS	1.000	1.000	0.972	0.084	1.000	0.082	1.000	0.992	1.000	0.987	1.000	0.979
	CC-SIS	0.999	1.000	1.000	1.000	0.999	0.998	1.000	0.991	1.000	0.983	1.000	0.974
		$\rho = 0.8$											
16	SIS	0.925	0.999	0.813	0.003	0.996	0.002	0.985	0.794	1.000	0.782	1.000	0.612
	SIRS	0.898	1.000	0.755	0.002	0.993	0.001	0.969	0.756	1.000	0.746	0.999	0.546
	DC-SIS	0.958	1.000	0.814	0.004	0.993	0.003	0.967	0.733	1.000	0.732	0.998	0.518
	CC-SIS	0.960	0.997	0.956	0.973	0.994	0.885	0.911	0.826	1.000	0.820	1.000	0.617
32	SIS	0.984	1.000	0.946	0.025	1.000	0.023	1.000	0.966	1.000	0.971	1.000	0.938
	SIRS	0.974	1.000	0.920	0.026	1.000	0.023	1.000	0.947	1.000	0.943	1.000	0.893
	DC-SIS	0.994	1.000	0.941	0.021	0.998	0.020	1.000	0.963	1.000	0.938	1.000	0.903
	CC-SIS	0.999	1.000	0.996	0.999	1.000	0.994	0.991	0.968	1.000	0.963	1.000	0.924
48	SIS	0.992	1.000	0.968	0.053	1.000	0.051	1.000	0.989	1.000	0.987	1.000	0.976
	SIRS	0.986	1.000	0.957	0.046	1.000	0.043	1.000	0.976	1.000	0.968	1.000	0.945
	DC-SIS	0.998	1.000	0.965	0.050	0.999	0.048	1.000	0.984	1.000	0.973	1.000	0.957
	CC-SIS	0.999	1.000	0.999	1.000	1.000	0.998	0.999	0.984	1.000	0.983	1.000	0.966

Table 3: The quantiles of M for Example 1.

	Case I: varying coefficient model					Case II: linear model				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	$\rho = 0.4$									
SIS	25.95	185.75	454.50	728.25	951.00	5.00	5.00	5.00	7.00	14.00
SIRS	60.95	223.75	475.00	744.25	943.00	5.00	5.00	6.00	7.00	22.00
DC-SIS	32.00	135.00	299.50	522.50	841.10	5.00	5.00	6.00	8.00	24.00
CC-SIS	5.00	5.00	5.00	5.00	7.00	5.00	5.00	6.00	7.00	29.05
	$\rho = 0.8$									
SIS	52.00	228.75	442.50	701.00	951.00	8.00	11.00	15.00	20.00	35.05
SIRS	51.95	208.00	440.00	675.00	922.20	8.00	12.00	15.00	22.00	51.00
DC-SIS	51.00	142.00	297.00	500.25	790.05	8.00	12.00	16.00	21.00	41.05
CC-SIS	6.00	8.00	10.00	13.00	20.00	8.00	11.00	15.00	20.00	44.00

varying coefficient model as a linear regression model and apply SIS, the constant coefficient β_{600} is indeed 0, and hence the true marginal correlation between x_{600} and y is 0. Therefore, the magnitude of the Pearson correlation for x_{600} is expected to be small, although x_{600} is functionally important as successfully detected by CC-SIS. In addition, SIRS and DC-SIS both fail to identify x_{600} likewise under the varying coefficient model setting.

The proportions p_a and p_j's for the important predictors are tabulated in Table 2. All p_a and p_j's of CC-SIS are close to one, even for the smallest $d = 16$, which illustrates the sure screening property. While the low p_{600} and p_a values for the other three screening procedures imply their failure of detecting x_{600}, and increasing the submodel size d does not help much.

Similar conclusions can be drawn from Table 3. SIS, SIRS and DC-SIS need large models to include all the true predictors due to the low rank of x_{600}. Consequently, the models with size d do not guarantee all the important predictors to be selected, even with the largest $d = 48$. CC-SIS, on the other hand, requires only fairly small models, and thus all of the important variables can be selected with any of the three choices of size d. Therefore, both ranking consistency and sure screening property are illustrated in this table.

In addition, by comparing the models with the two different ρ's, we observe that the ones with $\rho = 0.4$ typically perform better than those with $\rho = 0.8$ for all the four screening

procedures. This is because when the predictors are highly correlated ($\rho = 0.8$), the screening scores of some unimportant variables are inflated by the adjacent important ones, hence the unimportant predictors may be selected due to their strong correlation with the true predictors.

For Case II, the four screening procedures perform similarly well in terms of all the criteria. Thus CC-SIS is still valid for linear models, but it pays a price of computational cost. Therefore, if the underlying model is known to be linear, one may prefer SIS due to its easier implementation.

3.2. Two-stage approach for varying coefficient models and an application

Consider the varying coefficient model (2.1). Although CC-SIS can reduce the ultrahigh dimensionality p to the moderate scale d, a subsequent step is needed to further select the significant variables and recover the final sparse model. In this section, we discuss the entire variable section procedure, referred to as a two-stage approach.

In the screening stage, CC-SIS is conducted to obtain the submodel index set (2.6) with size $d = [n^{4/5}/\log(n^{4/5})]$. For the ease of presentation, we denote the screened submodel by

$$y = \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon. \tag{3.1}$$

The screened predictor $\mathbf{x} = (1, x_{s_1}, \ldots, x_{s_d})^{\mathrm{T}} \in \mathbb{R}^{d+1}$ where $s_i \in \widehat{\mathcal{M}}$ in (2.6), and the screened coefficient vector $\boldsymbol{\beta}(u) = (\beta_0(u), \beta_{s_1}(u), \ldots, \beta_{s_d}(u))^{\mathrm{T}} \in \mathbb{R}^{d+1}$.

In the post-screening variable selection stage, the modified penalized regression procedures are applied to further select important variables and estimate the coefficient function $\boldsymbol{\beta}(u)$ in model (3.1). Following the idea of the KLASSO method (Wang and Xia, 2009), we aim to estimate the $n \times (d+1)$ matrix

$$\mathbf{B} = \{\boldsymbol{\beta}(u_1), \ldots, \boldsymbol{\beta}(u_n)\}^{\mathrm{T}} = (\mathbf{b}_1, \ldots, \mathbf{b}_{d+1}),$$

where $\mathbf{b}_j = (\beta_j(u_1), \ldots, \beta_j(u_n))^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$ is the jth column of $\mathbf{B}$. The estimator $\widehat{\mathbf{B}}_\lambda$ of $\mathbf{B}$ is defined by

$$\widehat{\mathbf{B}}_\lambda = \mathrm{argmin}_{\mathbf{B}\in\mathbb{R}^{n\times(d+1)}} \left\{ \sum_{t=1}^{n}\sum_{i=1}^{n} \{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(u_t)\}^2 K_h(u_t - u_i) + n\sum_{j=1}^{d+1} p_\lambda(\|\mathbf{b}_j\|) \right\}, \quad (3.2)$$

where $\|\cdot\|$ is the Euclidean norm, $p_\lambda(\cdot)$ is the penalty function, and λ is the tuning parameter to be chosen by a data-driven method.

With a chosen λ, a modified iterative algorithm based on the local quadratic approximation (Fan and Li, 2001) is applied to solve the minimization problem (3.2):

1. Set the initial value $\widehat{\mathbf{B}}_\lambda^{(0)}$ to be the unpenalized estimator (Fan and Zhang, 2000b):

$$\widehat{\mathbf{B}}_\lambda^{(0)} = \mathrm{argmin}_{\mathbf{B}\in\mathbb{R}^{n\times(d+1)}} \left\{ \sum_{t=1}^{n}\sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(u_t))^2 K_h(u_t - u_i) \right\},$$

2. Denote the mth-step estimator of $\mathbf{B}$ by

$$\widehat{\mathbf{B}}_\lambda^{(m)} = \{\widehat{\boldsymbol{\beta}}_\lambda^{(m)}(u_1), \ldots, \widehat{\boldsymbol{\beta}}_\lambda^{(m)}(u_n)\}^{\mathrm{T}} = (\widehat{\mathbf{b}}_{\lambda,1}^{(m)}, \ldots, \widehat{\mathbf{b}}_{\lambda,d+1}^{(m)}).$$

   Then the $(m+1)$th-step estimator is $\widehat{\mathbf{B}}_\lambda^{(m+1)} = \{\widehat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_1), \ldots, \widehat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_n)\}^{\mathrm{T}}$, with

$$\widehat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_t) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} K_h(u_t - u_i) + \mathbf{D}^{(m)} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i y_i K_h(u_t - u_i) \right), \quad (3.3)$$

   where the matrix $\mathbf{D}^{(m)}$ is a $(d+1)\times(d+1)$ diagonal matrix with the jth diagonal component $\mathbf{D}_{jj}^{(m)} = \{p'_\lambda(\|\widehat{\mathbf{b}}_{\lambda,j}^{(m)}\|)\}/\{2\|\widehat{\mathbf{b}}_{\lambda,j}^{(m)}\|\}$.

