Journal of Multivariate Analysis 173 (2019) 480–493

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sufficient variable selection using independence measures for continuous response

Baoying Yang [a], Xiangrong Yin [b,*], Nan Zhang [c]

[a] Department of Statistics, College of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan 611756, China
[b] Department of Statistics, 319 Multidisciplinary Science Building, University of Kentucky, Lexington, KY 40536, USA
[c] Department of Statistics, 204 Statistics Building, 101 Cedar Street, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 17 November 2018
Available online 29 April 2019

AMS 2010 subject classifications:
primary 62H20
secondary 62G20
62G35

Keywords:
Distance correlation
Hilbert–Schmidt independence criterion
Independence test
Marginal feature screening
Sufficient variable selection

ABSTRACT

We propose two sufficient variable selection procedures, i.e., one- and two-stage approaches using independence measures for continuous response, illustrated by distance correlation and the Hilbert–Schmidt Independence Criterion correlation. We show the advantages of the proposed procedures over some existing marginal screening methods through simulations and a real data analysis. Our procedures are model-free and thus robust against model mis-specification. They are particularly useful when some active predictors are marginally independent of the response.

1. Introduction

Variable selection has become increasingly important in various research fields, as data are being collected at a relatively low cost due to modern technology. Many methods have been proposed over the last two decades, such as the least absolute shrinkage and selection operator (Lasso) [37], the smoothly clipped absolute deviation (SCAD) [10], and the Dantzig selector [2]. These methods have shown promise in dealing with high-dimensional data.

For ultrahigh-dimensional data, however, Fan and Lv [11] pointed out that the aforementioned methods have limitations due to the challenges of computational cost, statistical accuracy, and algorithmic stability. These concerns led to the sure independent screening (SIS) method in [11] for ultrahigh-dimensional data. The SIS method is based on the marginal Pearson correlation learning and is designed for linear regressions with Gaussian predictors and responses. SIS not only can speed up variable selection drastically but can also improve the estimation accuracy when dimensionality is ultrahigh. Many other methods have been developed in recent years, following SIS with specified models, both parametric and semi-parametric; see, e.g., [3,4,9,12,13,24,33]. However, specifying a correct model for ultrahigh-dimensional data may be challenging.

As the aforementioned model-specific screening procedures may not be robust to model mis-specification, model-free sure screening procedures have been developed; see, e.g., [1,8,19–21,23,25,26,32,41]. Fan and Lv [11] pointed out that the marginal screening procedure may miss some active predictors that are marginally independent of the response, and they proposed iterative sure independence screening (ISIS) to overcome the problem. Although this idea has been empirically

* Corresponding author.
E-mail address: yinxiangrong@uky.edu (X. Yin).

https://doi.org/10.1016/j.jmva.2019.04.006
0047-259X/

demonstrated in [9,13,23,41], its theoretical justification still remains unclear. Mai and Zou [25,26] discussed the subtle difference between variable selection and variable screening: the former uses fine methods to exactly select the active set of predictors, while the latter uses rough but fast methods to select a set containing the active set of predictors. The existing variable screening methods may not always select such a set, though they often work in practice. This motivates us to explore new procedures to overcome the drawback of existing screening approaches, and to seek theoretical guarantees that they select a set containing all active predictors.

In this paper, focusing on continuous response, we propose two new sufficient variable selection approaches based on theoretical results from the sufficient dimension reduction literature. These two new approaches translate conditional independence in sufficient variable selection to alternative measures of independence. The independence statistic is illustrated by distance correlation [35,36] and Hilbert–Schmidt Independence Criterion correlation [15]. Although fine statistical tests could be developed for these procedures, we only use the screening approach for the purpose of sufficient variable selection as it is fast and cost-efficient, even though the selected set may be larger than the set of active predictors. Our approach is model-free. Thus, it is robust against model mis-specification, which is an attractive property in practice. Also, our methods allow for arbitrary regression relationships, which makes them more effective than the model-specific marginal approaches. More importantly, our proposed procedures are advantageous when some active predictors are marginally independent of the response.

The rest of this paper is organized as follows. Section 2 describes both distance correlation and Hilbert–Schmidt Independence Criterion correlation for sufficient variable screening. Section 3 reports some of their theoretical properties, while Section 4 contains simulation studies and a real data application. A short discussion follows in Section 5. Related proofs of theorems and additional simulations can be found in the Online Supplement.

Throughout this paper, we assume that Y is a univariate or multivariate response variable, and $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is a $p \times 1$ vector. The notation $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{W}$ means that $\mathbf{U}$ and $\mathbf{V}$ are independent given $\mathbf{W}$.

2. Methodology

2.1. Sufficient variable selection

We adopt the following definition of sufficient variable selection from Yin and Hilafu [38].

Definition 1. If there is a $p \times q$ matrix $\mathbf{A}$ with $q \leq p$, where the columns of $\mathbf{A}$ consist of unit vectors, e_αs, whose αth element is 1, such that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{A}^\top\mathbf{X}$, then the column space of $\mathbf{A}$ is called a variable selection space. The intersection of all such spaces, if it satisfies the conditional independence condition above, is called the central variable selection space, denoted by $\mathcal{S}^V_{Y|\mathbf{X}}$.

Let $\mathbf{X}_\mathcal{D}$ be the set of X_k which are involved in $\mathcal{S}^V_{Y|\mathbf{X}}$ and $\mathbf{X}_{\bar{\mathcal{D}}}$ be its complement, where $\mathcal{D}$ and $\bar{\mathcal{D}}$ are the respective index sets. In this paper, we assume the existence of $\mathcal{S}^V_{Y|\mathbf{X}}$. Then Definition 1 is equivalent to $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}}|\mathbf{X}_\mathcal{D}$, where $\mathbf{X}_\mathcal{D}$ is the set of active variables, which is smallest and unique. Yin and Hilafu [38] concluded that the existence of a central subspace implies the existence of $\mathcal{S}^V_{Y|\mathbf{X}}$. Conditions for the existence of the central subspace were obtained by Cook [5] and Yin et al. [39]. In fact, the existence of the central subspace implies that the set of variables involved in that subspace is $\mathbf{X}_\mathcal{D}$. Therefore, the goal is to find $\mathbf{X}_\mathcal{D}$. Directly using the conditional independence, $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}}|\mathbf{X}_\mathcal{D}$, seems infeasible, because it is hard to decide which and how many variables should be included in the set $\mathbf{X}_\mathcal{D}$.

Note that the popular SIS in [11] and its family consider the independence, $Y \perp\!\!\!\perp X_\alpha$, then rank all X_αs in decreasing order based on the strength of the measure, and choose the first d (threshold value) predictors as an estimator containing $\mathbf{X}_\mathcal{D}$. However, there are fundamental differences between this approach and $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}}|\mathbf{X}_\mathcal{D}$. The former is looking at the marginal relation, while the latter is focused on a conditional relation. Fan and Lv [11] and Zhu et al. [41] pointed out that the marginal feature screening procedure may miss those predictors which are marginally unrelated but jointly related to the response. To partly eliminate this, they proposed an iterative procedure that computes the correlation between the response and the residual of the remaining $\mathbf{X}$s. The iterative procedure performs well empirically, but its theoretical justification remains unclear. We propose two novel sufficient variable selection procedures to achieve the conditional independence, based on a simplified version of Proposition 1 in [38] as below.

