Ann Inst Stat Math (2018) 70:283–301
https://doi.org/10.1007/s10463-016-0597-2

Model-free feature screening for ultrahigh-dimensional data conditional on some variables

Yi Liu[1,2] · **Qihua Wang**[1,3]

Received: 2 June 2016 / Revised: 18 November 2016 / Published online: 17 January 2017

Abstract In this paper, the conditional distance correlation (CDC) is used as a measure of correlation to develop a conditional feature screening procedure given some significant variables for ultrahigh-dimensional data. The proposed procedure is model free and is called conditional distance correlation-sure independence screening (CDC-SIS for short). That is, we do not specify any model structure between the response and the predictors, which is appealing in some practical problems of ultrahigh-dimensional data analysis. The sure screening property of the CDC-SIS is proved and a simulation study was conducted to evaluate the finite sample performances. Real data analysis is used to illustrate the proposed method. The results indicate that CDC-SIS performs well.

Keywords Conditional distance correlation · Feature selection · Sure screening property · High-dimensional data

Electronic supplementary material The online version of this article (doi:10.1007/s10463-016-0597-2) contains supplementary material, which is available to authorized users.

✉ Qihua Wang
qhwang@amss.ac.cn

1 Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

2 College of Science, China University of Petroleum, Qingdao 266580, China

3 Institute of Statistical Science, Shenzhen University, Shenzhen 518006, China

1 Introduction

With the development of modern technology, the collection and storage of ultrahigh-dimensional data become easier in various scientific areas, such as genomics, proteomics, and high-frequency finance, where the number of variables p may grow exponentially with the sample size n. One way to deal with large p is to use variable selection which assumes that only a small number of predictors are rescaled to the response, that is, the sparsity principle. However, the regulation methods may not perform well for ultrahigh-dimensional data, due to simultaneous challenges of computational expediency, statistical accuracy, and algorithm stability (Fan et al. 2009).

To tackle these difficulties, Fan and Lv (2008) proposed a two-stage procedure. First, a fast screening procedure is applied to reduce the ultrahigh dimensionality to a moderate scale that is smaller than or equal to the sample size n; then, regulation method can be used to obtain the final model. Several screening methods have been developed in the recent history. Fan and Lv (2008) introduced a marginal Pearson correlation measure in the linear model. Fan and Song (2010) extended the method to generalized linear model by ranking the maximum marginal likelihood estimates. Furthermore, Fan et al. (2011) explored the feature screening technique for ultrahigh-dimensional additive models. Zhu et al. (2011) proposed a robust correlation measure under the multi-index model. The methods mentioned above are based on model structures, which may cause incorrect results when the models are misspecified. Recently, Li et al. (2012) proposed a model-free feature screening technique based on the distance correlation (DC) studied in Szekely et al. (2007). This measure is robust to model misspecification and can be used for feature screening without specifying a regression model. Zhong et al. (2016) generalized the DC method to a robust one through the distance correlation between the predictors and the marginal distribution of the response.

In some practical problems, however, some predictors are known to be significant to response. A problem is that how to make feature screening for the remaining predictors given the significant predictors. An analogous problem is considered by Liu et al. (2014), where the conditional Pearson correlation coefficient is used as a measure to develop a conditional feature screening for the linear varying coefficient model. That is, given the exposure variables, the conditional Pearson correlation-based feature screening is developed for the predictors in the linear part. Fan et al. (2014) also proposed a screening method in linear varying coefficient models, that cannot adapt to the nonlinear situations yet.

How to construct a model-free conditional measure to screen the predictors is a very important task. This paper uses the conditional distance correlation (CDC) measure due to Wang et al. (2015) to develop conditional feature screening given some significant variables without assuming any model structure. The proposed procedure is referred to as conditional distance correlation sure independence screening (CDC-SIS for short). Wang et al. (2015) showed that the CDC equals zero if and only if two random vectors are independent conditional on some other variables. We systematically study the theoretical properties of the CDC-SIS and prove that with probability tending to one, all active predictors are selected, i.e., the sure screening property proposed in Fan and

Lv (2008) is proved. The finite sample performances of the proposed procedure via numerical simulation are studied.

The rest of this paper is organized as follows. In Sect. 2, we propose a new conditional feature screening procedure for ultrahigh-dimensional data and study its sure screening property. In Sect. 3, a simulation study is conducted to assess the finite sample performances. In Sect. 4, we illustrate the method through a real data example. The regularity conditions and technical proofs are given in Appendix.

2 Independence screening using CDC

2.1 The methodology

Let $Y \in R$ denote the response, W some significant predictor vector of Y, and $\mathbf{X} = (X_1, \ldots, X_p) \in R^p$ the remaining p-dimensional predictors. To highlight our method, we consider univariate W next without loss of generality. We consider the conditional distribution function of Y given $\mathbf{X}$ and W, denoted by $F(y|\mathbf{X}, W) = P(Y \le y|\mathbf{X}, W)$. Let

$$\mathcal{M}_* = \{j : F(y|\mathbf{X}, W) \text{ functionally depends on } X_j\},$$

be the index set of active predictors and it is natural to assume the sparsity, that is, only a small number of predictors in $\mathbf{X}$ are relevant to Y given W. Throughout the paper, we assume the cardinality of $\mathcal{M}_*$, $s_n = |\mathcal{M}_*|$ is smaller than the sample size n. Next, let us introduce the conditional distance correlation suggested by Wang et al. (2015).

For $t \in R$ and $s \in R$, the conditional joint characteristic function of X_j and Y given W is defined as

$$\Phi_{X_j,Y|W}(t, s) = E(e^{itX_j+isY}|W), \quad j = 1, \ldots, p, \tag{1}$$

where i is the imaginary unit. In addition, the conditional marginal characteristic functions of X_j and Y given W are defined as

$$\Phi_{X_j|W}(t) = \Phi_{X_j,Y|W}(t, 0), \quad \text{and} \quad \Phi_{Y|W}(s) = \Phi_{X_j,Y|W}(0, s), \quad j = 1, \ldots, p.$$

Wang et al. (2015) proposed conditional distance correlation for measuring the dependence between two random vectors given another random vector. Specifically, given W, the conditional distance of Y and X_j, $j = 1, \ldots, p$ is defined as

$$D^2(X_j, Y|W) = \int |\Phi_{X_j,Y|W}(t, s) - \Phi_{X_j|W}(t)\Phi_{Y|W}(s)|^2 w(t, s)\mathrm{d}t\mathrm{d}s, \tag{2}$$

where $w(t, s) = 1/(c_1^2\|t\|^2\|s\|^2)$, and $c_d = \pi^{(1+d)/2}/\Gamma((1 + d)/2)$. Throughout the article, $\|\cdot\|$ stands for the Euclidean norm.

