
Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

# Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates

Jingyuan Liu, Runze Li & Rongling Wu

Published online: 19 Mar 2014.

To cite this article: Jingyuan Liu, Runze Li & Rongling Wu (2014) Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates, Journal of the American Statistical Association, 109:505, 266-274, DOI: 10.1080/01621459.2013.850086

To link to this article: http://dx.doi.org/10.1080/01621459.2013.850086

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

# Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates

Jingyuan Liu, Runze Li, and Rongling Wu

This article 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. Supplementary materials for this article are available online.

KEY WORDS: Conditional correlation; Ranking consistency; Sure screening property; Ultrahigh dimensionality; Varying coefficient models.



## 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 3.2 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 the varying coefficient model for this analysis. There are thousands of single-nucleotide polymorphisms (SNPs) available in the FHS database, leading to the ultrahigh dimensionality. While only hundreds of samples are available, as is typical 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 smoothly clipped absolute deviation (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 integrating local polynomial regression techniques (Fan and Gijbels 1996) and least absolute shrinkage and selection operator (LASSO; Tibshirani 1996). Nevertheless, these variable selection procedures were developed for the varying coefficient models with fixed-dimensional covariates. As a result, they 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 ultrahigh 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 an optimal transformation function. Fan and Song (2010) modified SIS for 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 et al. (2011) proposed a sure independence ranking and screening (SIRS) procedure to select important predictors under the multi-index model setting. Li et al. (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 (DC), which is directly applicable for multiple response and grouped

Jingyuan Liu is Assistant Professor, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, China (E-mail: *jingyuan1230@gmail.com*). Runze Li is Distinguished Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rzli@psu.edu*). Rongling Wu is Professor, Department of Public Health Sciences, Penn State Hershey College of Medicine, Hershey, PA 17033 (E-mail: *RWu@phs.psu.edu*). The research of Runze Li was supported by National Institute on Drug Abuse (NIDA) grant P50-DA10075, National Cancer Institute (NCI) grant R01 CA168676, and National Natural Science Foundation of China grant 11028103. The research of Rongling Wu was supported by an NSF grant IOS-0923975 and an NIH grant UL1RR0330184. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, NIDA, and NCI.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.



DOI: 10.1080/01621459.2013.850086

predictors. He, Wang, and Hong (2013) proposed a quantile-adaptive model-free feature screening procedure for heterogenous data. Our article aims to develop a kernel regression-based screening method specifically for ultrahigh-dimensional varying coefficient models to reduce dimensionality.

Suppose that the varying coefficients in the varying coefficient models are functions 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 article, 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 article makes the following theoretical contributions 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 article 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 addition, 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. We also propose an iterative screening procedure in Section 3. Conclusion remark is given in Section 4, and the technical proofs are given in the online supplement.

## 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 univariate variable $u$.

Note that given $u$, the varying coefficient model becomes a linear regression model. Fan and Lv (2008) proposed an SIS procedure for linear regression models 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 $u$. Define the marginal utility for feature screening as

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

To estimate $\rho^*_{j0}$, 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 article, it is assumed that these five conditional means are nonparametric smooth functions of $u$. Therefore, the conditional correlation in (2.2) can be estimated through nonparametric regression 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

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

where $K_h(t) = h^{-1}K(t/h)$. Similarly, we may define kernel regression estimates $\widehat{E}(y^2|u)$, $\widehat{E}(x_j|u)$, $\widehat{E}(x_j^2|u)$, and $\widehat{E}(x_j y|u)$ for $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$, and $E(x_j y|u)$, respectively. The conditional covariance $\mathrm{cov}(x_j, y|u)$ can be estimated by $\widehat{\mathrm{cov}}(x_j, y|u) = \widehat{E}(x_j y|u) - \widehat{E}(x_j|u)\widehat{E}(y|u)$, 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 1.* We employ the kernel regression rather than local linear regression because local linear regression estimates 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 to guarantee that $|\hat{\rho}(x_j, y|u)| \leq 1$. In our numerical studies, we first select an optimal bandwidth for $E(x_j y|u)$ by using a plug-in method (Ruppert, Sheather, and Wang 1995), and then use this bandwidth for other four conditional means. We empirically study the impact of bandwidth selection on the performance of the proposed screening procedure in Section 3.1. For our simulation study, the proposed procedure performs quite well provided that the bandwidth lies within an appropriate range.