3. Iterate step 2 for $m = 1, 2, \ldots$ until convergence.

We can adopt various penalty functions to obtain different $\mathbf{D}^{(m)}$'s in (3.3). In this section, we consider the LASSO penalty, the adaptive LASSO penalty and the SCAD penalty. Specifically, the LASSO penalty (Tibshirani, 1996) yields $\mathbf{D}_{jj}^{(m)} = \lambda/\|\widehat{\mathbf{b}}_{\lambda,j}^{(m)}\|$; the adaptive LASSO (Zou, 2006) replaces λ with the coefficient-specific parameter, that is, $\mathbf{D}_{jj}^{(m)} = \lambda_j/\|\widehat{\mathbf{b}}_{\lambda,j}^{(m)}\|$,

where $\lambda_j = \lambda/\|\widehat{\mathbf{b}}_{\lambda,j}^{(0)}\|$; and the SCAD penalty (Fan and Li, 2004) gives

$$\mathbf{D}_{jj}^{(m)} = \frac{1}{2\|\widehat{\mathbf{b}}_j^{(m)}\|}\left\{\lambda I(\|\widehat{\mathbf{b}}_j^{(m)}\| \le \lambda) + \frac{(a\lambda - \|\widehat{\mathbf{b}}_j^{(m)}\|)_+ \cdot I(\|\widehat{\mathbf{b}}_j^{(m)}\| > \lambda)}{a-1}\right\}.$$

We next illustrate the proposed two-stage approach by an empirical analysis of Framingham Heart Study data.

Table 4: The sizes and MSPE of the nine models

	CC-SIS+LASSO	CC-SIS+ALASSO	CC-SIS+SCAD
AIC	43 (0.405)[1]	40 (0.401)	34 (0.380)
BIC	42 (0.395)	38 (0.400)	34 (0.380)
GCV	43 (0.405)	40 (0.401)	34 (0.380)

[1] The numbers in the parentheses are MSPE of the model.

Example 2. The Framingham Heart Study is a cardiovascular study beginning in 1948 under the direction of the National Heart, Lung and Blood Institute (Dawber et al., 1951; Jaquish, 2007). In our analysis, 349, 985 non-rare single-nucleotide polymorphisms (SNPs) are of interest, and the data from 977 patients are available. The goal is to detect the SNPs that are important to explaining the body mass index (BMI). For each SNP, both dominant effect and additive effect are considered, thus the dimensionality is 699, 970, much larger than the sample size 977. In addition, one may argue that the effect of SNPs on BMI might change with age. Therefore, the varying coefficient model (2.1) is appropriate, where y is BMI, $\mathbf{x}$ is the SNP vector, and u is age.

To select the significant SNPs, the proposed two-stage approach is applied, based on three penalties: LASSO, Adaptive LASSO (ALASSO) and SCAD, along with three tuning parameter selection criteria: AIC, BIC and GCV. The model sizes of the nine selected models are tabulated in Table 4, in which the smaller models are nested within the bigger ones, and the same size indicates the identical model. Thus, there are only five different models out of the nine selected models. The median squared prediction error (MSPE) of the nine models are reported in the parentheses of Table 4. One can see that CC-SIS+SCAD two-stage approach yields the sparsest model with size 34 and the smallest MSPE. Furthermore, the

pairwise likelihood ratio tests for the nested varying coefficient models (Fan, Zhang and Zhang, 2001) are conducted. The p-values are shown in Table 5, which indicate the sparsest model chosen by CC-SIS+SCAD is sufficient.

Table 5: The p-values of the pairwise generalized likelihood ratio tests

				H_1		
		Unpenalized	LASSO-AIC	LASSO-BIC	ALASSO-AIC	ALASSO-BIC
	LASSO-AIC	0.9952	·	·	·	·
	LASSO-BIC	0.9999	0.9462	·	·	·
H_0	ALASSO-AIC	0.9999	0.9998	0.9995	·	·
	ALASSO-BIC	0.9999	0.9967	0.9854	0.7481	·
	SCAD	0.9999	0.9991	0.9965	0.9516	0.9268

Table 6 contains the information of the SNPs chosen by CC-SIS+SCAD. The last column of the table indicates which effect (dominant or additive) of the selected SNP is significant. And Figure 1 is the plot of the estimated coefficient functions versus age, which depicts the age-dependent effects of the 34 chosen SNPs in Table 6.

3.3. Iterative CC-SIS

As is known, the proposed CC-SIS, which selects the variables all at once, suffers from two main problems (Fan and Lv, 2008). First, as the screening procedures are marginal utilities, CC-SIS likely fails to identify these important predictors who are marginally irrelevant to the response, but contribute to the response jointly with other variables. Second, some unimportant predictors may possess high marginal correlation with the response, and hence be falsely included in the screened submodel, merely because they are highly correlated with certain true predictors. To address these issues, we propose an iterative conditioning-correlation sure independence screening (ICC-SIS) for varying coefficient models. The ICC-SIS for choosing d predictors comprises the following steps:

1. Apply CC-SIS to each column of X, where X is the $n \times p$ matrix containing all the candidate covariates. Select d_1 predictors with the highest d_1 $\widehat{\rho}_j^*$ values, denoted by $\mathcal{X}_1 = \{x_{1_1}, \ldots, x_{1_{d_1}}\}$, where $d_1 \leq d$.

Table 6: Information of the significant SNPs

No.	Chromosome	SNP Name	Position	Effect
SNP_1	1	ss66379476	181239647	Additive
SNP_2	1	ss66516012	198313489	Additive
SNP_3	2	ss66282476	47658001	Additive
SNP_4	2	ss66085516	10151206	Dominant
SNP_5	3	ss66266272	29713029	Dominant
SNP_6	4	ss66346937	92071818	Additive
SNP_7	4	ss66137328	94451805	Additive
SNP_8	4	ss66159949	105978188	Additive
SNP_9	4	ss66353634	15504889	Dominant
SNP_{10}	4	ss66354801	115353605	Dominant
SNP_{11}	5	ss66078741	34192815	Dominant
SNP_{12}	5	ss66164865	99237174	Dominant
SNP_{13}	7	ss66524659	44465215	Additive
SNP_{14}	7	ss66155306	134464951	Additive
SNP_{15}	7	ss66261449	46646583	Dominant
SNP_{16}	8	ss66236850	32389924	Additive
SNP_{17}	8	ss66143305	122601829	Additive
SNP_{18}	8	ss66445258	14959676	Dominant
SNP_{19}	8	ss66517429	15818735	Dominant
SNP_{20}	9	ss66319388	16387155	Additive
SNP_{21}	11	ss66153510	13262887	Additive
SNP_{22}	11	ss66110771	13267430	Additive
SNP_{23}	11	ss66112931	103378701	Dominant
SNP_{24}	12	ss66470239	51255904	Additive
SNP_{25}	12	ss66323107	117659019	Additive
SNP_{26}	13	ss66041456	107914455	Dominant
SNP_{27}	14	ss66404926	24422783	Additive
SNP_{28}	15	ss66058021	44940166	Dominant
SNP_{29}	16	ss66064472	5022290	Dominant
SNP_{30}	19	ss66435333	5178008	Dominant
SNP_{31}	20	ss66272727	2700340	Additive
SNP_{32}	20	ss66176990	2723332	Dominant
SNP_{33}	21	ss66511535	15841940	Additive
SNP_{34}	22	ss66305798	16578327	Additive