The conditional distance variance of X_j and Y given W are, respectively,

$$D^2(X_j|W) = D^2(X_j, X_j|W), \;\; D^2(Y|W) = D^2(Y, Y|W). \tag{3}$$

Then the conditional distance correlation between X_j and Y given W is defined as

$$\rho^2(X_j, Y|W) = \frac{D^2(X_j, Y|W)}{\sqrt{D^2(X_j|W)D^2(Y|W)}}. \tag{4}$$

Define the marginal utility for feature screening as

$$\rho_{j0}^* = E(\rho^2(X_j, Y|W)), \quad j = 1, \dots, p.$$

A remarkable property of the marginal utility ρ_{j0}^* is that $\rho_{j0}^* = 0$ if and only if X_j and Y are independent, conditional on W. This measure allows our method to detect any nonlinear relationship between the response and predictors. This implies that when there is a nonlinear relationship between X_j and Y, ρ_{j0}^* is far away from zero, while the conditional Pearson correlation proposed by Liu et al. (2014) may be very small and even close to zero because that Pearson correlation can only detect the linear relationship between X_j and Y.

Suppose that $\{(\mathbf{X}_i, Y_i, W_i), i = 1, \dots, n\}$ are independent and identically distributed copies of $(\mathbf{X}, Y, W)$, and $\mathbf{X}_i = (X_{1i}, X_{2i}, \dots, X_{pi})$. Denote $d_{kl}^{X_j} = d(X_{jk}, X_{jl})$ as the Euclidean distance of X_{jk} and X_{jl} and, similarly, d_{kl}^Y for Y. Wang et al. (2015) establishes the following expression:

$$D^2(X_j, Y|W = w) = S_1(w) + S_2(w) - 2S_3(w), \tag{5}$$

where $S_j(w)$, $j = 1, 2, 3$ are defined as

$$\begin{aligned}
S_1(w) &= E(d_{kl}^{X_j} d_{kl}^Y | W_k = w, W_l = w),\\
S_2(w) &= E(d_{kl}^{X_j} | W_k = w, W_l = w) E(d_{kl}^Y | W_k = w, W_l = w),\\
S_3(w) &= E(d_{kl}^{X_j} d_{km}^Y | W_k = w, W_l = w, W_m = w).
\end{aligned}$$

To estimate $D^2(X_j, Y|W = w)$, we only need derive the sample estimators of $S_j(w)$, $j = 1, 2, 3$. Clearly, these conditional expectations can be estimated by the kernel smoothing method (Fan and Gijbels 1996). Let $K(\cdot)$ be a given kernel function, h a bandwidth, $a_k(w) = K_h(w - W_k) = K((w - W_k)/h)/h$ and $a(w) = \sum_{k=1}^n a_k(w)$, then the kernel regression estimates are given by

$$\begin{aligned}
\hat{S}_1(w) &= \sum_{k,l=1}^n d_{kl}^{X_j} d_{kl}^Y a_k(w) a_l(w)/a^2(w),\\
\hat{S}_2(w) &= \sum_{k,l=1}^n d_{kl}^{X_j} a_k(w) a_l(w) \sum_{k,l=1}^n d_{kl}^Y a_k(w) a_l(w)/a^4(w),\\
\hat{S}_3(w) &= \sum_{k,l,m=1}^n d_{kl}^{X_j} d_{km}^Y a_k(w) a_l(w) a_m(w)/a^3(w).
\end{aligned}$$

Substituting these estimates into (5), we obtain a nature estimator of $D^2(X_j, Y|W = w)$, denoted by $\hat{D}^2(X_j, Y|W = w) = \hat{S}_1(w) + \hat{S}_2(w) - 2\hat{S}_3(w)$. Similarly, we can define the sample conditional distance variances $\hat{D}^2(X_j|W = w)$ and $\hat{D}^2(Y|W = w)$. Accordingly, the sample conditional distance correlation is given by

$$\hat{\rho}^2(X_j, Y|W = w) = \frac{\hat{D}^2(X_j, Y|W = w)}{\sqrt{\hat{D}^2(X_j|W = w)\hat{D}^2(Y|W = w)}}, \tag{6}$$

which can be seen as a function of w, denoted by $\hat{\rho}_j^2(w)$. We now define an estimate of the marginal utility ρ_{j0}^* as

$$\hat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}_j^2(W_i), \quad j = 1, \ldots, p.$$

Based on $\hat{\rho}_j^*$, we select a set of important predictors with large $\hat{\rho}_j^*$,

$$\hat{\mathcal{M}} = \{j : 1 \le j \le p,\ \hat{\rho}_j^* > cn^{-\kappa}\},$$

where c and κ are prespecified threshold values. However, in practice, we often select the first d largest $\hat{\rho}_j^*$ with d taken to be smaller than the sample size n. Thus, we can reduce the dimensionality of the predictors from p to a moderate scale d. Liu et al. (2014) suggested setting $d = [n^{4/5}/\log(n^{4/5})]$ for ultrahigh-dimensional varying coefficient model, where $[a]$ refers to the integer part of a.

2.2 Sure screening property

We next study the theoretical properties of the proposed screening procedure CDC-SIS.

Theorem 1 *Under regularity conditions given in Appendix, suppose the bandwidth $h = O(n^{-\gamma})$, where $0 < \gamma < 1/2$, $0 \le \kappa < \gamma$, and ξ is a positive constant, then we have*

$$P(\max_{1\le j\le p} |\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa}) \le O(np\exp(-n^{\gamma-\kappa}/\xi)),$$
$$P(\mathcal{M}^* \subset \hat{\mathcal{M}}) \ge 1 - O(ns_n \exp(-n^{\gamma-\kappa}/\xi)).$$

Theorem 1 indicates that we can handle the nonpolynomial (NP) dimensionality of order $\log p = o(n^{\gamma-\kappa})$. In other words, the tail probability in Theorem 1 is exponentially small. Hence, $\hat{\mathcal{M}}$ can retain all important predictors with probability tending to 1. The following corollary gives the sure screening property of the CDC-SIS screening.