Proposition 1. *Let $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ be random vectors, and $\mathbf{X}^\top = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)$, then statement (i) or statement (ii) implies statement (iii):*

(i) $(Y, \mathbf{X}_2) \perp\!\!\!\perp \mathbf{X}_1$;
(ii) $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|Y$ *and* $Y \perp\!\!\!\perp \mathbf{X}_1$;
(iii) $Y \perp\!\!\!\perp \mathbf{X}_1|\mathbf{X}_2$.

Note that statement (iii) implies that $\Pr(Y|\mathbf{X}_1, \mathbf{X}_2) = \Pr(Y|\mathbf{X}_2)$. Therefore, if statement (iii) holds, then we can eliminate $\mathbf{X}_1$ without losing any regression information. After eliminating $\mathbf{X}_1$, we treat $\mathbf{X}_2$ as a new $\mathbf{X}$, split it, and then do a further test until nothing can be eliminated. Hence, in the end, the set contains $\mathbf{X}_\mathcal{D}$. Thus, statement (iii) is very

important. However, statement (iii) is just the goal of the sufficient variable selection, i.e., the conditional independence test $Y \perp\!\!\!\perp \mathbf{X}_{\widehat{\mathcal{D}}} | \mathbf{X}_{\mathcal{D}}$. Hence, it is difficult to test statement (iii) directly because we do not know $\mathbf{X}_2$ in advance.

Note that statement (iii) can be forced to hold if either of statement (i) or statement (ii) holds. Therefore, developing methods for testing statements (i) and (ii) is useful. To do so, we propose two sufficient variable selection approaches based on statements (i) and (ii), which we call one-stage sufficient variable selection and two-stage sufficient variable selection, respectively. It appears that statement (i) is naturally useful for a continuous response as the combination of Y and $\mathbf{X}$s makes sense, while statement (ii) is naturally useful for a categorical response as the first part of statement (ii) is a conditional test. Note that SIS [11] and its family only use the second part of statement (ii) for scalar $\mathbf{X}_1$, which is not sufficient to imply (iii). While statement (i) or statement (ii) implies statement (iii), the converse is not true. Hence, situations where (iii) holds while either (i) or (ii) fails are excluded.

In this paper, we use distance correlation (DC) [36] and Hilbert–Schmidt Independence Criterion (HSIC) correlation [15] to illustrate the two sufficient procedures for continuous response. Let $(\mathbf{U}, \mathbf{V}) \in \mathcal{U} \times \mathcal{V}$, and $\mathbf{U}$ and $\mathbf{V}$ be random vectors with dimensions d_u and d_v, respectively. Suppose that $(\mathbf{U}', \mathbf{V}')$ is an iid copy of $(\mathbf{U}, \mathbf{V})$. Suppose that $(\mathbf{U}_1, \mathbf{V}_1), \ldots, (\mathbf{U}_n, \mathbf{V}_n)$ is a random sample of $(\mathbf{U}, \mathbf{V})$. Next we review DC and HSIC correlation.

2.2. Distance correlation

Suppose that $\psi_{\mathbf{U}}$ and $\psi_{\mathbf{V}}$ are the respective characteristic functions of $\mathbf{U}, \mathbf{V}$, and $\psi_{\mathbf{U},\mathbf{V}}$ is their joint characteristic function. Following [36], the distance covariance between $\mathbf{U}$ and $\mathbf{V}$ is the nonnegative number $\mathrm{dcov}(\mathbf{U}, \mathbf{V})$ given by

$$\mathrm{dcov}^2(\mathbf{U}, \mathbf{V}) = \int_{\mathcal{U}\times\mathcal{V}} \|\psi_{\mathbf{U},\mathbf{V}}(t,s) - \psi_{\mathbf{U}}(t)\psi_{\mathbf{V}}(s)\|^2 w(t,s)dtds,$$

where $\|\psi\|^2 = \psi\bar{\psi}$ for a complex-valued function ψ with $\bar{\psi}$ being the conjugate of ψ, and

$$w(t,s) = \left(c_{d_u} c_{d_v} \|t\|_{d_u}^{1+d_u} \|s\|_{d_v}^{1+d_v}\right)^{-1},$$

with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$, where $\|a\|_d$ stands for the Euclidean norm of $a \in \mathbb{R}^d$.

The DC between $\mathbf{U}$ and $\mathbf{V}$ is defined as

$$\mathrm{dcorr}(\mathbf{U}, \mathbf{V}) = \frac{\mathrm{dcov}(\mathbf{U}, \mathbf{V})}{\sqrt{\mathrm{dcov}(\mathbf{U}, \mathbf{U}')\mathrm{dcov}(\mathbf{V}, \mathbf{V}')}}.$$

An important property is that $\mathrm{dcorr}(\mathbf{U}, \mathbf{V}) = 0$ if and only if $\mathbf{U}$ and $\mathbf{V}$ are independent. Székely et al. [36] expressed $\mathrm{dcov}^2(\mathbf{U}, \mathbf{V})$ as $\mathrm{dcov}^2(\mathbf{U}, \mathbf{V}) = S_1 + S_2 - 2S_3$, where $S_1 = \mathrm{E}\|\mathbf{U} - \mathbf{U}'\|_{d_u}\|\mathbf{V} - \mathbf{V}'\|_{d_v}$, $S_2 = \mathrm{E}\|\mathbf{U} - \mathbf{U}'\|_{d_u}\mathrm{E}\|\mathbf{V} - \mathbf{V}'\|_{d_v}$, and $S_3 = \mathrm{E}\{\mathrm{E}(\|\mathbf{U} - \mathbf{U}'\|_{d_u}|\mathbf{U})\mathrm{E}(\|\mathbf{V} - \mathbf{V}'\|_{d_v}|\mathbf{V})\}$. The respective sample versions of S_1, S_2 and S_3 are

$$\hat{S}_1 = \frac{1}{n^2}\sum_{i,j=1}^{n} \|\mathbf{U}_i - \mathbf{U}_j'\|_{d_u}\|\mathbf{V}_i - \mathbf{V}_j'\|_{d_v}, \quad \hat{S}_2 = \frac{1}{n^2}\sum_{i,j=1}^{n} \|\mathbf{U}_i - \mathbf{U}_j'\|_{d_u} \frac{1}{n^2}\sum_{i,j=1}^{n} \|\mathbf{V}_i - \mathbf{V}_j'\|_{d_v},$$

$$\hat{S}_3 = \frac{1}{n^3}\sum_{i,j,\ell=1}^{n} \|\mathbf{U}_i - \mathbf{U}_\ell'\|_{d_u}\|\mathbf{V}_j - \mathbf{V}_\ell'\|_{d_v}.$$

Thus, an estimator of $\mathrm{dcov}^2(\mathbf{U}, \mathbf{V})$ is $\widehat{\mathrm{dcov}}^2(\mathbf{U}, \mathbf{V}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$. Putting this into the formula of $\mathrm{dcorr}(\mathbf{U}, \mathbf{V})$, we obtain an estimator of DC.