The plug-in estimate of $\rho^*_{j0}$ 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 define 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 $\nu$. 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 is denoted by $\mathbb{U} = [a, b]$ with finite constants $a$ and $b$. Define the true model index set $\mathcal{M}_*$ with cardinality $p_*$ and its complement $\mathcal{M}_*^c$ by

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

Denote the truly important predictor vector by $\mathbf{x}_{\mathcal{M}_*}$, a vector consisting of $x_j$ with $j \in \mathcal{M}_*$. That is, if $\mathcal{M}_* = \{j_1, \ldots, j_{p_*}\}$, then $\mathbf{x}_{\mathcal{M}_*} = (x_{j_1}, \ldots, x_{j_{p_*}})^{\mathrm{T}}$. Similarly, define $\mathbf{x}_{\mathcal{M}_*^c}$, $\boldsymbol{\beta}_{\mathcal{M}_*}(u)$, and $\boldsymbol{\beta}_{\mathcal{M}_*^c}(u)$. Furthermore, define $\boldsymbol{\rho}_{\mathcal{M}_*}(u)$ to be a vector consisting of $\rho(x_j, y|u)$ with $j \in \mathcal{M}_*$. Denote by $\lambda_{\max}\{A\}$ and $\lambda_{\min}\{A\}$ the largest and smallest eigenvalues of the matrix $A$, respectively, and "$a_n > b_n$ uniformly in $n$" means "$\liminf_{n\to\infty}\{a_n - b_n\} > 0$." Furthermore, denote $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}^{\mathrm{T}}$ for notation simplicity.

The following two conditions are needed for Theorem 1, which characterizes the relationship of $\rho_{j0}^*$ between the truly important and unimportant predictors.

(B1) The following inequality holds uniformly in $n$:

$$\min_{j\in\mathcal{M}_*} \rho_{j0}^* > E\big\{\big(\lambda_{\max}\big\{\mathrm{cov}\big(\mathbf{x}_{\mathcal{M}_*}, \mathbf{x}_{\mathcal{M}_*^c}^{\mathrm{T}}|u\big)\mathrm{cov}\big(\mathbf{x}_{\mathcal{M}_*^c}, \mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}|u\big)\big\} \times \lambda_{\max}\big\{\boldsymbol{\rho}_{\mathcal{M}_*}^{\otimes 2}(u)\big\}\big)\big/\big(\lambda_{\min}^2\{\mathrm{cov}(\mathbf{x}_{\mathcal{M}_*}|u)\}\big)\big\}. \quad (2.7)$$

(B2) Assume that conditioning on $\mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{M}_*}(u)$ and $u$, $\mathbf{x}$ and $\varepsilon$ are independent. Further assume that the following linearity condition is valid:

$$E\big\{\mathbf{x}|\mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{M}_*}(u), u\big\} = \mathrm{cov}\big(\mathbf{x}, \mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}|u\big)\boldsymbol{\beta}_{\mathcal{M}_*}(u)\big\{\mathrm{cov}\big(\mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{M}_*}(u)|u\big)\big\}^{-1} \times \boldsymbol{\beta}_{\mathcal{M}_*}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{M}_*}. \quad (2.8)$$

Conditions (B1) and (B2) are adapted from Zhu et al. (2011). (B1) requires that the population level unconditioned-squared correlation cannot be too small. The first assumption in (B2) implies that $y$ depends on $\mathbf{x}$ through $\mathbf{x}_{\mathcal{M}_*}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{M}_*}(u)$, and (2.8) refers to the conditional linearity condition.