Corollary 1 *Under the conditions of Theorem 1, when* $\log p = o(n^{\gamma-\kappa})$*, we have that*

$$P(\max_{1\le j\le p} |\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa}) \to 0, \quad as\ n \to \infty,$$
$$P(\mathcal{M}^* \subset \hat{\mathcal{M}}) \to 1, \quad as\ n \to \infty.$$

CDC-SIS also provides an alternative to the varying coefficient models, although it is suggested in a general case that we do not require specifying the relationship between Y and $\mathbf{X}$ given W.

Remark 1 This proposed method is a nonparametric one and hence depends on the selection of the bandwidth. However, in practice, the bandwidth selection is not so critical since we use $\hat{\rho}_j^*$, which is the average of $\hat{\rho}_j^2(W_i)$ for $i = 1, 2, \ldots, n$ and hence is a global quantity. That is, the method is not sensible to the selection of bandwidth as long as the bandwidth satisfies the condition given for the feature screening property. The explanation is similar to Wang and Rao (2002).

3 Numerical studies

In this section, we conducted some numerical studies to evaluate the proposed method CDC-SIS, and compared it with the conditional Pearson correlation coefficient proposed by Liu et al. (2014) (CC-SIS), the nonparametric independence screening (NIS) method proposed by Fan et al. (2014), the SIS method proposed by Fan and Lv (2008), the DC-SIS method proposed by Li et al. (2012) and DC-RoSIS method proposed by Zhong et al. (2016). The last three methods are developed for unconditional feature screening. We compare our method with them to display the benefit of using prior knowledge of some significant predictor. The kernel function is taken to be $K(w) = 0.75(1 - w^2)_+$ and the bandwidth is taken to be $h = n^{-1/5}$ throughout this paper.

Similar to Liu et al. (2014), the variables $(W^*, \mathbf{X}^\top)^\top$ are generated from $N(0, \Sigma)$, where Σ is a $(p + 1) \times (p + 1)$ covariance matrix with element $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, p + 1$. We consider $\rho = 0.4$ and 0.8, respectively. Then we take $W = \Phi(W^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus, W follows a uniform distribution $U(0, 1)$ and is correlated with $\mathbf{X}$. We take p to be 1000, and the sample size n is 200; the model size d is chosen to be $d_i = i[n^{4/5}/log(n^{4/5})]$, $i = 1, 2, 3$, where $[a]$ denotes the integer part of a. All the simulations are based on 500 replications.

Following Li et al. (2012), we employ S, P_j and P_{All} to assess the performance of the CDC-SIS, where S, P_j and P_{All} are defined as follows:

- S: The minimal model size to include all active predictors. We report the 5, 25, 50, 75, and 95% quantiles of S out of 500 replications.
- P_j: The proportion of the j-th active predictor selected by the submodel $\hat{\mathcal{M}}$ with size d among 500 replications.
- P_{All}: The proportion of all active predictors selected by the submodel $\hat{\mathcal{M}}$ with size d among 500 replications.

Example 1 In this example, we consider the following linear varying coefficient model:

$$
\begin{aligned}
(1.1): Y = {} & \beta_2(W)X_2 + \beta_{100}(W)X_{100} + \beta_{400}(W)X_{400} + \beta_{600}(W)X_{600} \\
& + \beta_{1000}(W)X_{1000} + \epsilon,
\end{aligned}
$$

where the nonzero coefficient functions are defined by

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

We consider two error distributions, a standard norm $N(0, 1)$ and a standard Cauchy distribution which has a heavy tail.

Table 1 reports the quantile of S. It is seen that, when the model is indeed linear with a norm error, CDC-SIS has a comparable performance to CC-SIS and both outperform the unconditional methods SIS, DC-SIS and DC-RoSIS significantly. The NIS method performs a little bit worse. On the other hand, when the error distribution is heavily tailed, our method clearly outperforms the other methods. It is reasonable because the proposed method is model free, while CC-SIS and NIS are developed for linear varying coefficient model and are not robust to models with heavy tail error distribution. The unconditional methods are intuitively inefficient because they do not use the information of the significant(conditional) predictor. Table 2 reports the proportion P_j and P_{All}. All P_j and P_{All} of CDC-SIS are close to 1 as d increases, while the low value of P_{600} and P_{All} of the SIS, DC-SIS and DC-RoSIS imply that they rank X_{600} behind and regard it as an unimportant variable. This may be because that $\beta_{600}(W) = 2\sin(2\pi W)$ has mean 0 if W follows a $U(0, 1)$ distribution. Thus, the screening methods SIS, DC-SIS and DC-RoSIS are not suitable for varying the coefficient model, especially when the coefficient oscillates about zero.

Example 2 Similar to *Example* 1, we set

$$
\begin{aligned}
\beta_1(W) = 2I(W > 0.4), \quad \beta_2(W) = 1 + W, \\
\beta_3(W) = (2 - 3W)^2, \quad \beta_4(W) = 2\sin(2\pi W),
\end{aligned}
$$

and the error ϵ follows a standard normal distribution. The response is generated from the following three models.

$$
\begin{aligned}
&(2.1)\ : Y = \beta_1(W)X_1 + \beta_2(W)X_2 + \beta_3(W)I(X_{12} < 0) + \beta_4(W)X_{22} + \epsilon, \\
&(2.2)\ : Y = \beta_1(W)X_1X_2 + \beta_3(W)I(X_{12} < 0) + \beta_4(W)X_{22} + \epsilon, \\
&(2.3)\ : Y = \beta_1(W)X_1 + \beta_2(W)X_2 + \beta_3(W)I(X_{12} < 0) + \exp(|X_{22}|)\epsilon,
\end{aligned}
$$

where $I(X_{12} < 0)$ is an indicator function.