2.3. HSIC correlation

Following Gretton et al. [15], let $\mathcal{F}$ and $\mathcal{G}$ be the respective universal Reproducing Kernel Hilbert space (RKHS) on $\mathcal{U}$ and $\mathcal{V}$. For each point $\mathbf{U} \in \mathcal{U}$, there corresponds an element $\phi(\mathbf{U}) \in \mathcal{F}$ satisfying $K(\mathbf{U}, \mathbf{U}') = \langle\phi(\mathbf{U}), \phi(\mathbf{U}')\rangle_K$, where $K : \mathcal{U}\times\mathcal{U} \to \mathbb{R}$ is a positive definite kernel with inner-product $\langle\cdot,\cdot\rangle_K$. Similarly, for each point $\mathbf{V} \in \mathcal{V}$, there corresponds an element $\varphi(\mathbf{V}) \in \mathcal{G}$ satisfying $L(\mathbf{V}, \mathbf{V}') = \langle\varphi(\mathbf{V}), \varphi(\mathbf{V}')\rangle_L$, where $L : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ is a positive definite kernel with inner-product $\langle\cdot,\cdot\rangle_L$. The cross covariance operator $\mathrm{C}_{\mathbf{U},\mathbf{V}} : \mathcal{F} \to \mathcal{G}$ is defined for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$ via the bilinear form

$$\langle g, \mathrm{C}_{\mathbf{U},\mathbf{V}} f\rangle_{\mathcal{G}} = \mathrm{cov}\{f(\mathbf{U}), g(\mathbf{V})\} = \mathrm{E}_{\mathbf{U},\mathbf{V}}\{f(\mathbf{U})g(\mathbf{V})\} - \mathrm{E}_{\mathbf{U}}\{f(\mathbf{U})\}\mathrm{E}_{\mathbf{V}}\{g(\mathbf{V})\}.$$

The HSIC covariance between random vectors $\mathbf{U}$ and $\mathbf{V}$ can be formulated (See [15,16]) as

$$H(\mathbf{U}, \mathbf{V}) = \|\mathrm{C}_{\mathbf{U},\mathbf{V}}\|_{\mathcal{HS}} = \mathrm{E}\{K(\mathbf{U}, \mathbf{U}')L(\mathbf{V}, \mathbf{V}')\} + \mathrm{E}\{K(\mathbf{U}, \mathbf{U}')\}\mathrm{E}\{L(\mathbf{V}, \mathbf{V}')\} - 2\mathrm{E}[\mathrm{E}\{K(\mathbf{U}, \mathbf{U}')|\mathbf{U}\}\mathrm{E}\{L(\mathbf{V}, \mathbf{V}')|\mathbf{V}\}].$$

That is, the HSIC covariance is defined as the Hilbert–Schmidt norm of the cross-covariance operator $\mathrm{C}_{\mathbf{U},\mathbf{V}}$. The Hilbert–Schmidt norm of $\mathrm{C}_{\mathbf{U},\mathbf{V}}$ exists when the various expectations over the kernels are bounded, which is true as long as the kernels K and L are bounded [15]. It can be shown that $H(\mathbf{U}, \mathbf{V}) = 0$ if and only if $\mathbf{U}$ and $\mathbf{V}$ are statistically independent,

as long as the associated RKHSs $\mathcal{F}$ and $\mathcal{G}$ are universal [15], i.e., they are dense in the space of continuous functions with respect to the infinity norm [17].

Examples of kernels generating universal RKHS are, e.g., the Gaussian and the Laplace kernels [34]. In our algorithm we choose Gaussian kernel, which belongs to the class of translation-invariant kernels, i.e., the kernel functions with the property $K(\mathbf{U}, \mathbf{U}') = K(\mathbf{U} - \mathbf{U}')$. Let $H(\mathbf{U}) = H(\mathbf{U}, \mathbf{U})$. The HSIC correlation (HR) [40] between $\mathbf{U}$ and $\mathbf{V}$ is the nonnegative number $\mathrm{HR}(\mathbf{U}, \mathbf{V})$ defined by

$$\mathrm{HR}(\mathbf{U}, \mathbf{V}) = \begin{cases} \dfrac{H(\mathbf{U}, \mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}} & \text{if } H(\mathbf{U})H(\mathbf{V}) > 0, \\ 0 & \text{if } H(\mathbf{U})H(\mathbf{V}) = 0. \end{cases}$$

The function HR generalizes the idea of Pearson correlation in the sense that $\mathrm{HR}(\mathbf{U}, \mathbf{V}) = 0$ corresponds to the independence of $\mathbf{U}$ and $\mathbf{V}$. One empirical HSIC [15] is

$$H_n(\mathbf{U}, \mathbf{V}) = \frac{1}{n^2}\sum_{i,j} K_{ij}L_{ij} - \frac{2}{n^3}\sum_{i,j,k} K_{ij}L_{ik} + \frac{1}{n^4}\sum_{i,j,k,\ell} K_{ij}L_{k\ell},$$

where

$$K_{ij} = \exp\left\{-\frac{1}{2}(\mathbf{U}_i - \mathbf{U}_j)^\top \hat{\Sigma}_{\mathbf{U}}^{-1}(\mathbf{U}_i - \mathbf{U}_j)\right\} \quad \text{and} \quad L_{k\ell} = \exp\left\{-\frac{1}{2}(\mathbf{V}_k - \mathbf{V}_\ell)^\top \hat{\Sigma}_{\mathbf{V}}^{-1}(\mathbf{V}_k - \mathbf{V}_\ell)\right\}, \tag{1}$$

and $\hat{\Sigma}_{\mathbf{U}}$ and $\hat{\Sigma}_{\mathbf{V}}$ are the respective sample covariance matrices. Let $H_n(\mathbf{U}) = H_n(\mathbf{U}, \mathbf{U})$, the corresponding sample HR, HR_n, is then

$$\mathrm{HR}_n(\mathbf{U}, \mathbf{V}) = \begin{cases} \dfrac{H_n(\mathbf{U}, \mathbf{V})}{\sqrt{H_n(\mathbf{U})H_n(\mathbf{V})}} & \text{if } H_n(\mathbf{U})H_n(\mathbf{V}) > 0, \\ 0 & \text{if } H_n(\mathbf{U})H_n(\mathbf{V}) = 0. \end{cases}$$

Note that in practice, the sample covariance matrices in the kernel of (1) are singular for ultrahigh-dimensional data. In this case, there are two ways to solve this problem. One way is to normalize each component of $\mathbf{U}$ to have unit variance, and rewrite the kernel as

$$K_{ij} = \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{p}(U_{i\alpha} - U_{j\alpha})^2\right\},$$

which is equivalently to use product kernel [28,31]. Another way is to follow the method of Schafer and Strimmer [27] to shrink the inverse of the covariance matrix. This method is included in the `R` package `Corpcor`. Our simulations indicate that these two methods yield similar results. Thus we only report the latter one in the paper.