*Theorem 1.* Under conditions (B1) and (B2), it follows that

$$\liminf_{n\to\infty}\Big\{\min_{j\in\mathcal{M}_*}\rho_{j0}^* - \max_{j\in\mathcal{M}_*^c}\rho_{j0}^*\Big\} > 0. \quad (2.9)$$

The proof of Theorem 1 is similar to Theorem 1 of Zhu et al. (2011) and therefore is given in the supplementary material of this article.

The inequality (2.9) provides a clear separation between the important and unimportant predictors in terms of $\rho_{j0}^*$. It rules out the situation when certain unimportant predictors have large $\rho_{j0}^*$'s and are selected only because they are highly correlated with the true ones. This is a necessary condition for the ranking consistency property established below.

The following regularity conditions are used to establish the ranking consistency property and sure screening property of the CC-SIS.

(C1) Denote the density function of $u$ by $f(u)$. Assume that $f(u)$ has continuous second-order derivative on $\mathbb{U}$.

(C2) The kernel $K(\cdot)$ is a symmetric density function with finite support and is bounded uniformly over its support.

(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 $0 \le s < s_0$,

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

(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

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

Conditions (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, that is, $E(|K(u)|^r) < \infty$, for any $r > 0$. Although the Gaussian kernel function does not satisfy (C2), it can be shown that results in Theorems 2 and 3 are still valid for the Gaussian kernel function. Condition (C3) is relatively strong and only used to facilitate the technical proofs. Condition (C4) requires the mean-related quantities bounded and the variances positive, to guarantee that the conditional correlation is well defined. We first establish the ranking consistency property of CC-SIS.

*Theorem 2. (Ranking Consistency Property).* Under conditions (B1), (B2), (C1)–(C4), suppose the bandwidth $h \to 0$ but $nh^3 \to \infty$ as $n \to \infty$. Then for $p = o\{\exp(an)\}$ with some $a > 0$, we have

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

The proof of Theorem 2 is given in the online supplement of this article. Theorem 2 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 3. (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

$$\min_{j\in\mathcal{M}_*}\rho_{j0}^* \ge 2c_3 n^{-\kappa}, \tag{2.10}$$

then

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

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

The proof of Theorem 3 is given in the online supplement of this article. Condition (2.10) 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 $\rho_{j0}^*$'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 3 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. The kernel function is taken to be $K(u) = 0.75(1-u^2)_+$ in all the numerical studies.

For each simulation example (i.e., Examples 1 and 3), 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 $\Sigma$ is a $(p+1)\times(p+1)$ covariance matrix with element $\sigma_{ij} = \rho^{|i-j|}$, $i,j = 1,\ldots,p+1$. 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 error $\varepsilon$ 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 instance, $R_j$ for SIS is the average rank of the Pearson correlation between $x_j$ and $y$ in the decreasing order; $R_j$ for CC-SIS 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 submodels $\widehat{\mathcal{M}}$ with size $d$ that contain all the true predictors among 1000 simulations.
- $p_j$: The proportion of submodels $\widehat{\mathcal{M}}$ with size $d$ that contain $x_j$ among 1000 simulations.

The criteria are used to empirically verify the theoretical properties in Theorems 2 and 3. The ranking consistency of a screening procedure refers to the property that the screening scores of the 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 $p_a$ and $p_j$'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 the 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 model settings. In Case I, the nonzero coefficient functions are truly varying over $u$; 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 1:* The nonzero coefficient functions are defined by

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

*Case 2:* The nonzero coefficient functions are defined by

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

First consider Case I, in which data were generated from a varying coefficient model. 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 $\rho = 0.8$. However, the SIS 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 misspecify the varying coefficient model as a linear regression model and apply SIS, the constant coefficient $\beta_{600}$ 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