Models (2.1)–(2.3) are all nonlinear in X_{12}, and model (2.2) contains an interaction term X_1X_2, and model (2.3) is heteroscedastic. However, the CC-SIS and NIS

Table 1 The quantile of S

ϵ	Method	$\rho = 0.4$					$\rho = 0.8$				
		5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$N(0, 1)$	CDC-SIS	5.0000	5.0000	5.0000	5.0000	8.0000	6.0000	8.0000	11.0000	15.0000	24.0000
	CC-SIS	5.0000	5.0000	5.0000	5.0000	7.0000	6.0000	8.0000	11.0000	14.0000	21.0000
	NIS	6.0000	13.0000	39.0000	129.0000	438.0000	10.0000	22.0000	47.0000	127.0000	516.5000
	SIS	30.0000	189.0000	477.0000	795.0000	962.0000	60.0000	306.0000	556.5000	803.0000	968.5000
	DC-SIS	30.0000	134.0000	282.0000	506.5000	884.0000	49.0000	152.5000	316.5000	532.0000	824.0000
	DC-RoSIS	30.5000	160.0000	341.5000	613.0000	901.0000	59.5000	201.0000	391.5000	621.5000	872.5000
Cauchy	CDC-SIS	5.0000	13.0000	46.0000	143.5000	443.5000	11.0000	23.5000	63.0000	162.5000	461.0000
	CC-SIS	52.5000	215.0000	418.5000	673.5000	927.5000	59.5000	260.0000	475.0000	703.0000	922.0000
	NIS	286.0000	559.0000	756.0000	882.0000	979.0000	303.0000	569.0000	742.5000	882.0000	969.5000
	SIS	215.5000	474.0000	691.0000	852.5000	970.5000	207.5000	482.5000	690.0000	861.0000	970.5000
	DC-SIS	78.0000	259.0000	468.5000	723.5000	937.0000	81.5000	237.0000	447.0000	703.5000	918.0000
	DC-RoSIS	48.0000	235.0000	446.5000	740.0000	936.5000	68.0000	238.0000	441.5000	699.5000	901.0000

Table 2 The proportion of P_j and P_{All}

ϵ	Method	Size	$\rho = 0.4$						$\rho = 0.8$					
			P_j					P_{All}	P_j					P_{All}
			X_2	X_{100}	X_{400}	X_{600}	X_{1000}	All	X_2	X_{100}	X_{400}	X_{600}	X_{1000}	All
$N(0, 1)$	CDC-SIS	d_1	1.0000	1.0000	1.0000	0.9900	0.9980	0.9880	0.9200	1.0000	0.9760	0.9580	0.9800	0.8380
		d_2	1.0000	1.0000	1.0000	1.0000	0.9980	0.9980	0.9880	1.0000	0.9960	0.9960	0.9960	0.9760
		d_3	1.0000	1.0000	1.0000	1.0000	0.9980	0.9980	0.9960	1.0000	0.9980	1.0000	0.9980	0.9920
	CC-SIS	d_1	1.0000	1.0000	1.0000	0.9980	1.0000	0.9980	0.9360	1.0000	0.9800	0.9700	0.9840	0.8720
		d_2	1.0000	1.0000	1.0000	0.9980	1.0000	0.9980	0.9900	1.0000	1.0000	1.0000	0.9980	0.9880
		d_3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9960	1.0000	1.0000	1.0000	0.9980	0.9940
	NIS	d_1	0.9780	0.9580	0.3640	0.9760	0.9060	0.2880	0.7540	0.9720	0.2620	0.9840	0.9000	0.1640
		d_2	0.9860	0.9860	0.5120	0.9860	0.9500	0.4640	0.8700	0.9900	0.4660	0.9940	0.9560	0.3680
		d_3	0.9920	0.9880	0.5780	0.9920	0.9740	0.5400	0.9220	0.9960	0.5840	0.9940	0.9740	0.5140
	SIS	d_1	0.9980	1.0000	0.9480	0.0140	1.0000	0.0120	0.9320	1.0000	0.8480	0.0040	0.9980	0.0020
		d_2	1.0000	1.0000	0.9680	0.0560	1.0000	0.0540	0.9880	1.0000	0.9540	0.0200	1.0000	0.0200
		d_3	1.0000	1.0000	0.9760	0.0760	1.0000	0.0720	0.9960	1.0000	0.9800	0.0340	1.0000	0.0340
	DC-SIS	d_1	0.9980	1.0000	0.9240	0.0200	1.0000	0.0200	0.9620	1.0000	0.8240	0.0000	0.9940	0.0000
		d_2	1.0000	1.0000	0.9540	0.0560	1.0000	0.0540	0.9880	1.0000	0.9400	0.0160	0.9980	0.0160
		d_3	1.0000	1.0000	0.9640	0.0940	1.0000	0.0900	0.9900	1.0000	0.9600	0.0560	1.0000	0.0500
	DC-RoSIS	d_1	0.9960	1.0000	0.9080	0.0220	1.0000	0.0200	0.9560	1.0000	0.7960	0.0020	0.9900	0.0020
		d_2	1.0000	1.0000	0.9420	0.0580	1.0000	0.0540	0.9880	1.0000	0.9280	0.0180	0.9980	0.0160
		d_3	1.0000	1.0000	0.9580	0.0740	1.0000	0.0700	0.9880	1.0000	0.9520	0.0380	1.0000	0.0320
Cauchy	CDC-SIS	d_1	0.8180	0.8640	0.7080	0.6020	0.7460	0.2880	0.5320	0.8500	0.6740	0.6020	0.7300	0.1580
		d_2	0.8740	0.9120	0.7720	0.7120	0.8300	0.4300	0.6400	0.9080	0.7840	0.7380	0.8200	0.3320
		d_3	0.9040	0.9300	0.8180	0.7760	0.8720	0.5160	0.7060	0.9380	0.8180	0.7940	0.8680	0.4380