It is interesting to note that distance covariance and HSIC covariance are equivalent in that they are different kernel choices under the RKHS [30].

2.4. Algorithms

We describe algorithms using an independence index $\mathcal{I}(\mathbf{U}, \mathbf{V})$ for HR and DC, where $\mathbf{U}$ and $\mathbf{V}$ are two generic random vectors. Let X_α be the αth predictor in $\mathbf{X}$ and Y be the response variable. Suppose

$$\mathbf{X}_{-\alpha} = (X_1, \ldots, X_{\alpha-1}, X_{\alpha+1}, \ldots, X_p).$$

The three algorithms based on Proposition 1 are as follows.

(i) The *marginal feature selection* (FS_M, i.e., SIS procedure) calculates the marginal relation in the second part of statement (ii), i.e., the measure is $u_\alpha = \mathcal{I}(Y, X_\alpha)$.
(ii) The *one-stage sufficient variable selection* procedure (SVS_1) uses statement (i), i.e., the measure is $u_\alpha^* = \mathcal{I}\{(Y, \mathbf{X}_{-\alpha}), X_\alpha\}$.
(iii) The *two-stage sufficient variable selection* procedure (SVS_2) uses statement (ii), i.e., the two measures are $u_\alpha^{**} = \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y\}$ and u_α.

Calculating u_α and u_α^* is straightforward, however, u_α^{**} involves a conditional form. To overcome this conditional form, we use a slicing approach. If Y is a continuous, let

$$\mathcal{J} = \{[\ell_{s-1}, \ell_s) : \ell_{s-1} < \ell_s, s \in \{1, \ldots, S\}, \cup_{\ell=1}^{S}[\ell_{s-1}, \ell_s) \setminus \ell_0 = \mathbb{R}, \ell_0 = -\infty, \ell_S = \infty\},$$

where ℓ_s is the s/Sth sample quantile of Y. Note that the interval (l_0, l_1) is open, but we abuse the notation a little by writing the intervals $[l_{s-1}, l_s)$ for all s. Each $[l_{s-1}, l_s)$ is called a slice. Define $Z \in \{1, \ldots, S\}$ such that $Z = s$ if and only if $Y \in [\ell_{s-1}, \ell_s)$. If Y is discrete, i.e., $Y \in \{1, \ldots, S\}$, let $Z = Y$. Denote $u_{\alpha,s}^{\mathcal{J}} = \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Z = s\}$ for $s \in \{1, \ldots, S\}$. We have $u_\alpha^{**} = \sum_{s=1}^{S} u_{\alpha,s}^{\mathcal{J}}$.

Given a random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, we then construct their respective estimators as $\hat{u}_\alpha = \hat{\mathcal{I}}\{Y, X_\alpha\}$, $\hat{u}^*_\alpha = \hat{\mathcal{I}}\{(Y, \mathbf{X}_{-\alpha}), X_\alpha\}$, and $\hat{u}^{**}_\alpha = \hat{\mathcal{I}}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y\}$. The detailed algorithms are as follows.

Marginal feature selection (FS_M)

1. Calculate $\hat{u}_\alpha = \hat{\mathcal{I}}\{Y, X_\alpha\}$ for $\alpha \in \{1, \ldots, p\}$. The estimate $\mathbf{X}_{\hat{\mathcal{D}}_1}$ is the set of X_αs with the largest d values of $\hat{u}_\alpha$, where d is the preselected value.
2. The estimate $\mathbf{X}_{\hat{\mathcal{D}}}$ is the same as $\mathbf{X}_{\hat{\mathcal{D}}_1}$.

One-stage sufficient variable selection (SVS_1)

1. Calculate $\hat{u}_\alpha$ for $\alpha \in \{1, \ldots, p\}$. The estimate $\mathbf{X}_{\hat{\mathcal{D}}_1}$ is the set of X_αs with the largest d_1 values of $\hat{u}_\alpha$.
2. Calculate $\hat{u}^*_\alpha = \hat{\mathcal{I}}\{(Y, \mathbf{X}_{-\alpha}), X_\alpha\}$ for $\alpha \in \{1, \ldots, p\}$. The estimate $\mathbf{X}_{\hat{\mathcal{D}}^*}$ is the set of X_αs with the largest d_2 values of $\hat{u}^*_\alpha$ but not in $\mathbf{X}_{\hat{\mathcal{D}}_1}$ of Step 1.
3. The final estimate $\mathbf{X}_{\hat{\mathcal{D}}}$ is the union of these two sets, i.e., $\mathbf{X}_{\hat{\mathcal{D}}_1} \cup \mathbf{X}_{\hat{\mathcal{D}}^*}$.

Two-stage sufficient variable selection (SVS_2)

1. Calculate $\hat{u}_\alpha$ for $\alpha \in \{1, \ldots, p\}$. The estimate $\mathbf{X}_{\hat{\mathcal{D}}_1}$ is the set of X_αs with the largest d_1 values of $\hat{u}_\alpha$.
2. Obtain the conditional set:
   (a) Slice Y into S non-overlapping slices.
   (b) Calculate $\hat{u}^{**}_\alpha = \sum_{s=1}^{S} \hat{u}^{\mathcal{J}}_{\alpha,s}$ with $\hat{u}^{\mathcal{J}}_{\alpha,s} = \hat{\mathcal{I}}\{(X_\alpha, \mathbf{X}_{-\alpha})|Z = s\}$.
   (c) The estimate $\mathbf{X}_{\hat{\mathcal{D}}^{**}}$ is the set of X_αs with the largest d_2 values of $\hat{u}^{**}_\alpha$ but not in $\mathbf{X}_{\hat{\mathcal{D}}_1}$ of Step 1.
3. The final estimate $\mathbf{X}_{\hat{\mathcal{D}}}$ is the union of these two sets, i.e., $\mathbf{X}_{\hat{\mathcal{D}}_1} \cup \mathbf{X}_{\hat{\mathcal{D}}^{**}}$.

Note that in SVS_1, we add the marginal screening procedure (which is implied by statement (i)). This is because in practice the marginal relation typically does play an important role. Note as well that in SVS_2, the number of slices S can be subjective. For computational ease, we choose $S = 2$ in the simulation study, which seems to work well. Li et al. [23] stressed that the choice of threshold value d is very important, and they suggested a user-specified model size d. Based on their suggestion, we use $d = \lfloor p/\ln(p) \rfloor$ when $n > p$, and $d = \lfloor n/\ln(n) \rfloor$ or $n - 1$ when $n < p$. Avoiding the ad hoc choice of d, Kong et al. [21] developed a sequential method for variable screening. Alternatively, one can estimate d through an independent test, such as a permutation or a bootstrap test.