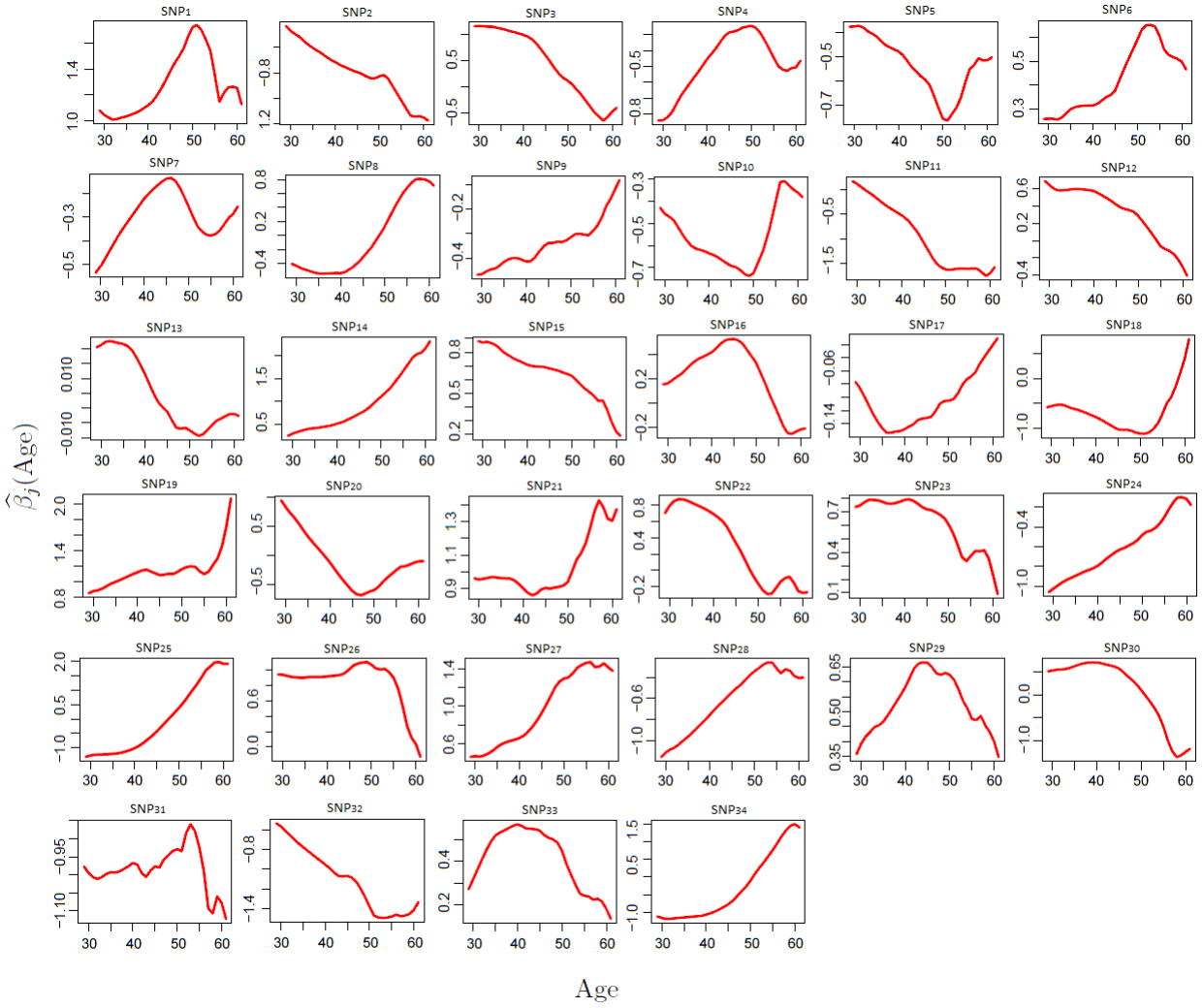

Figure 1: The estimated coefficient functions of significant SNPs

2. Denote $X_s = (x_{1_1}, \ldots, x_{1_{d_1}})$ to be the $n \times d_1$ matrix of selected predictors, and X_r to be the complement of X_s with dimension $n \times (p - d_1)$. Compute the projection of X_r onto the orthogonal complement space of X_s by $X_{proj} = (I_n - X_s(X_s^T X_s)^{-1} X_s^T) X_r$.

3. Apply CC-SIS to each column of X_{proj}, and select d_2 predictors $\mathcal{X}_2 = \{x_{2_1}, \ldots, x_{2_{d_2}}\}$ in the same fashion as step 1, where $d_1 + d_2 \leq d$.

4. Repeat step 2 and 3 until the kth step where $d_1 + d_2 + \ldots + d_k \geq d$. And the selected predictors are $\mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \ldots \bigcup \mathcal{X}_k$.

In the algorithm of ICC-SIS, $d_1, \ldots, d_k$ are chosen by users according to the desired computational complexity. Two steps are often sufficient in practice to achieve satisfactory result: The marginally important predictors are selected in the first step, and the jointly important but marginally uncorrelated predictors are identified afterwards. In addition, if $d_1 = d$, ICC-SIS becomes CC-SIS. We next examine the performance of the proposed iterative procedure.

Table 7: The quantiles of M for Example 3.

	5%	25%	50%	75%	95%
CC-SIS	5.00	17.00	68.50	226.00	654.10
ICC-SIS	5.95	11.00	11.00	11.00	17.00

Table 8: The selecting rates p_a and p_j's for Example 3.

	CC-SIS						ICC-SIS					
d	p_1	p_2	p_3	p_4	p_5	p_a	p_1	p_2	p_3	p_4	p_5	p_a
16	1.000	1.000	0.244	1.000	1.000	0.244	1.000	1.000	1.000	0.990	1.000	0.990
32	1.000	1.000	0.362	1.000	1.000	0.362	1.000	1.000	1.000	0.998	1.000	0.998
48	1.000	1.000	0.431	1.000	1.000	0.431	1.000	1.000	1.000	1.000	1.000	1.000

Example 3. This example demonstrates the advantage of ICC-SIS over CC-SIS when some covariates are jointly active in the presence of other covariates but are marginally uncorrelated with the response. We define the true model index set $\mathcal{M}_* = \{1, 2, 3, 4, 5\}$, and the nonzero coefficient functions as follows:

$$\beta_1(u) = 2 + \cos\left\{\frac{\pi(6u-5)}{3}\right\}, \quad \beta_2(u) = 3 - 3u, \quad \beta_3(u) = -2 + \frac{(2-3u)^3}{4}$$
$$\beta_4(u) = \sin(\frac{9u^2}{2}) + 1, \quad \beta_5(u) = \exp\{3u^2/(3u^2+1)\}.$$

Moreover, the correlation ρ in the covariance matrix of $\mathbf{x}$ is taken to be 0.4. Under this model setting, $\widehat{\rho}_3^*$ is approximately 0, but x_3 is still jointly correlated with y according to the

construction of $\beta_3(u)$. Table 7 and 8 compare the performances of CC-SIS and two-step ICC-SIS. From the tables one can see that ICC-SIS is able to select x_3 which is easily overlooked by CC-SIS. The rankings of $\widehat{\rho}_j^*$'s are not reported because in each iteration of ICC-SIS, the $\widehat{\rho}_j^*$'s of the remaining predictors will change after the previously chosen predictors are removed from the X matrix.

4. Summary

In this paper we proposed a feature screening procedure CC-SIS specifically for ultrahigh dimensional varying coefficient models. The screening criterion $\widehat{\rho}^*$ was constructed based on the conditional correlation which can be estimated by the kernel smoothing technique. We systematically studied the ranking consistency and sure screening property of CC-SIS, and conducted several numerical examples to verify them empirically. The Monte Carlo simulations also showed that CC-SIS can indeed be improved by the iterative algorithm ICC-SIS under certain situations. Furthermore, a two-stage approach, based on CC-SIS and modified penalized regressions, was developed to estimate sparse varying coefficient model with high dimensional covariates.

Appendices

Appendix A: Some Technical Lemmas

In this section, we introduce the following lemmas which are used repeatedly in the proofs of Theorems 1 and 2.

Lemma 1. For any random variable X, the following two statements are equivalent:

(A) *There exists* $H > 0$ *such that* $Ee^{tX} < \infty$ *for all* $|t| < H$.

(B) *There exist* $r > 0$ *and* $T > 0$ *such that* $Ee^{s(X-EX)} \leq e^{rs^2}$.