Table 2 continued

ϵ	Method	Size	$\rho = 0.4$						$\rho = 0.8$					
			P_j					P_{All}	P_j					P_{All}
			X_2	X_{100}	X_{400}	X_{600}	X_{1000}	All	X_2	X_{100}	X_{400}	X_{600}	X_{1000}	All
	CC-SIS	d_1	0.3560	0.3820	0.3420	0.2100	0.3040	0.0100	0.2000	0.3600	0.3200	0.2160	0.3020	0.0080
		d_2	0.4300	0.4420	0.3980	0.2800	0.4100	0.0220	0.2740	0.4380	0.3960	0.3000	0.3860	0.0360
		d_3	0.4720	0.5040	0.4560	0.3600	0.4780	0.0440	0.3200	0.5120	0.4360	0.3520	0.4360	0.0420
	NIS	d_1	0.0800	0.0560	0.0340	0.0540	0.0560	0.0000	0.0440	0.0680	0.0260	0.0720	0.0440	0.0020
		d_2	0.1300	0.0920	0.0540	0.0900	0.0860	0.0020	0.0720	0.1220	0.0440	0.1080	0.0780	0.0020
		d_3	0.1620	0.1180	0.0740	0.1200	0.1000	0.0060	0.0900	0.1540	0.0740	0.1340	0.1040	0.0020
	SIS	d_1	0.2080	0.2740	0.1700	0.0140	0.2680	0.0000	0.1640	0.2680	0.1320	0.0120	0.2440	0.0000
		d_2	0.2780	0.3440	0.2360	0.0380	0.3360	0.0020	0.2220	0.3520	0.1880	0.0340	0.3200	0.0020
		d_3	0.3160	0.4000	0.2960	0.0600	0.4020	0.0040	0.2660	0.4080	0.2280	0.0620	0.3680	0.0080
	DC-SIS	d_1	0.6660	0.8220	0.5480	0.0180	0.7820	0.0080	0.6260	0.8420	0.4440	0.0120	0.7600	0.0040
		d_2	0.7340	0.8640	0.6340	0.0320	0.8280	0.0220	0.7220	0.8920	0.5880	0.0240	0.8140	0.0120
		d_3	0.7840	0.8840	0.6920	0.0560	0.8640	0.0320	0.7740	0.9120	0.6320	0.0480	0.8540	0.0240
	DC-RoSIS	d_1	0.8800	0.9800	0.7540	0.0240	0.9660	0.0160	0.8180	0.9800	0.5840	0.0080	0.9180	0.0020
		d_2	0.9260	0.9960	0.8460	0.0320	0.9820	0.0280	0.9060	0.9980	0.7380	0.0160	0.9660	0.0100
		d_3	0.9460	0.9960	0.8760	0.0540	0.9880	0.0520	0.9420	1.0000	0.8100	0.0380	0.9820	0.0300

methods which perform well in linear varying coefficient model are not suitable in these nonlinear cases. The quantile of S is reported in Table 3. We can see that CDC-SIS performs better than the other five screening methods, in particular when models deviate far from the linear model. P_j and P_{All} are reported in Table 4. The performance of CC-SIS is not too bad in model (2.1). P_1, P_2 and P_{22} are all equal to 1, and P_{12} is a little lower, that is because X_1, X_2, X_{22} are the linear parts, and X_{12} is the nonlinear part of the response. However, CC-SIS has little chance to identify the important predictors X_1, X_2 in model (2.2) and X_{12}, X_{22} in model (2.3). NIS has a poor performance, mainly because the predictor X_{12} presents in an index function, and NIS cannot find it out. The other three unconditional methods clearly cannot select all important predictors with the nonlinear and varying coefficient interaction.

In this paper, we only consider univariate W, however, W can be extended to multivariate very directly. In this subsection, we study the applicability of the proposed method in the case of multivariate W.

Example 3 $Y = \beta_1(\mathbf{W}^\top\gamma)X_1X_2+\beta_2(\mathbf{W}^\top\gamma)I(X_3<0)+\epsilon$, with $\mathbf{W} = (W_1, W_2)^\top$ is a two-dimensional index vector, $\gamma = [1, 1]^\top$ is the index coefficient, and $(\mathbf{W}^\top, \mathbf{X}^\top)^\top$ are generated as described before with $\rho = 0.4$. Set

$$\beta_1(u) = \exp(5u)/(1+\exp(5u)), \quad \beta_2(u) = \sin(\pi u),$$

and the error ϵ follows a standard normal distribution.

We compare the performances of different methods in Tables 5 and 6. Similar to Example 2, in terms of the quantile of S, the size of CDC-SIS is much smaller than the others; on the other hand, the proportion P_j and P_{All} are much closer to 1. In summary, CDC-SIS outperforms the other methods in the case of multivariate W setup under consideration.

Based on the referees' suggestions, we further consider a simulation setup which satisfies the assumptions for the NIS. The results are listed in the supplement with the same setting as in Example 3 in Fan et al. (2014). The results show that CDC-SIS, CC-SIS and NIS perform well and behave better than the unconditional screening methods SIS, DC-SIS and DC-RoSIS. NIS behaves comparably to CDC-SIS and CC-SIS according to the top 50% quantiles of S. However, in terms of the 75 and 95% quantiles of S, the NIS method needs a larger model size to include all active predictors than the CDC-SIS and CC-SIS methods. Moreover, P_{All} of NIS is a little lower than those of CDC-SIS and CC-SIS, but all these three methods outperform the unconditional screening methods significantly. For more details, please see the supplemental material.

4 Real data analysis

In this section, we illustrate the performance of our method through a real data analysis on Boston Housing Data (Harrison and Rubinfeld 1978). The sample size $n = 506$ in

Table 3 The quantile of S

Model	Method	$\rho = 0.4$					$\rho = 0.8$				
		5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
(2.1)	CDC-SIS	4.0000	4.0000	7.0000	17.0000	130.5000	7.0000	12.0000	19.0000	55.0000	269.5000
	CC-SIS	4.0000	5.0000	12.0000	37.0000	211.5000	9.0000	17.0000	37.0000	116.5000	426.0000
	NIS	23.5000	135.0000	318.0000	595.0000	875.5000	48.0000	192.0000	438.5000	731.0000	941.5000
	SIS	681.5000	891.0000	969.0000	994.0000	1000.0000	432.5000	714.5000	880.0000	968.0000	997.0000
	DC-SIS	24.0000	96.5000	208.0000	413.5000	765.0000	32.5000	119.0000	240.5000	461.0000	806.0000
	DC-RoSIS	30.0000	123.0000	288.0000	491.5000	816.5000	47.0000	139.5000	284.5000	480.5000	819.0000
(2.2)	CDC-SIS	4.0000	7.0000	22.0000	72.5000	297.5000	6.0000	8.0000	12.0000	21.0000	132.5000
	CC-SIS	5.0000	23.5000	100.0000	318.5000	716.5000	7.0000	11.0000	19.0000	45.5000	226.5000
	NIS	68.5000	230.5000	444.0000	685.0000	923.0000	44.0000	189.5000	378.0000	661.0000	908.0000
	SIS	805.0000	956.0000	991.0000	999.0000	1000.0000	601.5000	881.0000	961.5000	992.0000	1000.0000
	DC-SIS	27.0000	75.5000	147.0000	271.0000	557.5000	18.0000	71.5000	133.0000	242.0000	639.5000
	DC-RoSIS	60.0000	149.5000	244.0000	388.5000	648.0000	38.0000	102.0000	195.0000	323.0000	681.5000
(2.3)	CDC-SIS	6.0000	20.0000	58.5000	157.5000	522.5000	7.5000	19.0000	58.0000	141.5000	468.5000
	CC-SIS	13.0000	95.5000	239.5000	489.5000	850.0000	13.5000	87.0000	200.0000	420.5000	798.5000
	NIS	73.0000	247.5000	462.0000	704.5000	933.5000	66.5000	221.5000	446.0000	697.5000	916.5000
	SIS	519.5000	821.0000	945.0000	989.0000	1000.0000	281.0000	658.5000	867.5000	971.5000	1000.0000
	DC-SIS	13.5000	46.0000	116.5000	339.0000	827.0000	27.5000	111.0000	269.5000	538.0000	858.0000
	DC-RoSIS	43.0000	117.5000	214.5000	408.5000	814.5000	57.0000	164.5000	319.5000	546.0000	856.0000