Our simulations below indicate that the power of DC/HR (our extra step in SVS_1 and SVS_2) for picking up active variables that are marginally related to the response is small. However, they are powerful in picking up active variables that are marginally independent of the response but are conditionally dependent of the response. In contrast, the marginal screening method is not powerful in selecting the active variables that are marginally independent of the response but conditionally dependent of the response. Thus, the two procedures complement each other and the combination of the two, in fact, significantly improves the result of the overall percentage of correctly selecting active variables. We propose to use $d_1 = 0.95d$ and $d_2 = 0.05d$, which works well in our simulations. Certainly, if the number of conditional important predictors increases, then we should increase d_2.

3. Theoretical properties

We now study theoretical properties of the screening procedures of $\hat{u}_\alpha$, $\hat{u}^*_\alpha$ and $\hat{u}^{**}_\alpha$. To study $\hat{u}_\alpha$, we only focus on HR, as Li et al. [23] had the result for DC. Note the fact that $u_\alpha = \mathcal{I}(Y, X_\alpha) = 0$ if and only if Y and X_α are independent with $\alpha \in \{1, \ldots, p\}$. This fact guarantees that HR ranks the marginal active predictors above the marginal inactive ones, i.e., $\max_{\alpha \in \tilde{\mathcal{D}}_1} u_\alpha < \min_{\alpha \in \mathcal{D}_1} u_\alpha$, where $\mathcal{D}_1$ is the marginal active set, and separates the marginal active ones from the marginal inactive ones. Hence, the quantity u_α can be used for marginal feature screening. We assume the following two conditions.

(C1): The universal kernels K and L are bounded and non-negative.
(C2): $\min_{\alpha \in \mathcal{D}_1} u_\alpha \geq 2cn^{-\nu}$ for some $\nu \in [0, 1/2)$ and $c > 0$.

Condition (C1) makes sure that the HSIC norm is finite, because various expectations, (i.e., $\mathrm{E}_{\mathbf{X},\mathbf{X}'}\{K(\mathbf{X}, \mathbf{X}')\}$ and $\mathrm{E}_{\mathbf{Y},\mathbf{Y}'}\{L(\mathbf{Y}, \mathbf{Y}')\}$) over bounded kernels K and L are also bounded [15]. Without loss of generality, we assume that K and L are bounded by 1, i.e., $\|K\| \leq 1$ and $\|L\| \leq 1$. Condition (C2) means that the marginal HR of active predictors cannot be too small. This assumption is equivalent to the condition (3) in [11] and condition (C2) in [23]. Condition (C2) reflects the signal strength of individual active predictor, which in turn controls the rate of probability error in selecting the active

predictors [41]. Define the selected set as

$$\hat{\mathcal{D}}_1 = \{\alpha : \hat{u}_\alpha \geq cn^{-\nu}\}.$$

Let c_1 and c_2 be positive generic constants. We have the following result.

Theorem 1. *Under condition* (C1)*, let* $\nu \in (0, 1/2)$*, there exists a positive constant* $c_1 > 0$ *such that*

$$\Pr\Big(\max_{1\leq\alpha\leq p} |\hat{u}_\alpha - u_\alpha| \geq cn^{-\nu}\Big) \leq O[p\{\exp(-c_1 n^{1-2\nu})\}]$$

Furthermore, under condition (C2)*, denote* $\delta = \min_{\alpha\in\mathcal{D}_1} u_\alpha - \max_{\alpha\in\bar{\mathcal{D}}_1} u_\alpha$*, i.e.,* $\delta \geq 2cn^{-\nu} > 0$*. Then we have*

$$\Pr\Big(\max_{\alpha\in\bar{\mathcal{D}}_1} \hat{u}_\alpha < \min_{\alpha\in\mathcal{D}_1} \hat{u}_\alpha\Big) \geq 1 - O[p\{\exp(-c_1 n^{1-2\nu})\}] \tag{2}$$

Under conditions (C1) *and* (C2)*, we have that*

$$\Pr(\mathcal{D}_1 \subseteq \hat{\mathcal{D}}_1) \geq 1 - O[s_n\{\exp(-c_1 n^{1-2\nu})\}], \tag{3}$$

where s_n *is the cardinality of* $\mathcal{D}_1$*.*

The proof of Theorem 1 is in the Online Supplement. The sure screening property holds for HR FS_M procedure. It also indicates that we can handle the non-polynomial (NP) dimensionality of order $\ln p = o(n^{1-2\nu})$ with $\nu \in (0, 1/2)$, meaning that, if $\ln p = o(n^{1-2\nu})$ with $\nu \in (0, 1/2)$, then (2) indicates that the probability of $\max_{\alpha\in\bar{\mathcal{D}}_1} \hat{u}_\alpha < \min_{\alpha\in\mathcal{D}_1} \hat{u}_\alpha$ approaches 1 as $n \to \infty$. That is, $\hat{u}_\alpha$ always ranks the active predictors above the inactive ones in probability; and furthermore, in such a case, (3) indicates that the size of true active predictors can change with p, as it is no larger than p.

To study $\hat{u}^*_\alpha$, we focus on HR only, leaving related conditions and results for DC in the Online Supplement. Note that $u^*_\alpha = \mathcal{I}\{(Y, \mathbf{X}_{-\alpha}), X_\alpha\} = 0$ if and only if $(Y, \mathbf{X}_{-\alpha}) \perp\!\!\!\perp X_\alpha$, which guarantees $Y \perp\!\!\!\perp X_\alpha|\mathbf{X}_{-\alpha}$ based on Proposition 1. This fact makes sure that the quantity u^*_α can attain our ultimate goal $Y \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{D}}^*}|\mathbf{X}_{\mathcal{D}^*}$, where $\mathbf{X}_{\mathcal{D}^*}$ is the set of predictors whose $u^*_\alpha > 0$. Condition (C1) is still needed to ensure the existence of the related (HSIC) norm. Condition (C2) is replaced by the following condition.

(C2*): $\min_{\alpha\in\mathcal{D}^*} u^*_\alpha \geq 2cn^{-\nu}$ for some $\nu \in [0, 1/2)$ and $c > 0$.

Condition (C2*) reflects that the resulting HR for active variables cannot be too small. Define the selected set

$$\hat{\mathcal{D}}^* = \{\alpha : \hat{u}^*_\alpha \geq cn^{-\nu}\}.$$

We have the following result.