Proof. It is easy to show (A) under Condition (B). Thus, we focus on showing (B) under Condition (A), which implies that $\mathrm{Var}(X)$ is finite. By Taylor's expansion, for a small $s > 0$, we have

$$Ee^{s(X-EX)} = 1 + E\{s(X - EX)\} + \frac{1}{2}E\{s^2(X - EX)^2\} + o(s^2) = 1 + \frac{1}{2}s^2\mathrm{Var}(X) + o(s^2),$$

and hence

$$\log Ee^{s(X-EX)} = \log\{1 + \frac{1}{2}s^2\mathrm{Var}(X) + o(s^2)\} \leq \frac{1}{2}s^2\mathrm{Var}(X) + o(s^2) \leq s^2\mathrm{Var}(X).$$

The first inequality above is because $\log(1+x) < x$ for all $x > -1$. Therefore, (B) is derived by taking $r \geq \mathrm{Var}(X)$. □

Lemma 2. Suppose that X is a random variable with $E(e^{a|X|}) < \infty$ for some $a > 0$. Then for any $M > 0$, there exist positive constant b and c such that

$$P(|X| \geq M) \leq be^{-cM}.$$

Proof. For any nondecreasing and nonnegative function $g(x)$ and any real number x,

$$\begin{aligned} P(X \geq x) &\leq P\{g(X) \geq g(x)\} = E\{I(g(X) \geq g(x))\} \\ &\leq E\left\{\frac{g(X)}{g(x)} \cdot I(g(X) \geq g(x))\right\} \\ &= \frac{1}{g(x)}E\{g(X) \cdot I(g(X) \geq g(x))\} \\ &\leq \frac{Eg(X)}{g(x)}. \end{aligned}$$

Take $g(x) = e^{ax}$, then we have

$$P(|X| \geq M) \leq P(X \geq M) + P(-X \geq M) \leq \frac{Ee^{aX}}{e^{aM}} + \frac{Ee^{-aX}}{e^{aM}} = be^{-cM},$$

where $b > 0$ such that $Ee^{a|X|} \leq b/2$, and $c = a$. □

Lemma 2 is used to control the tail distribution of x_j and y.

Lemma 3.(Hoeffding's inequality) Suppose that an independent random sample $\{X_i,\ i = 1,\ldots,n\}$ satisfies $P(X_i \in [a_i, b_i]) = 1$ for some a_i and b_i, for all $i = 1,\ldots,n$. Then, for any $\varepsilon > 0$, we have

$$P(|\overline{X} - E(\overline{X})| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \tag{A.1}$$

where $\overline{X} = (X_1 + \cdots + X_n)/n$.

Lemma 4. Suppose that $a(u)$ and $b(u)$ are two uniformly-bounded functions of u. That is, there exist $M_5 > 0$, $M_6 > 0$ such that

$$\sup_{u\in\mathbb{U}}|a(u)| \leq M_5, \quad \sup_{u\in\mathbb{U}}|b(u)| \leq M_6.$$

For a given $u \in \mathbb{U}$, $\hat{A}(u)$ and $\hat{B}(u)$ are estimates of $a(u)$ and $b(u)$ based on a sample with size n. Suppose for any small $\varepsilon \in (0,1)$, there exist positive constants c_1, c_2 and s, such that

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P(|\hat{A}(u) - a(u)| \geq \varepsilon) \leq c_1(1 - \frac{\varepsilon s}{c_1})^n, \\
&\sup_{u\in\mathbb{U}} P(|\hat{B}(u) - b(u)| \geq \varepsilon) \leq c_2(1 - \frac{\varepsilon s}{c_2})^n.
\end{aligned} \tag{A.2}$$

Then

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u) - a(u)b(u)| \geq \varepsilon\right) \leq C_1(1 - \frac{\varepsilon s}{C_1})^n, \\
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)^2 - a(u)^2| \geq \varepsilon\right) \leq C_2(1 - \frac{\varepsilon s}{C_2})^n,
\end{aligned}$$

$$\sup_{u\in\mathbb{U}} P\left(|\{\hat{A}(u) - \hat{B}(u)\} - \{a(u) - b(u)\}| \geq \varepsilon\right) \leq C_3(1 - \frac{\varepsilon s}{C_3})^n,$$

where $C_1 = \max\{2c_1 + c_2, c_1 + 2c_2 + 2c_2M_5, 2c_1M_6\}$, $C_2 = \max\{3c_1 + 2c_1M_5, 2c_1M_6\}$, and $C_3 = \max\{2c_1, 2c_2, c_1 + c_2\}$. Moreover, suppose $b(u)$ is uniformly bounded away from 0 (i.e.,

there is $M_7 > 0$ such that $\inf_{u\in\mathbb{U}} |b(u)| > M_7$), then

$$\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)/\hat{B}(u) - a(u)/b(u)| \geq \varepsilon\right) \leq C_4(1 - \frac{\varepsilon s}{C_4})^n,$$

where $C_4 = \max\{c_1 + c_2 + c_5,\ 2c_1/M_8, 2c_2M_5/(M_7M_8)\}$ with some positive constants c_5 and M_8 defined in the proof. If we further assume $b(u) \geq 0$, then

$$\sup_{u\in\mathbb{U}} P\left(\left|\sqrt{\hat{B}(u)} - \sqrt{b(u)}\right| \geq \varepsilon\right) \leq C_5(1 - \frac{\varepsilon s}{C_5})^n,$$

where $C_5 = \max\{c_2 + c_5, c_2/(\sqrt{M_7} + \sqrt{M_8})\}$.

Proof. For any $\varepsilon \in (0, 1)$, since $\sup_{u\in\mathbb{U}} |a(u)| \leq M_5$,

$$\begin{aligned}
\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)| \geq M_5 + \varepsilon\right) &\leq \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u) - a(u)| + |a(u)| \geq M_5 + \varepsilon\right) \\
&\leq \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u) - a(u)| \geq \varepsilon\right) \\
&\leq c_1(1 - \frac{\varepsilon s}{c_1})^n
\end{aligned}$$

by (A.2). Similarly, it can be proven that $\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)| \geq M_6 + \varepsilon\right) \leq c_2(1 - \varepsilon s/c_2)^n$. Thus, $\hat{A}(u)$ and $\hat{B}(u)$ are bounded in probability.

Consider $\hat{A}(u)\hat{B}(u)$. For any $\varepsilon \in (0, 1)$,

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u) - a(u)b(u)| \geq \varepsilon\right) \qquad \text{(A.3)} \\
= &\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u) - \hat{A}(u)b(u) + \hat{A}(u)b(u) - a(u)b(u)| \geq \varepsilon\right) \\
\leq &\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)| \cdot |\hat{B}(u) - b(u)| + |b(u)| \cdot |\hat{A}(u) - a(u)| \geq \varepsilon\right) \\
\leq &\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)| \cdot |\hat{B}(u) - b(u)| \geq \frac{\varepsilon}{2}\right) + \sup_{u\in\mathbb{U}} P\left(|b(u)| \cdot |\hat{A}(u) - a(u)| \geq \frac{\varepsilon}{2}\right),
\end{aligned}$$

where the first term

$$\begin{aligned}
& \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right) \\
= & \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2},\ |\hat{A}(u)|\geq M_5+\varepsilon\right) \\
& +\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2},\ |\hat{A}(u)|< M_5+\varepsilon\right) \\
\leq & \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)|\geq M_5+\varepsilon\right)+\sup_{u\in\mathbb{U}} P\left((M_5+\varepsilon)\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right) \\
\leq & c_1(1-\frac{\varepsilon s}{c_1})^n+c_2(1-\frac{\varepsilon s}{2c_2(M_5+1)})^n\leq c_3(1-\frac{\varepsilon s}{c_3})^n,
\end{aligned}$$

where $c_3=\max\{c_1+c_2,\ 2c_2(M_5+1)\}$. We next deal with the second term.

$$\sup_{u\in\mathbb{U}} P\left(|b(u)|\cdot|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2}\right)\leq\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2M_6}\right)\leq c_1(1-\frac{\varepsilon s}{2c_1M_6})^n.$$

Therefore, (A.3) becomes

$$\begin{aligned}
& \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u)-a(u)b(u)|\geq\varepsilon\right) \\
\leq & c_3(1-\frac{\varepsilon s}{c_3})^n+c_1(1-\frac{\varepsilon s}{2c_1M_6})^n\leq C_1(1-\frac{\varepsilon s}{C_1})^n, \qquad \text{(A.4)}
\end{aligned}$$

where $C_1=\max\{c_3+c_1,2c_1M_6\}=\max\{2c_1+c_2,c_1+2c_2+2c_2M_5,2c_1M_6\}$. In addition, set $\hat{B}(u)\equiv\hat{A}(u)$, $b(u)\equiv a(u)$, and hence $c_1=c_2$. Let $C_2=\max\{3c_1+2c_1M_5,2c_1M_6\}$, then (A.4) indicates $\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)^2-a(u)^2|\geq\varepsilon\right)\leq C_2(1-\varepsilon s/C_2)^n$.