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

| Method | Low correlation: $\rho = 0.4$ | | | | | High correlation: $\rho = 0.8$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R_2$ | $R_{100}$ | $R_{400}$ | $R_{600}$ | $R_{1000}$ | $R_2$ | $R_{100}$ | $R_{400}$ | $R_{600}$ | $R_{1000}$ |
| | | | | | Case I: varying coefficient model | | | | | |
| SIS | 3.5 | 1.5 | 6.5 | 461.3 | 2.2 | 7.9 | 1.8 | 14.0 | 468.1 | 3.1 |
| SIRS | 3.7 | 1.6 | 10.7 | 486.8 | 2.1 | 9.4 | 1.8 | 15.7 | 454.5 | 3.2 |
| DC-SIS | 3.1 | 1.6 | 10.1 | 350.6 | 2.2 | 6.3 | 1.9 | 12.9 | 341.8 | 3.4 |
| CC-SIS | 2.7 | 2.1 | 3.8 | 3.8 | 3.3 | 6.8 | 2.1 | 6.9 | 5.9 | 3.9 |
| | | | | | Case II: linear model | | | | | |
| SIS | 3.1 | 5.4 | 1.7 | 5.4 | 1.7 | 5.2 | 12.6 | 1.9 | 12.9 | 2.2 |
| SIRS | 3.2 | 6.3 | 1.8 | 5.8 | 1.8 | 5.6 | 14.1 | 2.0 | 14.4 | 2.3 |
| DC-SIS | 3.2 | 6.4 | 1.8 | 6.8 | 1.8 | 5.6 | 13.6 | 2.1 | 15.0 | 2.1 |
| CC-SIS | 3.1 | 6.6 | 1.7 | 7.0 | 1.7 | 9.2 | 12.3 | 1.7 | 12.7 | 1.9 |

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, comparing the models with the two $\rho$'s, those 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

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

| $d$ | Method | Low correlation: $\rho = 0.4$ | | | | | | High correlation: $\rho = 0.8$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $p_2$ | $p_{100}$ | $p_{400}$ | $p_{600}$ | $p_{1000}$ | $p_a$ | $p_2$ | $p_{100}$ | $p_{400}$ | $p_{600}$ | $p_{1000}$ | $p_a$ |
| | | | | | | Case I: varying coefficient model | | | | | | | |
| 16 | SIS | 0.99 | 1.00 | 0.95 | 0.03 | 0.99 | 0.03 | 0.93 | 0.99 | 0.81 | 0.00 | 0.99 | 0.00 |
| | SIRS | 0.99 | 1.00 | 0.90 | 0.01 | 1.00 | 0.01 | 0.90 | 1.00 | 0.76 | 0.00 | 0.99 | 0.00 |
| | DC-SIS | 0.99 | 1.00 | 0.92 | 0.02 | 1.00 | 0.02 | 0.96 | 1.00 | 0.81 | 0.00 | 0.99 | 0.00 |
| | CC-SIS | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.96 | 1.00 | 0.96 | 0.97 | 0.99 | 0.89 |
| 32 | SIS | 1.00 | 1.00 | 0.98 | 0.07 | 1.00 | 0.07 | 0.98 | 1.00 | 0.95 | 0.03 | 1.00 | 0.02 |
| | SIRS | 0.99 | 1.00 | 0.94 | 0.03 | 1.00 | 0.03 | 0.97 | 1.00 | 0.92 | 0.03 | 1.00 | 0.02 |
| | DC-SIS | 1.00 | 1.00 | 0.95 | 0.05 | 1.00 | 0.05 | 0.99 | 1.00 | 0.94 | 0.02 | 1.00 | 0.02 |
| | CC-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| 48 | SIS | 1.00 | 1.00 | 0.99 | 0.09 | 1.00 | 0.09 | 0.99 | 1.00 | 0.97 | 0.05 | 1.00 | 0.05 |
| | SIRS | 1.00 | 1.00 | 0.96 | 0.04 | 1.00 | 0.03 | 0.99 | 1.00 | 0.96 | 0.05 | 1.00 | 0.04 |
| | DC-SIS | 1.00 | 1.00 | 0.97 | 0.08 | 1.00 | 0.08 | 1.00 | 1.00 | 0.97 | 0.05 | 1.00 | 0.05 |
| | CC-SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | | | Case II: linear model | | | | | | | |
| 16 | SIS | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 0.96 | 0.99 | 0.79 | 1.00 | 0.78 | 1.00 | 0.61 |
| | SIRS | 0.99 | 0.96 | 1.00 | 0.96 | 1.00 | 0.93 | 0.96 | 0.75 | 1.00 | 0.74 | 0.99 | 0.54 |
| | DC-SIS | 0.99 | 0.96 | 1.00 | 0.96 | 0.99 | 0.92 | 0.97 | 0.73 | 1.00 | 0.73 | 0.99 | 0.52 |
| | CC-SIS | 1.00 | 0.96 | 1.00 | 0.95 | 1.00 | 0.91 | 0.91 | 0.82 | 1.00 | 0.82 | 1.00 | 0.62 |
| 32 | SIS | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.97 | 1.00 | 0.97 | 1.00 | 0.94 |
| | SIRS | 1.00 | 0.98 | 1.00 | 0.99 | 1.00 | 0.97 | 1.00 | 0.95 | 1.00 | 0.94 | 1.00 | 0.89 |
| | DC-SIS | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 0.97 | 1.00 | 0.96 | 1.00 | 0.94 | 1.00 | 0.90 |
| | CC-SIS | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 0.96 | 0.99 | 0.97 | 1.00 | 0.96 | 1.00 | 0.92 |
| 48 | SIS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.98 |
| | SIRS | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 0.97 | 1.00 | 0.96 |
| | DC-SIS | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 0.97 | 1.00 | 0.96 |
| | CC-SIS | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 | 0.97 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 0.97 |