Theorem 2. *Under condition* (C1) *with* $\nu \in (0, 1/2)$*, there exists a positive constant* $c_1 > 0$ *such that*

$$\Pr\Big(\max_{1\leq\alpha\leq p} |\hat{u}^*_\alpha - u^*_\alpha| \geq cn^{-\nu}\Big) \leq O[p\{\exp(-c_1 n^{1-2\nu})\}].$$

Furthermore, under condition (C2*)*, denote* $\delta = \min_{\alpha\in\mathcal{D}^*} u^*_\alpha - \max_{\alpha\in\bar{\mathcal{D}}^*} u^*_\alpha$*, i.e.,* $\delta \geq 2cn^{-\nu} > 0$*. Then we have*

$$\Pr\Big(\max_{\alpha\in\bar{\mathcal{D}}^*} \hat{u}^*_\alpha < \min_{\alpha\in\mathcal{D}^*} \hat{u}^*_\alpha\Big) \geq 1 - O[p\{\exp(-c_1 n^{1-2\nu})\}]. \tag{4}$$

Under conditions (C1) *and* (C2*)*, we have that*

$$\Pr(\mathcal{D}^* \subseteq \hat{\mathcal{D}}^*) \geq 1 - O[s_n\{\exp(-c_1 n^{1-2\nu})\}], \tag{5}$$

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

The proof of Theorem 2 is in the Online Supplement. Combining this result with Theorem 1, we find that the screening property holds for the HR SVS_1 procedure. Again, we can handle the non-polynomial (NP) dimensionality of order $\ln p = o(n^{1-2\nu})$ with $\nu \in (0, 1/2)$. That is, if $\ln p = o(n^{1-2\nu})$ with $\nu \in (0, 1/2)$, then (4) indicates that the probability of $\max_{\alpha\in\bar{\mathcal{D}}^*} \hat{u}^*_\alpha < \min_{\alpha\in\mathcal{D}^*} \hat{u}^*_\alpha$ approaches 1 as $n \to \infty$, separating the active predictors from the inactive predictors. And (5), in such a case, indicates that the size of true active predictors can change with p, as it is no larger than p.

To study $\hat{u}^{**}_\alpha$, we focus on HR only, leaving the related conditions and results for DC in the Online Supplement. Note that $u^{**}_\alpha = \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y\}$ for $\alpha \in \{1, \ldots, p\}$. Let

$$\hat{u}^{**}_\alpha = \hat{\mathcal{I}}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y\} = \sum_{s=1}^{S} \hat{u}^{\mathcal{J}}_{\alpha,s} = \sum_{s=1}^{S} \hat{\mathcal{I}}\{(X_\alpha, \mathbf{X}_{-\alpha})|Z = s\}.$$

Let $\mathcal{D}^{**}$ be the set of predictors whose $u^{**}_\alpha > 0$. We need replace condition (C2) by the following condition.

(C2**): $\min_{\alpha\in\mathcal{D}^{**}} u^{**}_\alpha \geq 2cn^{-\nu}$ for some $\nu \in [0, 1/2)$ and $c > 0$.

Condition (C2**) reflects that the resulting HR for active variables cannot be too small. Additionally, we need the following conditions.

(C3**): Observations within each slice are iid.
(C4**): For any interval $[b_1, b_2)$, we have

$$\inf_{y\in[b_1,b_2)} \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y=y\} \le \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y\in[b_1,b_2)\} \le \sup_{y\in[b_1,b_2)} \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y=y\}.$$

Furthermore, for any $\epsilon > 0$, if

$$1/S - \epsilon \le \Pr\{Y \in [b_1, b_2)\} \le 1/S + \epsilon,$$

then for any $y_1, y_2 \in [b_1, b_2)$,

$$|\mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y=y_1\} - \mathcal{I}\{(X_\alpha, \mathbf{X}_{-\alpha})|Y=y_2\}| \le \epsilon/2.$$

Note that although assumption (C3**) appears strong, it is in fact quite common in different areas, e.g., in sufficient dimension reduction [6,7,14,22] and in screening methods [26]. This condition ensures the rank consistency of $\hat{u}^{\mathcal{J}}_{\alpha,s}$ in each slice, which also simplifies the proof. Condition (C4**) is equivalent to the condition (C2) in [26], which will be used to provide us the exponential inequality. Condition (C4**) is to make sure that $\mathcal{I}\{(\mathbf{X}_{-\alpha}, X_\alpha)|Y \in [b_1, b_2)\}$ approximates the goal $\mathcal{I}\{(\mathbf{X}_{-\alpha}, X_\alpha)|Y=y\}$ accurately. Define the select set

$$\hat{\mathcal{D}}^{**} = \{\alpha : \hat{u}^{**}_\alpha \ge cn^{-\nu}\}.$$

We have the following result.

Theorem 3. *Under conditions* (C1), (C3**), *and* (C4**) *with* $\nu \in (0, 1/2)$, *there exist positive constants* $c_1, c_2 > 0$ *such that*

$$\Pr\Big(\max_{1\le\alpha\le p} |\hat{u}^{**}_\alpha - u^{**}_\alpha| \ge cn^{-\nu}\Big) \le O[Sp\{\exp(-c_1 n^{1-2\nu}/S^2) + \exp(-c_2 n^{1-2\nu}/S^2)\}]. \tag{6}$$

Furthermore, under condition (C2**), *denote* $\delta = \min_{\alpha\in\mathcal{D}^{**}} u^{**}_\alpha - \max_{\alpha\in\tilde{\mathcal{D}}^{**}} u^{**}_\alpha$. *Then we have*

$$\Pr\Big(\max_{\alpha\in\tilde{\mathcal{D}}^{**}} \hat{u}^{**}_\alpha < \min_{\alpha\in\mathcal{D}^{**}} \hat{u}^{**}_\alpha\Big) \ge 1 - O[Sp\{\exp(-c_1 n^{1-2\nu}/S^2) + \exp(-c_2 n^{1-2\nu}/S^2)\}]. \tag{7}$$

Under conditions (C1), (C2**), (C3**), *and* (C4**), *we have that*

$$\Pr(\mathcal{D}^{**} \subseteq \hat{\mathcal{D}}^{**}) \ge 1 - O[s_n S\{\exp(-c_1 n^{1-2\nu}/S^2) + \exp(-c_2 n^{1-2\nu}/S^2)\}], \tag{8}$$

where s_n *is the cardinality of* $\mathcal{D}^{**}$, *and* S *is the slice number of* Y.

Note first that we can simplify (6)–(8) by getting rid of the larger of c_1 and c_2 in the right-hand side of (6)–(8). The proof of Theorem 3 is in the Online Supplement. Combining this result with the result of Theorem 1, we conclude that the screening property holds for SVS_2. In fact, S does not have to be fixed. For instance, if $S \le \ln n$ [26], if $\ln p = o(n^{1-2\nu}/S^2)$ with $\nu \in (0, 1/2)$, then (7) indicates that the probability of $\max_{\alpha\in\tilde{\mathcal{D}}^{**}} \hat{u}^{**}_\alpha < \min_{\alpha\in\mathcal{D}^{**}} \hat{u}^{**}_\alpha$ approaches 1 as $n \to \infty$, separating the active predictors from the inactive predictors. And (8), in such a case, indicates that the size of true active predictors can change with p, as it is no larger than p.

4. Numerical studies

In this section, we assess the performance of our proposed approaches through simulation studies. We repeat each experiment 200 times, and report the results based on the following criteria.