We next consider $\hat{A}(u)-\hat{B}(u)$. Denote $C_3=\max\{2c_1,\ 2c_2,\ c_1+c_2\}$. For any $\varepsilon\in(0,1)$,

$$\begin{aligned}
& \sup_{u\in\mathbb{U}} P\left(|\{\hat{A}(u)-\hat{B}(u)\}-\{a(u)-b(u)\}|\geq\varepsilon\right) \\
\leq & \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2}\right)+\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right) \\
\leq & c_1(1-\frac{\varepsilon s}{2c_1})^n+c_2(1-\frac{\varepsilon s}{2c_2})^n\leq C_3(1-\frac{\varepsilon s}{C_3})^n.
\end{aligned}$$

We now Consider $\hat{A}(u)/\hat{B}(u)$. First show that $\hat{B}(u)$ is uniformly bounded away from

0 with probability tending to 1. Since $\inf_{u\in\mathbb{U}}|b(u)| > M_7 > 0$, there exists some constant $\delta_0 \in (0,1)$ such that $M_8 \equiv M_7 - \delta_0 > 0$. Then

$$\begin{aligned}
\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)| \le M_8\right) &\le \sup_{u\in\mathbb{U}} P\left(|b(u)| - |\hat{B}(u) - b(u)| \le M_7 - \delta_0\right) \\
&\le \sup_{u\in\mathbb{U}} P\left(|\hat{B}(u) - b(u)| \ge \delta_0\right) \\
&\le c_2'(1 - \frac{\delta_0 s'}{c_2'})^n
\end{aligned}$$

for some positive constants c_2' and s' by (A.2). Take $c_4 = sc_2'/(\delta_0 s')$, then for $\varepsilon \in (0,1)$,

$$\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)| \le M_8\right) \le c_2'(1 - \frac{\delta_0 s'}{c_2'})^n \le c_2'(1 - \frac{s}{c_4})^n \le c_2'(1 - \frac{\varepsilon s}{c_4})^n \le c_5(1 - \frac{\varepsilon s}{c_5})^n,$$

where $c_5 = \max\{c_2', c_4\}$. Hence,

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon\right) \\
= &\sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| \le M_8\right) + \sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| > M_8\right) \\
\le &\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)| \le M_8\right) + \sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| > M_8\right) \\
\le &\; c_5(1 - \frac{\varepsilon s}{c_5})^n + \sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| > M_8\right). \qquad \text{(A.5)}
\end{aligned}$$

The second term of (A.5) is

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| > M_8\right) \\
\le &\sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{\hat{B}(u)}\right| + \left|\frac{a(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \ge \varepsilon, |\hat{B}(u)| > M_8\right) \\
\le &\sup_{u\in\mathbb{U}} P\left(\frac{1}{M_8}\cdot|\hat{A}(u) - a(u)| \ge \frac{\varepsilon}{2}\right) + \sup_{u\in\mathbb{U}} P\left(\frac{|a(u)|}{|\hat{B}(u)|\cdot|b(u)|}\cdot|\hat{B}(u) - b(u)| \ge \frac{\varepsilon}{2}\right) \\
\le &\; c_1(1 - \frac{\varepsilon s M_8}{2c_1})^n + c_2(1 - \frac{\varepsilon s M_7 M_8}{2c_2 M_5})^n. \qquad \text{(A.6)}
\end{aligned}$$

Thus (A.5) is simplified as

$$\begin{aligned}\sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)}-\frac{a(u)}{b(u)}\right|\geq\varepsilon\right) &\leq c_5(1-\frac{\varepsilon s}{c_5})^n+c_1(1-\frac{\varepsilon s M_8}{2c_1})^n+c_2(1-\frac{\varepsilon s M_7 M_8}{2c_2 M_5})^n\\ &\leq C_4(1-\frac{\varepsilon s}{C_4})^n,\end{aligned}$$

where $C_4=\max\{c_1+c_2+c_5,\ 2c_1/M_8, 2c_2M_5/(M_7M_8)\}$.

At last, consider $\sqrt{\hat{B}(u)}$ if well defined. In the similar fashion to (A.5) and (A.6),

$$\begin{aligned}&\sup_{u\in\mathbb{U}} P\left(\left|\sqrt{\hat{B}(u)}-\sqrt{b(u)}\right|\geq\varepsilon\right)\\ \leq\ &\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)|\leq M_8\right)+\sup_{u\in\mathbb{U}} P\left(\frac{|\hat{B}(u)-b(u)|}{\sqrt{\hat{B}(u)}+\sqrt{b(u)}}\geq\varepsilon, |\hat{B}(u)|>M_8\right)\\ \leq\ &c_5(1-\frac{\varepsilon s}{c_5})^n+\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)-b(u)|\geq\varepsilon(\sqrt{M_7}+\sqrt{M_8})\right)\\ \leq\ &c_5(1-\frac{\varepsilon s}{c_5})^n+c_2(1-\frac{\varepsilon s(\sqrt{M_7}+\sqrt{M_8})}{c_2})^n\leq C_5(1-\frac{\varepsilon s}{C_5})^n,\end{aligned}$$

where $C_5=\max\{c_2+c_5, c_2/(\sqrt{M_7}+\sqrt{M_8})\}$. □

Lemma 4 is used in the proof of the ranking consistency property. Now we propose Lemma 5 below to facilitate the proof of the sure screening property.

Lemma 5. Under the same conditions as Lemma 4, suppose that for any $\varepsilon\in(0,1)$,

$$\begin{aligned}&\sup_{u\in\mathbb{U}} P(|\hat{A}(u)-a(u)|\geq\varepsilon)\leq c_1\exp(-\frac{\varepsilon}{h}),\\ &\sup_{u\in\mathbb{U}} P(|\hat{B}(u)-b(u)|\geq\varepsilon)\leq c_2\exp(-\frac{\varepsilon}{h}).\end{aligned}$$

Then

$$\begin{aligned}
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u) - a(u)b(u)| \geq \varepsilon\right) \leq C_6 \exp(-\frac{\varepsilon}{C_6 h}),\\
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)^2 - a(u)^2| \geq \varepsilon\right) \leq C_7 \exp(-\frac{\varepsilon}{C_7 h}),\\
&\sup_{u\in\mathbb{U}} P\left(|\{\hat{A}(u) - \hat{B}(u)\} - \{a(u) - b(u)\}| \geq \varepsilon\right) \leq C_8 \exp(-\frac{\varepsilon}{C_8 h}),\\
&\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)/\hat{B}(u) - a(u)/b(u)| \geq \varepsilon\right) \leq C_9 \exp(-\frac{\varepsilon}{C_9 h)},\\
&\sup_{u\in\mathbb{U}} P\left(|\sqrt{\hat{B}(u)} - \sqrt{b(u)}| \geq \varepsilon\right) \leq C_{10} \exp(-\frac{\varepsilon}{C_{10} h}),
\end{aligned}$$

where $C_6 = \max\{2c_1+c_2, 2M_5+2, 2M_6\}$, $C_7 = \max\{3c_1, 2M_5+2, 2M_6\}$, $C_8 = \max\{c_1+c_2, 2\}$, $C_9 = \max\{c_1+c_2+c_7, 2/M_8, 2M_5/(M_7M_8)\}$, and $C_{10} = \max\{c_2+c_7, 1/(\sqrt{M_7}+\sqrt{M_8})\}$, with some positive constant c_7 defined in the proof.

Proof. For any $\varepsilon \in (0,1)$,

$$\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)| \geq M_5 + \varepsilon\right) \leq c_1 \exp(-\frac{\varepsilon}{h}), \quad \sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)| \geq M_6 + \varepsilon\right) \leq c_2 \exp(-\frac{\varepsilon}{h}).$$

Therefore, similar to the proof of Lemma 4,

$$\begin{aligned}
& \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)\hat{B}(u) - a(u)b(u)| \geq \varepsilon\right)\\
\leq\ & \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)| \geq M_5 + \varepsilon\right) + \sup_{u\in\mathbb{U}} P\left((M_5+\varepsilon)\cdot|\hat{B}(u) - b(u)| \geq \frac{\varepsilon}{2}\right)\\
& + \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u) - a(u)| \geq \frac{\varepsilon}{2M_6}\right)\\
\leq\ & c_1 \exp(-\frac{\varepsilon}{h}) + c_2 \exp(-\frac{\varepsilon}{2h(M_5+1)}) + c_1 \exp(-\frac{\varepsilon}{2hM_6})\\
\leq\ & C_6 \exp(-\frac{\varepsilon}{C_6 h}),
\end{aligned}$$

where $C_6 = \max\{2c_1 + c_2, 2M_5 + 2, 2M_6\}$. Let $C_7 = \max\{3c_1, 2M_5 + 2, 2M_6\}$, it follows that $\sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)^2 - a(u)^2| \geq \varepsilon\right) \leq C_7 \exp\{-\varepsilon/(C_7 h)\}$, by setting $\hat{B}(u) \equiv \hat{A}(u)$, $b(u) \equiv a(u)$,