Table 3. The quantiles of $M$ for Example 1

| | Low correlation: $\rho = 0.4$ | | | | | High correlation: $\rho = 0.8$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | 5% | 25% | 50% | 75% | 95% | 5% | 25% | 50% | 75% | 95% |
| | | | | | Case I: varying coefficient model | | | | | |
| SIS | 26.0 | 185.8 | 454.5 | 728.3 | 951.0 | 52.0 | 228.8 | 442.5 | 701.0 | 951.0 |
| SIRS | 61.0 | 223.8 | 475.0 | 744.3 | 943.0 | 52.0 | 208.0 | 440.0 | 675.0 | 922.2 |
| DC-SIS | 32.0 | 135.0 | 299.5 | 522.5 | 841.1 | 51.0 | 142.0 | 297.0 | 500.3 | 790.1 |
| CC-SIS | 5.0 | 5.0 | 5.0 | 5.0 | 7.0 | 6.0 | 8.0 | 10.0 | 13.0 | 20.0 |
| | | | | | Case II: linear model | | | | | |
| SIS | 5.0 | 5.0 | 5.0 | 7.0 | 14.0 | 8.0 | 11.0 | 15.0 | 20.0 | 35.1 |
| SIRS | 5.0 | 5.0 | 6.0 | 7.0 | 22.0 | 8.0 | 12.0 | 15.0 | 22.0 | 51.0 |
| DC-SIS | 5.0 | 5.0 | 6.0 | 8.0 | 24.0 | 8.0 | 12.0 | 16.0 | 21.0 | 41.1 |
| CC-SIS | 5.0 | 5.0 | 6.0 | 7.0 | 29.1 | 8.0 | 11.0 | 15.0 | 20.0 | 44.0 |

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.

Furthermore, we study the performance of CC-SIS when the kernel is oversmooth with a larger bandwidth $h_L = 1.25h$ and undersmooth with a smaller $h_S = 0.75h$, where $h$ is the optimal bandwidth chosen by the plug-in method introduced in the last section. The average rank, the selecting rate, and the minimum model size are very similar to Tables 1, 2, and 3, thus are omitted to save space. Therefore, CC-SIS is stable with respect to the bandwidth selection provided that the chosen bandwidth lies within the right range.