Table 4 The proportion of P_j and P_{All}

Model	Method	Size	$\rho = 0.4$					$\rho = 0.8$				
			P_j				P_{All}	P_j				P_{All}
			X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
(2.1)	CDC-SIS	d_1	1.0000	1.0000	0.7460	1.0000	0.7460	1.0000	1.0000	0.4380	1.0000	0.4380
		d_2	1.0000	1.0000	0.8320	1.0000	0.8320	1.0000	1.0000	0.6380	1.0000	0.6380
		d_3	1.0000	1.0000	0.8880	1.0000	0.8880	1.0000	1.0000	0.7240	1.0000	0.7240
	CC-SIS	d_1	1.0000	1.0000	0.5920	1.0000	0.5920	1.0000	1.0000	0.2380	1.0000	0.2380
		d_2	1.0000	1.0000	0.7260	1.0000	0.7260	1.0000	1.0000	0.4620	1.0000	0.4620
		d_3	1.0000	1.0000	0.8020	1.0000	0.8020	1.0000	1.0000	0.5700	1.0000	0.5700
	NIS	d_1	1.0000	1.0000	0.0340	0.9940	0.0340	1.0000	1.0000	0.0100	1.0000	0.0100
		d_2	1.0000	1.0000	0.0760	0.9960	0.0760	1.0000	1.0000	0.0300	1.0000	0.0300
		d_3	1.0000	1.0000	0.1260	0.9980	0.1260	1.0000	1.0000	0.0520	1.0000	0.0520
	SIS	d_1	1.0000	1.0000	0.0000	0.0380	0.0000	1.0000	1.0000	0.0020	0.0160	0.0000
		d_2	1.0000	1.0000	0.0000	0.0760	0.0000	1.0000	1.0000	0.0040	0.0360	0.0000
		d_3	1.0000	1.0000	0.0000	0.0920	0.0000	1.0000	1.0000	0.0080	0.0700	0.0000
	DC-SIS	d_1	1.0000	1.0000	0.3820	0.0700	0.0260	1.0000	1.0000	0.0960	0.1240	0.0100
		d_2	1.0000	1.0000	0.5160	0.1480	0.0700	1.0000	1.0000	0.1620	0.2800	0.0500
		d_3	1.0000	1.0000	0.5800	0.1920	0.1120	1.0000	1.0000	0.2180	0.3820	0.0800
	DC-RoSIS	d_1	1.0000	1.0000	0.4020	0.0640	0.0260	1.0000	1.0000	0.1100	0.0660	0.0040
		d_2	1.0000	1.0000	0.5260	0.1020	0.0560	1.0000	1.0000	0.1740	0.1500	0.0260
		d_3	1.0000	1.0000	0.5940	0.1480	0.0780	1.0000	1.0000	0.2440	0.2160	0.0580
(2.2)	CDC-SIS	d_1	0.5800	0.7340	0.9060	1.0000	0.4420	0.9960	0.9980	0.6720	1.0000	0.6680
		d_2	0.7100	0.8440	0.9480	1.0000	0.6040	0.9980	1.0000	0.8400	1.0000	0.8380
		d_3	0.7840	0.8720	0.9560	1.0000	0.6780	0.9980	1.0000	0.8820	1.0000	0.8800

Table 4 continued

Model	Method	Size	$\rho = 0.4$					$\rho = 0.8$				
			P_j				P_{All}	P_j				P_{All}
			X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
	CC-SIS	d_1	0.3120	0.5500	0.8040	1.0000	0.2000	0.9840	0.9920	0.4520	1.0000	0.4400
		d_2	0.4100	0.6420	0.8940	1.0000	0.3080	0.9920	0.9960	0.6620	1.0000	0.6560
		d_3	0.4740	0.6780	0.9300	1.0000	0.3680	0.9940	1.0000	0.7760	1.0000	0.7700
	NIS	d_1	0.2260	0.3040	0.0600	1.0000	0.0080	0.4140	0.4980	0.0320	1.0000	0.0080
		d_2	0.3060	0.3980	0.1140	1.0000	0.0140	0.5460	0.6180	0.0700	1.0000	0.0280
		d_3	0.3860	0.4580	0.1580	1.0000	0.0320	0.6200	0.6640	0.1180	1.0000	0.0540
	SIS	d_1	0.1780	0.3840	0.0000	0.0540	0.0000	0.9160	0.9440	0.0000	0.0180	0.0000
		d_2	0.2360	0.4560	0.0000	0.0900	0.0000	0.9500	0.9580	0.0000	0.0420	0.0000
		d_3	0.2760	0.5140	0.0000	0.1060	0.0000	0.9660	0.9700	0.0000	0.0680	0.0000
	DC-SIS	d_1	0.4660	0.5180	0.6280	0.1360	0.0120	0.9900	0.9880	0.4760	0.0860	0.0400
		d_2	0.6200	0.6340	0.7260	0.2360	0.0640	0.9980	0.9960	0.5860	0.1900	0.1200
		d_3	0.7040	0.7040	0.7820	0.3340	0.1220	0.9980	0.9980	0.6560	0.2700	0.1740
	DC-RoSIS	d_1	0.2260	0.3040	0.6380	0.0520	0.0000	0.9720	0.9660	0.5000	0.0420	0.0120
		d_2	0.3980	0.4320	0.7220	0.1160	0.0040	0.9900	0.9800	0.6120	0.0820	0.0400
		d_3	0.5060	0.5500	0.7700	0.1800	0.0320	0.9940	0.9920	0.6760	0.1360	0.0820
(2.3)	CDC-SIS	d_1	1.0000	1.0000	0.4200	0.5160	0.2080	1.0000	1.0000	0.3780	0.6080	0.2100
		d_2	1.0000	1.0000	0.5240	0.6760	0.3580	1.0000	1.0000	0.5000	0.7440	0.3640
		d_3	1.0000	1.0000	0.6000	0.7440	0.4560	1.0000	1.0000	0.5760	0.7960	0.4500
	CC-SIS	d_1	0.9980	1.0000	0.2160	0.3060	0.0680	0.9920	0.9980	0.1760	0.3620	0.0620
		d_2	0.9980	1.0000	0.3140	0.3720	0.1160	0.9940	0.9980	0.2780	0.4480	0.1160
		d_3	0.9980	1.0000	0.3660	0.4260	0.1420	0.9980	1.0000	0.3560	0.5120	0.1580