(i) $\mathcal{S}$: the minimum size to include all active predictors. We report the 5%, 25%, 50%, 75%, 95% quantiles $\mathcal{S}$ of out of 200 replications.
(ii) $\mathcal{P}_s$: the proportion that an individual active predictor is selected out of 200 replicates.
(iii) $\mathcal{P}_a$: the proportion that all active predictors are selected out of 200 replicates.

Note that results are better if $\mathcal{P}_a$ and $\mathcal{P}_s$ are closer to 1. We use DC, DC_1, DC_2 to represent DC based on marginal procedure, one stage procedure and two stage procedure, respectively, while HR, HR_1 and HR_2 are correspondingly based on HR.

Example 1. The models come from Fan and Lv [11] and Li et al. [23]:

(1.a) $Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + c_4\beta_4 X_{22} + \epsilon$;
(1.b) $Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) + c_4\beta_3 X_{22} + \epsilon$;
(1.c) $Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) X_{22} + \epsilon$;
(1.d) $Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + \exp(c_4|X_{22}|)\epsilon$,

Table 1
Example 1: $\mathcal{P}_s$ and $\mathcal{P}_a$.

Model		$r = 0.5$					$r = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
		X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
$n = 100, p = 30, d = \lfloor p/\ln(p) \rfloor$											
	DC	1.000	0.960	0.890	0.990	0.840	1.000	0.975	0.360	0.975	0.325
(1.*a*)	DC_1	1.000	0.955	0.910	0.990	0.855	0.995	0.970	0.395	0.960	0.345
	DC_2	1.000	0.945	0.875	0.990	0.810	0.995	0.965	0.415	0.940	0.350
	HR	0.995	0.935	0.885	0.985	0.810	0.995	0.975	0.465	0.945	0.405
	HR_1	0.995	0.895	0.850	0.980	0.730	0.995	0.955	0.400	0.915	0.310
	HR_2	0.995	0.900	0.850	0.980	0.735	0.990	0.955	0.420	0.915	0.320
	DC	0.640	0.625	0.975	1.000	0.465	0.605	0.625	0.670	0.985	0.310
(1.*b*)	DC_1	0.605	0.610	0.970	1.000	0.420	0.530	0.575	0.710	0.985	0.270
	DC_2	0.570	0.550	0.960	1.000	0.360	0.515	0.535	0.630	0.985	0.205
	HR	0.760	0.700	0.955	1.000	0.550	0.775	0.790	0.730	0.970	0.475
	HR_1	0.700	0.645	0.945	1.000	0.465	0.735	0.745	0.665	0.965	0.400
	HR_2	0.705	0.645	0.950	1.000	0.475	0.725	0.745	0.675	0.965	0.405
	DC	0.910	0.835	0.805	1.000	0.595	0.840	0.850	0.255	1.000	0.155
(1.*c*)	DC_1	0.880	0.830	0.745	1.000	0.515	0.800	0.795	0.230	1.000	0.120
	DC_2	0.875	0.800	0.755	1.000	0.500	0.795	0.780	0.310	1.000	0.190
	HR	0.950	0.910	0.865	1.000	0.740	0.935	0.930	0.475	1.000	0.390
	HR_1	0.925	0.900	0.845	1.000	0.685	0.910	0.910	0.420	1.000	0.335
	HR_2	0.925	0.900	0.840	1.000	0.680	0.910	0.910	0.430	1.000	0.345
	DC	1.000	0.895	0.835	0.995	0.745	0.960	0.910	0.530	0.880	0.390
(1.*d*)	DC_1	0.995	0.880	0.830	0.985	0.705	0.955	0.905	0.545	0.855	0.385
	DC_2	0.995	0.870	0.830	0.970	0.685	0.950	0.895	0.535	0.840	0.365
	HR	0.980	0.880	0.820	1.000	0.730	0.940	0.885	0.540	0.945	0.445
	HR_1	0.975	0.855	0.810	0.995	0.680	0.950	0.875	0.510	0.920	0.390
	HR_2	0.975	0.860	0.815	0.995	0.685	0.950	0.875	0.530	0.920	0.405
$n = 100, p = 300, d = \lfloor n/\ln(n) \rfloor$											
	DC	1.000	0.980	0.915	0.995	0.890	1.000	1.000	0.585	1.000	0.585
(1.*a*)	DC_1	1.000	0.975	0.925	0.995	0.895	1.000	1.000	0.570	1.000	0.570
	DC_2	1.000	0.975	0.910	0.995	0.880	1.000	1.000	0.550	1.000	0.550
	HR	0.995	0.930	0.875	0.990	0.810	0.995	0.995	0.600	0.985	0.585
	HR_1	0.995	0.920	0.865	0.990	0.790	0.995	0.990	0.585	0.975	0.555
	HR_2	0.995	0.920	0.865	0.990	0.790	0.995	0.990	0.575	0.975	0.545
	DC	0.590	0.585	0.985	0.995	0.445	0.905	0.885	0.875	0.995	0.735
(1.*b*)	DC_1	0.570	0.590	0.985	0.995	0.435	0.875	0.870	0.860	0.990	0.690
	DC_2	0.570	0.560	0.985	0.995	0.420	0.875	0.865	0.855	0.990	0.685
	HR	0.655	0.680	0.985	0.995	0.505	0.930	0.910	0.880	0.985	0.760
	HR_1	0.620	0.655	0.980	0.995	0.475	0.930	0.890	0.875	0.980	0.725
	HR_2	0.620	0.655	0.980	0.995	0.475	0.930	0.890	0.880	0.980	0.735
	DC	0.820	0.840	0.655	0.995	0.420	0.965	0.960	0.595	0.995	0.545
(1.*c*)	DC_1	0.810	0.820	0.635	0.995	0.395	0.955	0.955	0.530	0.995	0.480
	DC_2	0.810	0.815	0.635	0.995	0.390	0.955	0.955	0.535	0.995	0.485
	HR	0.910	0.890	0.745	1.000	0.600	0.985	0.990	0.670	0.985	0.635
	HR_1	0.895	0.890	0.740	1.000	0.580	0.985	0.985	0.635	0.985	0.595
	HR_2	0.895	0.890	0.740	1.000	0.580	0.985	0.985	0.630	0.985	0.590
	DC	0.970	0.840	0.720	0.960	0.610	0.985	0.955	0.570	0.945	0.510
(1.*d*)	DC_1	0.965	0.835	0.695	0.955	0.580	0.980	0.940	0.550	0.935	0.470
	DC_2	0.965	0.835	0.710	0.955	0.590	0.980	0.940	0.560	0.935	0.480
	HR	0.935	0.805	0.740	0.965	0.585	0.955	0.910	0.575	0.955	0.505
	HR_1	0.930	0.775	0.725	0.960	0.550	0.950	0.905	0.555	0.955	0.480
	HR_2	0.930	0.775	0.740	0.960	0.560	0.950	0.905	0.560	0.955	0.485

where $\mathbf{1}$ is an indicator function. Models (1.b) and (1.c) contain an interaction term X_1X_2, and model (1.d) is heteroscedastic. Following [11,23], we choose $\beta_j = (-1)^{\xi}(4 \ln n/\sqrt{n} + |\eta|)$ for $j \in \{1, 2, 3, 4\}$, where $\xi \sim$ Bernoulli(0.4) and $\eta \sim \mathcal{N}(0, 1)$. We set $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$. The predictor vector is $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = r^{|i-j|}$ in two cases: (a) $r = 0.5$, (b) $r = 0.8$. The error term is $\epsilon \sim \mathcal{N}(0, 1)$.