### 3.2 Two-Stage Approach for Varying Coefficient Models and an Application

Consider the varying coefficient model (3.1). 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 $d = [n^{4/5}/\log(n^{4/5})]$. With slight abuse of notation, 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 $j$th column of $\mathbf{B}$. The estimator $\widehat{\mathbf{B}}_\lambda$ of $\mathbf{B}$ is defined by

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

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

With a chosen $\lambda$, 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 2000):

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

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

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

Table 4. The sizes and MSPE of the nine models

| | CC-SIS+LASSO | CC-SIS+ALASSO | CC-SIS+SCAD |
|---|---|---|---|
| AIC | 43 (0.405)[a] | 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) |

NOTE: [a]The numbers in the parentheses are MSPE of the model.

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

| | | $H_1$ | | | | |
|---|---|---|---|---|---|---|
| | | Unpenalized | LASSO-AIC | LASSO-BIC | ALASSO-AIC | ALASSO-BIC |
| $H_0$ | LASSO-AIC | 0.9952 | – | – | – | – |
| | LASSO-BIC | 0.9999 | 0.9462 | – | – | – |
| | 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 |

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

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

where the matrix $\mathbf{D}^{(m)}$ is a $(d+1)\times(d+1)$ diagonal matrix with the $j$th 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 $\lambda$ 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

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

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

*Example 2*. The FHS is a cardiovascular study beginning in 1948 under the direction of the National Heart, Lung and Blood Institute (Dawber, Meadors, and Moore 1951; Jaquish 2007). In our analysis, 349,985 nonrare SNPs are of interest, and the data from 977 subjects are available. The goal is to detect the SNPs that are important to explaining the 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: Akaike information criterion (AIC), Bayesian information criterion (BIC), and generalized cross-validation (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.

Figure 1 is the plot of the estimated coefficient functions versus age, which depicts the age-dependent effects of the 34 chosen SNPs.

### 3.3 Iterative CC-SIS

Similar to the SIS (Fan and Lv 2008), the proposed CC-SIS has one major weakness. Since the screening procedure is based on a marginal utility, CC-SIS likely fails to identify those important predictors that are marginally irrelevant to the response, but contribute to the response jointly with other variables. To address this issue, 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 and 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 $d_1 \le d$.
2. Denote $\mathbf{X}_s$ to be the $n\times d_1$ matrix of selected predictors, and $\mathbf{X}_r$ to be the complement of $\mathbf{X}_s$ with dimension $n\times(p-d_1)$. For any given $u\in\mathbb{U}$, compute the weighted projection matrix of $\mathbf{X}_r$ by $\mathbf{X}_{\mathrm{proj}}(u) = \mathbf{X}_r - \mathbf{X}_s(\mathbf{X}_s^{\mathrm{T}}\mathbf{W}(u)\mathbf{X}_s)^{-1}\mathbf{X}_s^{\mathrm{T}}\mathbf{W}(u)\mathbf{X}_r$, where $\mathbf{W}(u)$ is the $n\times n$ diagonal weight matrix with the $i$th diagonal element $\omega_i(u) = K_h(u_i-u)/\{\sum_{i=1}^n K_h(u_i-u)\}$. Note that the matrix $\mathbf{X}_{\mathrm{proj}}(u)$ depends on $u$.
3. For each $u_i$, $i = 1, \ldots, n$, and $j = 1, \ldots, p-d_1$, compute the sample conditional correlation $\widehat{\rho}(x_{j,\mathrm{proj}}, y|u_i)$ by (2.3) and (2.4) using the $j$th column of $\mathbf{X}_{\mathrm{proj}}(u_i)$ and $y$. The screening criterion for the $j$th remaining predictor is $\widehat{\rho}_{j,\mathrm{proj}}^* = \frac{1}{n}\sum_{i=1}^n \widehat{\rho}^2(x_{j,\mathrm{proj}}, y|u_i)$. Select $d_2$ predictors $\mathcal{X}_2 = \{x_{2_1}, \ldots, x_{2_{d_2}}\}$ by ranking $\widehat{\rho}_{j,\mathrm{proj}}^*$'s, where $d_1+d_2 \le d$.
4. Repeat Steps 2 and 3 until the $k$th step when $d_1 + d_2 + \cdots + d_k \ge d$. And the selected predictors are $\mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \cdots \bigcup \mathcal{X}_k$.