Table 4 continued

Model	Method	Size	$\rho = 0.4$					$\rho = 0.8$				
			P_j				P_{All}	P_j				P_{All}
			X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
	NIS	d_1	0.8640	0.8780	0.0240	0.4060	0.0040	0.7580	0.7980	0.0180	0.4640	0.0060
		d_2	0.9160	0.9220	0.0500	0.4780	0.0120	0.8340	0.8480	0.0620	0.5760	0.0220
		d_3	0.9280	0.9360	0.0720	0.5120	0.0240	0.8660	0.8820	0.0860	0.6400	0.0320
	SIS	d_1	0.9900	0.9980	0.0000	0.1120	0.0000	0.9980	0.9960	0.0020	0.0720	0.0000
		d_2	0.9940	1.0000	0.0000	0.1320	0.0000	0.9980	1.0000	0.0100	0.1200	0.0000
		d_3	0.9960	1.0000	0.0000	0.1620	0.0000	0.9980	1.0000	0.0160	0.1520	0.0000
	DC-SIS	d_1	0.9980	1.0000	0.2320	0.3640	0.0760	1.0000	1.0000	0.0640	0.3660	0.0200
		d_2	1.0000	1.0000	0.3440	0.5680	0.2020	1.0000	1.0000	0.1180	0.5440	0.0640
		d_3	1.0000	1.0000	0.4040	0.6780	0.2600	1.0000	1.0000	0.1580	0.6500	0.0980
	DC-RoSIS	d_1	0.9980	1.0000	0.2680	0.0520	0.0060	1.0000	1.0000	0.0740	0.0700	0.0120
		d_2	0.9980	1.0000	0.3720	0.1000	0.0300	1.0000	1.0000	0.1360	0.1320	0.0200
		d_3	1.0000	1.0000	0.4340	0.1600	0.0660	1.0000	1.0000	0.1780	0.2100	0.0400
		d_3	1.0000	1.0000	0.4040	0.6780	0.2600	1.0000	1.0000	0.1580	0.6500	0.0980
	DC-RoSIS	d_1	0.9980	1.0000	0.2680	0.0520	0.0060	1.0000	1.0000	0.0740	0.0700	0.0120
		d_2	0.9980	1.0000	0.3720	0.1000	0.0300	1.0000	1.0000	0.1360	0.1320	0.0200
		d_3	1.0000	1.0000	0.4340	0.1600	0.0660	1.0000	1.0000	0.1780	0.2100	0.0400

Table 5 The quantile of S

Method	$\rho = 0.4$				
	5%	25%	50%	75%	95%
CDC-SIS	3.0000	4.0000	8.0000	29.0000	143.0000
CC-SIS	3.0000	8.0000	28.0000	107.0000	525.0000
SIS	216.5000	605.5000	840.0000	971.0000	1000.0000
DC-SIS	27.5000	135.5000	277.0000	536.0000	876.0000
DC-RoSIS	43.0000	179.5000	337.5000	599.5000	883.0000

this dataset. We treat MEDV (the median value of owner-occupied homes) as response, and log(DIS) (the weighted distances to five Boston employment centres) as the significant variable. It is reasonable because the geographical accessibility to employment is an important factor to consider when buying houses. The other 13 predictors are included, such as CRIM (per capita crime rate by town), NOX (nitric oxides concentration), LSTAT (lower status of the population) and so on.

Inspired by Fan et al. (2014), to evaluate our method in a high-dimensional setting, we expand the dataset by adding the artificial predictors:

$$X_j = \frac{Z_j + tU}{1+t}, \quad j = 14, 15, \ \ldots, p,$$

where $p = 1000$, $t = 2$, and Z_j, $j = 14, 15, \ \ldots, p$ are i.i.d. standard normal variables and U follows the standard uniform distribution. In this artificial example, we repeat the experiment 500 times. The results given in Table 7 are very appealing because our method can rank the 13 active variables before the artificial predictors. This implies that our method is very useful in high-dimensional data analysis.

Acknowledgements Wang's research was supported by the National Natural Science Foundation of China (General Program 11171331 and Key Program 11331011) and the National Natural Science Foundation for Creative Research Groups in China (61621003), a Grant from the Key Lab of Random Complex Structure and Data Science, CAS and Natural Science Fund of SZU.

Appendix

We first establish the following regularity conditions:

(C1) Denote the density function of W by $f(\cdot)$, and assume that it has continuous second derivatives. The support of W is assumed to be bounded and is denoted by $\mathcal{W} = [a, b]$ with finite constants a and b.

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

Table 6 The proportion of P_j and P_{All}

Method	Size	$\rho = 0.4$			
		P_j			P_{All}
		X_1	X_2	X_3	All
CDC-SIS	d_1	0.8180	0.9600	0.8160	0.6540
	d_2	0.8900	0.9780	0.8840	0.7780
	d_3	0.9140	0.9900	0.9100	0.8320
CC-SIS	d_1	0.5120	0.9100	0.7540	0.3880
	d_2	0.6200	0.9440	0.8360	0.5180
	d_3	0.6740	0.9600	0.8620	0.5840
SIS	d_1	0.0300	0.0720	0.0140	0.0000
	d_2	0.0440	0.1040	0.0320	0.0060
	d_3	0.0520	0.1380	0.0480	0.0100
DC-SIS	d_1	0.7380	0.6880	0.0360	0.0240
	d_2	0.8480	0.8280	0.0700	0.0620
	d_3	0.8980	0.8740	0.1000	0.0920
DC-RoSIS	d_1	0.5140	0.4820	0.0280	0.0100
	d_2	0.6480	0.6380	0.0540	0.0320
	d_3	0.7540	0.7280	0.0820	0.0600

Table 7 The quantile of S

Method	5%	25%	50%	75%	95%
CDC-SIS	13.0000	13.0000	13.0000	13.0000	13.0000

(C3) The random variables $\mathbf{X}$ and Y satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0, such that for $0 \leq s < s_0$,

$$\sup_{W \in \mathcal{W}} \max_{1 \leq j \leq p} E(\exp(sX_j^2|W)) < \infty,$$
$$\sup_{W \in \mathcal{W}} E(\exp(sY^2|W)) < \infty,$$

(C4) $\min_{j \in \mathcal{M}^*} \rho_{j0}^* \geq 2cn^{-\kappa}$ for some constant $c > 0$ and $0 \leq \kappa < 1/2$.