and $c_1 = c_2$. In addition,

$$\begin{aligned} & \sup_{u\in\mathbb{U}} P\left(|\{\hat{A}(u)-\hat{B}(u)\}-\{a(u)-b(u)\}|\geq\varepsilon\right) \\ \leq\ & \sup_{u\in\mathbb{U}} P\left(|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2}\right)+\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right) \\ \leq\ & c_1\exp(-\frac{\varepsilon}{2h})+c_2\exp(-\frac{\varepsilon}{2h})\leq C_8\exp(-\frac{\varepsilon}{C_8 h}), \end{aligned}$$

where $C_8 = \max\{c_1+c_2, 2\}$. Now consider $\hat{A}(u)/\hat{B}(u)$. Similar as (A.5), for $\varepsilon\in(0,1)$, there exists $c_2'' > 0$ such that

$$\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)|\leq M_8\right)\leq c_2''\exp(-\frac{\delta_0}{h})\leq c_2''\exp(-\frac{\varepsilon}{c_6 h})\leq c_7\exp(-\frac{\varepsilon}{c_7 h}),$$

where $c_6 = 1/\delta_0$ and $c_7 = \max\{c_2'', c_6\}$. Then use the same technique as (A.6),

$$\begin{aligned} & \sup_{u\in\mathbb{U}} P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)}-\frac{a(u)}{b(u)}\right|\geq\varepsilon\right) \\ \leq\ & c_7\exp(-\frac{\varepsilon}{c_7 h})+c_1\exp(-\frac{\varepsilon M_8}{2h})+c_2\exp(-\frac{\varepsilon M_7 M_8}{2hM_5}) \\ \leq\ & C_9\exp(-\frac{\varepsilon}{C_9 h}), \end{aligned}$$

where $C_9 = \max\{c_1+c_2+c_7, 2/M_8, 2M_5/(M_7M_8)\}$. At last, if $\sqrt{\hat{B}(u)}$ is well defined,

$$\begin{aligned} & \sup_{u\in\mathbb{U}} P\left(\left|\sqrt{\hat{B}(u)}-\sqrt{b(u)}\right|\geq\varepsilon\right) \\ \leq\ & \sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)|\leq M_8\right)+\sup_{u\in\mathbb{U}} P\left(|\hat{B}(u)-b(u)|\geq\varepsilon(\sqrt{M_7}+\sqrt{M_8})\right) \\ \leq\ & c_7\exp(-\frac{\varepsilon}{c_7 h})+c_2\exp(-\frac{\varepsilon(\sqrt{M_7}+\sqrt{M_8}))}{h})\leq C_{10}\exp\{-\frac{\varepsilon}{C_{10}h}\}, \end{aligned}$$

where $C_{10} = \max\{c_2+c_7, 1/(\sqrt{M_7}+\sqrt{M_8})\}$. □

Appendix B: Proof of Theorem 1 and 2

For both proofs in Appendix B, we denote C as a generic constant depending on the context, which can vary from line to line.

Proof of Theorem 1. The proof consists of three steps.

Step 1. Prove for any $\varepsilon \in (0,1)$ and $1 \le j \le p$, we can find some positive constants C and s such that

$$\sup_{u\in[a,b]} P(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \ge \varepsilon) \le C(1 - \frac{s\varepsilon}{C})^n. \tag{B.1}$$

Define

$$Z_1(u) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{u_i - u}{h}), \quad Z_2(u) = \frac{1}{n}\sum_{i=1}^{n} y_i K(\frac{u_i - u}{h}),$$
$$Z_3(u) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} K(\frac{u_i - u}{h}), \quad Z_4(u) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 K(\frac{u_i - u}{h}),$$
$$Z_5(u) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 K(\frac{u_i - u}{h}), \quad Z_6(u) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} y_i K(\frac{u_i - u}{h}).$$

Thus, $\hat{\rho}(x_j, y|u)$ can be written as

$$\hat{\rho}(x_j, y|u) = \frac{Z_1(u)Z_6(u) - Z_2(u)Z_3(u)}{\sqrt{\{Z_1(u)Z_4(u) - Z_3(u)^2\}\{Z_1(u)Z_5(u) - Z_2(u)^2\}}}.$$

To show (B.1), we first show $\sup_{u\in[a,b]} P(|Z_2(u) - hf(u)m(u)| \ge \varepsilon) \le 4(1 - s\varepsilon/4)^n$, where $f(u)$ is the density function of u, and $m(u) = E(y|u)$. Since y_i's are not necessarily bounded, we facilitate the proof by truncating y_i's. For any $M > 0$,

$$\begin{aligned}
& \sup_{u\in[a,b]} P(|Z_2(u) - hf(u)m(u)| \ge \varepsilon) \\
= & \sup_{u\in[a,b]} P(|Z_2(u) - hf(u)m(u)| \ge \varepsilon, \max|y_i| \le M) \\
& + \sup_{u\in[a,b]} P(|Z_2(u) - hf(u)m(u)| \ge \varepsilon, \max|y_i| \ge M) \\
\le & \sup_{u\in[a,b]} P\left(|Z_2(u) - hf(u)m(u)| \ge \varepsilon, \max|y_i| \le M\right) \\
& + P(|y_i| \ge M \text{ for some } i) \\
\le & \sup_{u\in[a,b]} P_1(u) + \sup_{u\in[a,b]} P_2(u) + P_3,
\end{aligned} \tag{B.2}$$

where

$$P_1(u) = P(Z_2(u) - hf(u)m(u) \geq \varepsilon, \max|y_i| \leq M),$$
$$P_2(u) = P(Z_2(u) - hf(u)m(u) \leq -\varepsilon, \max|y_i| \leq M),$$
$$P_3 = P(|y_i| \geq M \text{ for some } i) \leq nP(|y| \geq M).$$

First consider $P_1(u)$. The following arguments are all under the condition $\max|y_i| \leq M$, which is omitted for notation simplicity. For any $t > 0$, by Markov's Inequality,

$$\begin{aligned} P_1(u) &\leq P\left(\exp\{t(Z_2(u) - hf(u)m(u))\} \geq \exp(t\varepsilon)\right) \\ &\leq E[\exp\{tZ_2(u) - thf(u)m(u)\}]/\exp(t\varepsilon) \\ &= \exp(-t\varepsilon) \cdot \exp\{-thf(u)m(u)\} \cdot E\{\exp(tZ_2(u))\}, \end{aligned} \tag{B.3}$$

where

$$\begin{aligned} E\{\exp(tZ_2(u))\} &= E\left[\exp\left\{t \cdot \frac{1}{n}\sum_{i=1}^{n} y_i K(\frac{u_i - u}{h})\right\}\right] \\ &= E\left[\prod_{i=1}^{n} \exp\left\{\frac{t}{n} y_i K(\frac{u_i - u}{h})\right\}\right] \\ &= \left[E\left\{\exp\left(\frac{t}{n} y_i K(\frac{u_i - u}{h})\right)\right\}\right]^n. \end{aligned}$$

Set the arbitrary positive number t to be $t = ns$ with positive constant s to be specified later, and define $\varphi(s) = E\{\exp(sy_iK(\frac{u_i-u}{h}))\}$. Then (B.3) becomes

$$P_1(u) \leq \left[\exp(-s\varepsilon) \cdot \exp\{-shf(u)m(u)\} \cdot \varphi(s)\right]^n. \tag{B.4}$$

Now we deal with the last two terms of (B.4):

$$\exp\{-shf(u)m(u)\}\varphi(s) = E\left[\exp\left\{s\left(y_iK(\frac{u_i - u}{h}) - hf(u)m(u)\right)\right\}\right] \equiv I_1(u)I_2(u), \tag{B.5}$$

where

$$\begin{aligned} I_1(u) &= \exp\left\{ s\left(E\{y_i K(\frac{u_i - u}{h})\} - hf(u)m(u) \right) \right\} \\ I_2(u) &= E\left[\exp\left\{ s\left(y_i K(\frac{u_i - u}{h}) - E\{y_i K(\frac{u_i - u}{h})\} \right) \right\} \right] \end{aligned}$$

First consider $I_1(u)$. By Taylor's expansion, for x close to 0, we have

$$\exp(x) = 1 + x + o(|x|) \leq 1 + x + |x| \leq 1 + 2|x|. \tag{B.6}$$

Under the uniformly bounded conditions $\max |y_i| \leq M$, (C1), (C2), and (C4), choose $s > 0$ small enough so that (B.6) can be applied to $I_1(u)$ for any given $u \in [a, b]$:

$$I_1(u) \leq 1 + 2s \left| E\left\{ y_i K(\frac{u_i - u}{h}) \right\} - hf(u)m(u) \right|.$$