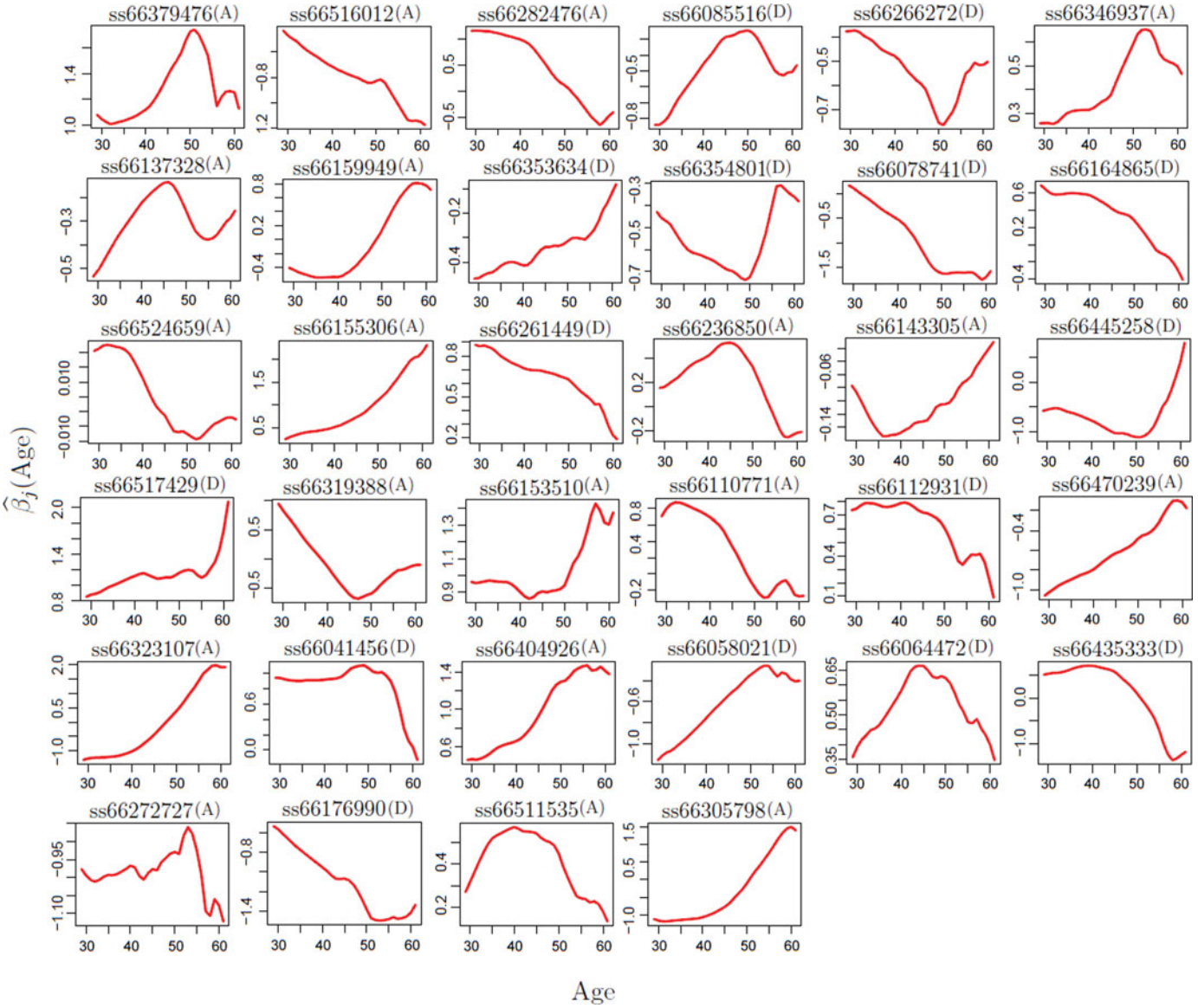


Figure 1. The estimated coefficient functions of the chosen 34 SNPs indexed by the SNP names. The A or D in the parentheses indicates which effect (A-additive, D-dominant) of the chosen SNP is significant.

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 a satisfactory result: the marginally important predictors are selected in the first step, and the jointly important but marginally uncorrelated predictors are identified afterward. In addition, if $d_1 = d$, ICC-SIS becomes CC-SIS. Motivation of the ICC-SIS and some theoretical insights into the ICC-SIS are given in the supplementary material. We next examine the performance of the ICC-SIS by the following simulation example.

Table 6. The quantiles of $M$ for Example 3

| | 5% | 25% | 50% | 75% | 95% |
|---|---|---|---|---|---|
| CC-SIS | 5.0 | 17.0 | 68.5 | 226.0 | 654.1 |
| ICC-SIS | 5.0 | 9.0 | 9.0 | 11.0 | 17.0 |

*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:

$$\beta_1(u) = 2 + \cos\{\pi(6u - 5)/3\}, \quad \beta_2(u) = 3 - 3u,$$
$$\beta_3(u) = -2 + 0.25(2 - 3u)^3$$
$$\beta_4(u) = \sin(9u^2/2) + 1, \quad \beta_5(u) = \exp\{3u^2/(3u^2 + 1)\}.$$

Moreover, the correlation $\rho$ 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 $\beta_3(u)$. Tables 6 and 7 compare the performances of CC-SIS and two-step ICC-SIS. From the tables one can see that ICC-SIS is able to select $x_3$ that 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

Table 7. 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 | 1.00 | 1.00 | 0.24 | 1.00 | 1.00 | 0.24 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.99 |
| 32 | 1.00 | 1.00 | 0.36 | 1.00 | 1.00 | 0.36 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 48 | 1.00 | 1.00 | 0.43 | 1.00 | 1.00 | 0.43 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

after the previously chosen predictors are removed from the X matrix.

## 4. SUMMARY

In this article, 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 that 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 models with high-dimensional covariates.