Proof of Theorem 1 The proof consists of three steps. We denote the positive constants c and C as generic constants depending on the context, which can vary from line to line.

Step 1. For some $0 \leq \kappa < 1/2$, we first prove

$$\max_{1 \leq j \leq p} \sup_{w \in [a,b]} P(|\hat{\rho}^2(X_j, Y|W = w) - \rho^2(X_j, Y|W = w)| \geq cn^{-\kappa}) \leq C \exp\left(-\frac{n^{-\kappa}}{Ch}\right). \tag{7}$$

Refer to the Supplemental material for the proof of Step 1.

Step 2. We prove $P(\max_{1\le j\le p}|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-k}) \le O(np\exp(-n^{\gamma-\kappa}/\xi))$.
Note that

$$P(|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa}) \le P(|\hat{\rho}_j^* - \rho_j^*| + |\rho_j^* - \rho_{j0}^*| \ge cn^{-\kappa})$$
$$\le P(|\hat{\rho}_j^* - \rho_j^*| \ge cn^{-\kappa}/2) + P(|\rho_j^* - \rho_{j0}^*| \ge cn^{-\kappa}/2).$$

By the definitions of $\hat{\rho}_j^*$, $\rho_j^* = \frac{1}{n}\sum_{i=1}^n \rho_j^2(W_i)$ with $\rho_j^2(w) = \rho^2(X_j, Y|W = w)$ and the result of Step 1, we have, for $j = 1, 2, \ldots, p$

$$\begin{aligned}
P(|\hat{\rho}_j^* - \rho_j^*| \ge cn^{-\kappa}/2) &= P\left(|\frac{1}{n}\sum_{i=1}^n \hat{\rho}_j^2(W_i) - \frac{1}{n}\sum_{i=1}^n \rho_j^2(W_i)| \ge cn^{-\kappa}/2\right)\\
&\le \sum_{i=1}^n P(|\hat{\rho}_j^2(W_i) - \rho_j^2(W_i)| \ge cn^{-\kappa}/2)\\
&\le Cn\exp\left(-\frac{n^{-\kappa}}{Ch}\right)\\
&= O(n\exp(-n^{\gamma-\kappa}/\xi)), \qquad (8)
\end{aligned}$$

where ξ is a positive constant, and $0 \le \kappa < \gamma$. By Hoeffding's inequality, for $j = 1, 2, \ldots, p$, it follows that

$$\begin{aligned}
P(|\rho_j^* - \rho_{j0}^*| \ge cn^{-\kappa}/2) &= P\left(|\frac{1}{n}\sum_{i=1}^n \rho_j^2(W_i) - E\rho_j^2(W_i)| \ge cn^{-\kappa}/2\right)\\
&\le 2\exp(-nc^2n^{-2\kappa}/2)) = O(\exp(-n^{1-2\kappa}/\xi)). \qquad (9)
\end{aligned}$$

Eq. (8) dominates Eq. (9). Hence, for $j = 1, 2, \ldots, p$, we get

$$P(|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa}) \le O(n\exp(-n^{\gamma-\kappa}/\xi)).$$

We thus have

$$P\left(\max_{1\le j\le p}|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-k}\right) \le O(np\exp(-n^{\gamma-\kappa}/\xi)).$$

Step 3. We prove $P(\mathcal{M}^* \subset \hat{\mathcal{M}}) \ge 1 - O(ns_n\exp(-n^{\gamma-\kappa}/\xi))$.
If $\mathcal{M}^* \not\subset \hat{\mathcal{M}}$, then there exist some $j \in \mathcal{M}^*$ such that $\hat{\rho}_j^* < cn^{-\kappa}$, due to $\min_{j\in\mathcal{M}^*}\rho_{j0}^* \ge 2cn^{-\kappa}$, $|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa}$ for some $j \in \mathcal{M}^*$, indicating that

$$\{\mathcal{M}^* \not\subset \hat{\mathcal{M}}\} \subset \{|\hat{\rho}_j^* - \rho_{j0}^*| \ge cn^{-\kappa} \quad \text{for some} \quad j \in \mathcal{M}^*\}.$$

Consequently,

$$\begin{aligned} P\{\mathcal{M}^* \subset \hat{\mathcal{M}}\} &\geq P\{\max_{j\in\mathcal{M}^*} |\hat{\rho}_j^* - \rho_{j0}^*| < cn^{-\kappa}\} \\ &= 1 - P\{\max_{j\in\mathcal{M}^*} |\hat{\rho}_j^* - \rho_{j0}^*| \geq cn^{-\kappa}\} \\ &\geq 1 - s_n P\{|\hat{\rho}_j^* - \rho_{j0}^*| \geq cn^{-\kappa}\} \\ &\geq 1 - O(ns_n \exp(-n^{\gamma-\kappa}/\xi)). \end{aligned}$$

□

References

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications*, Monographs on Statistics and Applied Probability, vol. 66. Chapman and Hall, London.

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

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

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

Fan, J., Feng, Y., Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, *106*(494), 544–557.

Fan, J., Ma, Y., Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*(507), 1270–1284.

Harrison, D., Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

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

Liu, J., Li, R., Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, *109*(505), 266–274.

Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Wang, Q. H., Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, *30*(3), 896–924.

Wang, X., Pan, W., Hu, W., Tian, Y., Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, *110*(512), 1726–1734.

Zhong, W., Zhu, L., Li, R., Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, *26*(1), 69–95.

Zhu, L. P., Li, L., Li, R., Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*(496), 1464–1475.