Denote $\Delta(u, h) = E\left\{ y_i K((u_i - u)/h) \right\} - hf(u)m(u)$. Note that $E(y_i|u_i) = m(u_i)$. Thus, $\Delta(u, h) = E\left\{ m(u_1) K((u_1 - u)/h) \right\} - hf(u)m(u)$. Note that

$$h^{-1}\Delta(u, h) = \int \{m(u + th)f(u + th) - f(u)m(u)\} K(t)\, dt.$$

Since $K(t)$ is symmetric, $\int tK(t)\, dt = 0$. Moreover,

$$\begin{aligned} & \lim_{h \to 0} h^{-2}[m(u + th)f(u + th) - f(u)m(u) - \{m'(u)f(u) + m(u)f'(u)\}th] \\ \to\ & \{m''(u) + 2m'(u)f'(u) + f''(u)\}t^2. \end{aligned}$$

Thus, it follows by using the dominated convergence theorem and $m''(u)+2m'(u)f'(u)+f''(u)$ being uniformly bounded by Condition (C1) and (C4) that $h^{-3}\Delta(u, h)$ is uniformly bounded by some constant C_0 for $u \in [a, b]$. This implies that

$$\sup_{u \in [a,b]} I_1(u) \leq 1 + sCh^3, \quad \text{as} \quad h \to 0$$

by setting $C = 2C_0$. Since $h \to 0$ as $n \to \infty$, it follows that $\sup_{u \in [a,b]} I_1(u) < 1 + \varepsilon s/16$ for

large enough n.

Furthermore, according to Lemma 1, $I_2(u)$ is uniformly bounded by $\exp(rs^2)$ for some constant $r > 0$, and (B.6) guarantees that $\exp(rs^2) \leq 1 + 2rs^2 < 1 + \varepsilon s/16$, as long as $0 < s < \varepsilon/(32r)$. That is, $\sup_{u\in[a,b]} I_2(u) < 1 + \varepsilon s/16$. Thus, for sufficiently small $s > 0$ and large n, (B.5) satisfies

$$\sup_{u\in[a,b]} \exp\{-shf(u)m(u)\} \cdot \varphi(s) \leq \sup_{u\in[a,b]} I_1(u) \cdot \sup_{u\in[a,b]} I_2(u) < (1 + \varepsilon s/16)^2 < 1 + \varepsilon s/4,$$

and (B.4) is simplified using Taylor's expansion again:

$$\sup_{u\in[a,b]} P_1(u) \leq \{\exp(-\varepsilon s)(1 + \varepsilon s/4)\}^n \leq \{(1 - \varepsilon s + \varepsilon s/2)(1 + \varepsilon s/4)\}^n \leq (1 - \varepsilon s/4)^n$$

Similarly, $\sup_{u\in[a,b]} P_2(u) \leq (1 - s\varepsilon/4)^n$.

Now deal with P_3. According to condition (C3) and Lemma 2, there exist some positive constants m_1 and m_2 such that for any $M > 0$, $P(|y| \geq M) \leq m_1 \exp(-m_2 M)$. Hence, $P_3 \leq nm_1 \exp(-m_2 M) \leq 2(1 - \varepsilon s/4)^n$, by taking $M = m_2^{-1} \log(nm_1/\{2(1 - \varepsilon s/4)^n\})$.

Therefore, (B.2) is simplified as

$$\sup_{u\in[a,b]} P(|Z_2(u) - hf(u)m(u)| \geq \varepsilon) \leq \sup_{u\in[a,b]} P_1(u) + \sup_{u\in[a,b]} P_2(u) + P_3 \leq 4(1 - s\varepsilon/4)^n.$$

All the other desired inequalities are obtained in the same fashion:

$$\sup_{u\in[a,b]} P(|Z_q(u) - hf(u)m(u)| \geq \varepsilon) \leq 4(1 - s\varepsilon/4)^n, q = 1, 3, 4, 5, 6,$$

by defining $m(u) = 1$, $E(x_j|u)$, $E(x_j^2|u)$, $E(y^2|u)$ and $E(x_j y|u)$, respectively. Furthermore, by Lemma 4, there exists some $C > 0$ such that

$$\sup_{u\in[a,b]} P\left(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \geq \varepsilon\right) \leq C \cdot (1 - \frac{s\varepsilon}{C})^n.$$

Step 2. For any $\varepsilon \in (0,1)$, derive the upper bound of $\max_{1\le j\le p} P(|\widehat{\rho}_j^* - \rho_{j0}^*| \ge \varepsilon)$. Notice that

$$\begin{aligned} P(|\widehat{\rho}_j^* - \rho_{j0}^*| \ge \varepsilon) &\le P(|\widehat{\rho}_j^* - \rho_j^*| + |\rho_j^* - \rho_{j0}^*| \ge \varepsilon) \\ &\le P(|\widehat{\rho}_j^* - \rho_j^*| \ge \varepsilon/2) + P(|\rho_j^* - \rho_{j0}^*| \ge \varepsilon/2). \end{aligned} \tag{B.7}$$

The first term of (B.7)

$$\begin{aligned} P\left(|\widehat{\rho}_j^* - \rho_j^*| \ge \varepsilon/2\right) &= P\left(\left|\frac{1}{n}\sum_{i=1}^n \hat{\rho}^2(x_j, y|u_i) - \frac{1}{n}\sum_{i=1}^n \rho^2(x_j, y|u_i)\right| \ge \varepsilon/2\right) \\ &\le P\left(\frac{1}{n}\sum_{i=1}^n \left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \ge \varepsilon/2\right) \\ &= P\left(\sum_{i=1}^n \left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \ge n\varepsilon/2\right) \\ &\le \sum_{i=1}^n P\left(\left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \ge \varepsilon/2\right) \\ &\le n \sup_{u\in[a,b]} P\left(\left|\hat{\rho}^2(x_j, y|u) - \rho^2(x_j, y|u)\right| \ge \varepsilon/2\right) \\ &\le nC(1 - \frac{s\varepsilon}{C})^n. \end{aligned} \tag{B.8}$$

The last inequality in (B.8) is indicated by Step 1 and Lemma 4. And the second term of (B.7) is bounded by $2\exp(-n\varepsilon^2/8)$ based on Lemma 3. Thus

$$\max_{1\le j\le p} P(|\widehat{\rho}_j^* - \rho_{j0}^*| \ge \varepsilon) \le nC(1 - \frac{s\varepsilon}{C})^n + 2\exp(-\frac{n\varepsilon^2}{8}).$$

Step 3. Prove $P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*} \widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}_j^*\} > 0\right) = 1$. Under condition (2.7),

there exists some $\delta > 0$ such that $\min_{j\in\mathcal{M}_*} \rho^*_{j0} - \max_{j\in\mathcal{M}_*^c} \rho^*_{j0} = \delta$. Then we have

$$\begin{aligned}
& P\left(\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j \le \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j\right) = P\left(\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \min_{j\in\mathcal{M}_*} \rho^*_{j0} + \delta \le \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \rho^*_{j0}\right) \\
= \;& P\left(\{\max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \rho^*_{j0}\} - \{\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \min_{j\in\mathcal{M}_*} \rho^*_{j0}\} \ge \delta\right) \\
\le \;& P\left(\max_{j\in\mathcal{M}_*^c} |\widehat{\rho}^*_j - \rho^*_{j0}| + \max_{j\in\mathcal{M}_*} |\widehat{\rho}^*_j - \rho^*_{j0}| \ge \delta\right) \\
\le \;& P\left(2\max_{1\le j\le p} |\widehat{\rho}^*_j - \rho^*_{j0}| \ge \delta\right) = P\left(\max_{1\le j\le p} |\widehat{\rho}^*_j - \rho^*_{j0}| \ge \delta/2\right) \le \sum_{j=1}^{p} P\left(|\widehat{\rho}^*_j - \rho^*_{j0}| \ge \delta/2\right) \\
\le \;& pnC(1 - \frac{\delta s}{2C})^n + 2p\exp(-\frac{n\delta^2}{32}).
\end{aligned}$$

The last inequality is the direct result from Step 2, and it goes to 0 as $n \to \infty$, for $p = o(\exp(an))$ where $a < \min\{\log(2C/(2C - \delta s)), \delta^2/32\}$. Then by Fatou's Lemma,

$$P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j\} \le 0\right) \le \lim_{n\to\infty} P\left(\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j \le 0\right) = 0.$$

In other words,

$$P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}^*_j\} > 0\right) = 1.$$

Therefore, the ranking consistency is proved. □

We next establish the sure screening property of CC-SIS. To begin with, we need to redefine the chosen set $\widehat{\mathcal{M}}$ based on an explicit cutoff $c_3 n^{-\kappa}$, where $0 \le \kappa < \gamma$, i.e.

$$\widehat{\mathcal{M}} = \{j : \; \widehat{\rho}^*_j \ge c_3 n^{-\kappa}, \; 1 \le j \le p\}. \tag{B.9}$$

Proof of Theorem 2. The proof consists of three steps.