## SUPPLEMENTARY MATERIALS

The supplementary materials consist of three parts. First, some lemmas used in the proof of Theorem 2 and 3 are presented and proved. The second part consists of the proof of Theorem 1. In the third part, we present the motivation of ICC-SIS as well as a brief theoretical justification of this method.

*[Received March 2013. Revised September 2013.]*

## 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. [272]

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. [266]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, New York: Chapman and Hall. [266,267]

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. [266,271]

——— (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723. [272]

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. [266,267,269,272]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [266]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193. [272]

Fan, J., and Zhang, W. (2000), "Simultaneous Confidence Bands and Hypotheses Testing in Varying-Coefficient Models," *Scandinavian Journal of Statistics*, 27, 1491–1518. [271]

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. [266]

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. [267]

Jaquish, C. (2007), "The Framingham Heart Study, on its Way to Becoming the Gold Standard for Cardiovascular Genetic Epidemiology," *BMC Medical Genetics*, 8, 63. [272]

Li, G., Peng, H., Zhang, J., and Zhu, L.-X. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877 [266]

Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Model," *The Annals of Statistics*, 36, 261–286. [266]

Li, R., Zhong, W., and Zhu, L. P. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [266,269]

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. [267]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [272]

Wang, H., and Xia, Y. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, 104, 747–757. [266,271]

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. [266]

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. [266,267,268,269]

Zou, H. (2006), "The Adaptive LASSO and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [272]