Step 1. Prove for any $\varepsilon \in (0, 1)$ and $1 \le j \le p$, we can find some positive constant C such that

$$\sup_{u\in[a,b]} P(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \ge \varepsilon) \le C\exp(-\frac{\varepsilon}{Ch}).$$

As in Theorem 1, we first show $\sup_{u\in[a,b]} P(|Z_2(u)-hf(u)m(u)|\geq\varepsilon)\leq C\exp(-\varepsilon/h)$. Notice that

$$\begin{aligned}&\sup_{u\in\mathbb{U}} P(|Z_2(u)-hf(u)m(u)|\geq\varepsilon)\\ \leq\ &\sup_{u\in\mathbb{U}} P\left(|Z_2(u)-EZ_2(u)|+|EZ_2(u)-hf(u)m(u)|\geq\varepsilon\right).\end{aligned}\tag{B.10}$$

The bias term in (B.10) satisfies $\sup_{u\in[a,b]}|EZ_2(u)-hf(u)m(u)|\leq Ch^3<\varepsilon/2$ for any fixed $\varepsilon\in(0,1)$ and large n by the same techniques as that for $I_1(u)$. Consequently, (B.10) becomes

$$\begin{aligned}\sup_{u\in[a,b]} P(|Z_2(u)-hf(u)m(u)|\geq\varepsilon)\ &\leq\ \sup_{u\in[a,b]} P\left(|Z_2(u)-E(Z_2(u))|\geq\varepsilon/2\right)\\ &\leq\ \sup_{u\in[a,b]} P_4(u)+P_3,\end{aligned}\tag{B.11}$$

where

$$P_4(u)=P(|Z_2(u)-E(Z_2(u))|\geq\varepsilon/2,\max|y_i|\leq M)\leq 2\exp(-\frac{n\varepsilon^2}{8M^2M_4^2})$$

based on Lemma 3, and $P_3=nP(|y|\geq M)\leq nm_1\exp(-m_2M)$ is identical with that in the proof of Theorem 1. Therefore,

$$\begin{aligned}&\sup_{u\in[a,b]} P(|Z_2(u)-hf(u)m(u)|\geq\varepsilon)\leq 2\exp(-\frac{n\varepsilon^2}{8M^2M_4^2})+nm_1\exp(-m_2M)\\ =\ &2\exp(-\frac{\varepsilon}{h}\cdot\frac{n\varepsilon h}{8M^2M_4^2})+m_1\exp(-\frac{\varepsilon}{h}\cdot\frac{h(Mm_2-\log n)}{\varepsilon})\end{aligned}\tag{B.12}$$

Since $h=O(n^{-\gamma})$, we take $M=O(n^{\tau})$, where $\gamma<\tau<(1-\gamma)/2$, then for large n,

$$\frac{n\varepsilon h}{8M^2M_4^2}=Cn^{1-\gamma-2\tau}>1,\quad\text{and}\quad\frac{h(Mm_4-\log n)}{\varepsilon}=C(n^{\tau-\gamma}-n^{-\gamma}\log n)>1,$$

thus (B.12) is simplified as

$$\sup_{u\in[a,b]} P(|Z_2(u)-hf(u)m(u)|\geq\varepsilon)\leq C\exp(-\frac{\varepsilon}{h}).$$

By Lemma 5 and the same technique as the proof of Theorem 1, there exists some $C>0$

such that

$$\sup_{u\in[a,b]} P\left(|\hat{\rho}(x_j,y|u)-\rho(x_j,y|u)|\geq\varepsilon\right)\leq C\exp(-\frac{\varepsilon}{Ch}).$$

Step 2. To show that $P\left(\max_{1\leq j\leq p}|\widehat{\rho}_j^*-\rho_{j0}^*|>c_3n^{-\kappa}\right)\leq O\{np\exp(-n^{\gamma-\kappa}/\xi)\}$. Similar to step 2 of the proof of Theorem 1,

$$\begin{aligned}\max_{1\leq j\leq p} P(|\widehat{\rho}_j^*-\rho_{j0}^*|\geq c_3n^{-\kappa}) &\leq P(|\widehat{\rho}_j^*-\rho_j^*|\geq c_3n^{-\kappa}/2)+P(|\rho_j^*-\rho_{j0}^*|\geq c_3n^{-\kappa}/2)\\ &\leq nC\exp(-\frac{c_3n^{-\kappa}}{2Ch})+2\exp(-c_3^2\frac{n^{1-2\kappa}}{8})\\ &= O\{n\exp(-n^{\gamma-\kappa}/\xi)\} \qquad \text{(B.13)}\end{aligned}$$

where ξ is a positive constant, and $0\leq\kappa<\gamma$. The last equation is because the first term dominates the second when $h=O(n^{-\gamma})$. Hence,

$$P\left(\max_{1\leq j\leq p}|\widehat{\rho}_j^*-\rho_{j0}^*|\geq c_3n^{-\kappa}\right)\leq\sum_{j=1}^{p}P(|\widehat{\rho}_j^*-\rho_{j0}^*|\geq c_3n^{-\kappa})\leq O\{np\exp(-n^{\gamma-\kappa}/\xi)\}.$$

This completes the proof of the first part of Theorem 2.

Step 3: Furthermore under condition (2.8), prove

$$P(\mathcal{M}_*\subset\widehat{\mathcal{M}})\geq 1-O\{ns_n\exp(-n^{\gamma-\kappa}/\xi)\}.$$

According to the definition of $\widehat{\mathcal{M}}$ in (B.9) and condition (2.8),

$$\begin{aligned}&P(\mathcal{M}_*\subset\widehat{\mathcal{M}})=P\left(\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*\geq c_3n^{-\kappa}\right)=P\left(\min_{j\in\mathcal{M}_*}\rho_{j0}^*-\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*\leq\min_{j\in\mathcal{M}_*}\rho_{j0}^*-c_3n^{-\kappa}\right)\\ \geq\ & P\left(\min_{j\in\mathcal{M}_*}\rho_{j0}^*-\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*\leq 2c_3n^{-\kappa}-c_3n^{-\kappa}\right)\geq P\left(\max_{j\in\mathcal{M}_*}|\rho_{j0}^*-\widehat{\rho}_j^*|\leq c_3n^{-\kappa}\right)\\ =\ & 1-P\left(\max_{j\in\mathcal{M}_*}|\rho_{j0}^*-\widehat{\rho}_j^*|\geq c_3n^{-\kappa}\right)\geq 1-s_n\max_{1\leq j\leq p}P\left(|\rho_{j0}^*-\widehat{\rho}_j^*|\geq c_3n^{-\kappa}\right)\\ \geq\ & 1-O\{ns_n\exp(-n^{\gamma-\kappa}/\xi)\}.\end{aligned}$$

The last inequality is due to (B.13). Consequently, this establishes the sure screening prop-

erty of CC-SIS. □

References

Dawber, T. R., Meadors, G. F. and Moore, F. E., Jr. (1951). "Epidemiological Approaches to Heart Disease: the Framingham Study," *American Journal of Public Health*, **41**, 279–286.

Fan, J., Feng, Y. and Song, R. (2011), "Nonparametric independence screening in sparse ultra-high dimensional additive models," *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, New York, NY.

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

Fan, J., and Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, **99**, 710–723.

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

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

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized likelihood ratio statistics and Wilks phenomenon," *The Annals of Statistics*, **29**, 153–193.

Hall, P. and Miller, H. (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, **18**, 533–550.

He, X., Wang, L. and Hong, H. G. (2013). "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *The Annals of Statistics*, **41**, 342–369.

Jaquish, C. (2007). "The Framingham Heart Study, on Its Way to Becoming the Gold Standard for Cardiovascular Genetic Epidemiology," *BMC Medical Genetics*, **8**, 63.

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

Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Model," *The Annals of Statistics*, **36**, 261–286.

Li, R., Zhong, W. and Zhu, L.P. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, **107**, 1129–1139.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, **90**, 1257–1270.

Tibshirani, R. (1996), "Regression shrinkage and selection via LASSO," *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Wang, H. and Xia. Y. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, **104**, 747–757.

Wang, L., Li, H. and Huang, J. (2008), "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, **103**, 1556–1569.

Zhu, L.P., Li, L., Li, R., and Zhu, L.X. (2011), "Model-free feature screening for ultrahigh dimensional data," *Journal of the American Statistical Association*, **106**, 1464–1475.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, **101**, 1418–1429.