We set $n = 100$, $p = 30$, $d = \lfloor p/\ln(p) \rfloor$ and $n = 100$, $p = 300$, $d = \lfloor n/\ln(n) \rfloor$, respectively. Tables 1 and 2 report the results. Table 1 indicates that the HR approach has better performance than DC does in almost all cases when $n > p$. For $n < p$, the DC approach seems to have comparable/slightly better performance than HR does, but the DC approach is computationally more efficient than HR. Thus we recommend to use HR when $n > p$, and to use DC when $n < p$.

Table 2
Example 1: The quantiles of the minimum model size S out of 200 replications.

		$r = 0.5$					$r = 0.8$				
S		5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$n = 100, p = 30$											
	DC	4.000	4.000	5.000	7.000	13.000	5.000	8.000	11.000	17.000	25.050
(1.*a*)	HR	4.000	4.000	5.000	8.000	16.050	5.000	7.000	10.000	17.000	26.000
	DC	5.000	6.000	9.000	13.000	22.100	5.000	8.000	10.000	14.000	19.000
(1.*b*)	HR	4.000	5.000	8.000	11.000	21.100	4.000	6.000	9.000	12.000	20.050
	DC	4.000	6.000	8.000	10.000	19.000	7.000	10.000	12.000	15.250	23.000
(1.*c*)	HR	4.000	4.000	6.000	9.000	18.000	5.000	7.000	10.000	14.000	23.050
	DC	4.000	4.000	6.000	9.000	20.000	4.950	6.000	10.000	15.000	25.000
(1.*d*)	HR	4.000	4.000	5.000	10.000	21.050	4.000	7.000	9.000	17.000	25.000
$n = 100, p = 300$											
	DC	4.000	4.000	5.000	10.000	53.000	5.000	9.000	16.000	43.000	147.050
(1.*a*)	HR	4.000	4.000	6.000	14.000	94.100	4.000	8.000	18.000	52.000	169.550
	DC	5.000	11.000	30.000	63.000	165.300	5.000	8.000	12.000	22.000	82.000
(1.*b*)	HR	4.000	8.000	20.500	55.250	173.050	4.000	7.000	11.000	21.000	70.200
	DC	6.000	13.000	25.000	55.500	156.200	7.000	13.000	20.000	38.250	89.450
(1.*c*)	HR	4.000	7.000	15.000	44.250	137.450	5.000	7.000	15.000	36.000	112.100
	DC	4.000	7.000	13.500	45.250	146.300	5.000	9.750	21.000	69.750	156.050
(1.*d*)	HR	4.000	6.000	15.000	55.500	192.400	4.950	7.000	21.000	66.250	194.000

Table 2 reports model sizes for Example 1. The fact that HR is better than DC for $n > p$ is mainly due to the way the distance is computed: DC only calculates the Euclidean distance, while HR uses the kernel (smoothing) to calculate the distance. When $n > p$, a kernel method is usually finer than the Euclidean distance. When $n < p$, a kernel method may not be good enough for smoothing as the sample size is not large enough.

We compare our methods with the following approaches from the literature: (i) the Lasso [37], whose size is determined by 10-fold cross-validation; (ii) ISIS [11]; (iii) QaSIS [19], the quantile adaptive model-free variable screening, which is proposed to estimate marginal quantile regression (denoted by Q) nonparametrically using a B-spline approximation, considering that Y and X_j are independent if and only if $Q_a(Y|X_j) - Q_a(Y) = 0$; (iv) NIS [9], the nonparametric independence screening in sparse ultrahigh-dimensional additive models using a B-spline basis; (v) SIRS; (vi) ISIRS [41]; and (vii)"DC-seq" [21], a sequential method based on DC without specifying the model size d. Results corresponding to these existing methods are displayed in Table 3; their performances are not as stable as DC and HR.

We also consider the case of $n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$. The simulation results are in Table 4. Additional simulation results for Example 1 are in the Online Supplement. Example 1 indicates that FS_M, SVS_1, and SVS_2 have similar performance. However, the next example shows the advantages of SVS_1, and SVS_2.

Example 2. (2.a) This is Example 4 in [41]: $Y = \beta^\top \mathbf{X} + \epsilon$. Set $\epsilon \sim \mathcal{N}(0, 1)$ and $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$, σ_{ij} satisfies $\sigma_{ii} = 1$ for $i \in \{1, \ldots, p\}$, $\sigma_{i4} = \sigma_{4j} = \rho^{1/2}$ for $i \neq 4$, and $\sigma_{ij} = \rho$ for $i \neq j, i \neq 4, j \neq 4$. In the model, all predictors except for X_4 are equally correlated with coefficient ρ, while X_4 has correlation $\rho^{1/2}$ with all other $p - 1$ predictors. Let $\beta = (5, 5, 5, -15\rho^{1/2}, 0, \ldots, 0)$. Although X_4 is marginal independent of Y, it is an active predictor when $\rho \neq 0$.

(2.b) A nonlinear model $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4^3 + \epsilon$. Set $\epsilon \sim \mathcal{N}(0, 1)$ and $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$, and σ_{ij} satisfies $\sigma_{ii} = 1$ for $i \in \{1, \ldots, p\}$, $\sigma_{i4} = \sigma_{4j} = \rho^{7/10}$ for $i \neq 4$, and $\sigma_{ij} = \rho$ for $i \neq j, i \neq 4, j \neq 4$. Let $\beta = (5, 5, 5, -3\rho^{7/10}, 0, \ldots, 0)$.

(2.c) A nonlinear model $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 \exp X_4 + \epsilon$. Set $\epsilon \sim \mathcal{N}(0, 1)$ and $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$, and σ_{ij} satisfies $\sigma_{ii} = 1$ for $i \in \{1, \ldots, p\}$, $\sigma_{i4} = \sigma_{4j} = \rho^{7/10}$ for $i \neq 4$, and $\sigma_{ij} = \rho$ for $i \neq j, i \neq 4, j \neq 4$. Let $\beta = (5, 5, 5, -15\rho^{7/10} \exp(-1/2), 0, \ldots, 0)$.

In Example 2, X_4 is marginally independent of Y; however, it is an active predictor when $\rho \neq 0$. In particular, the relationship between the active predictor X_4 and Y is linear in Example (2.a). The relationship in Example (2.b) and (2.c) is nonlinear. We consider the cases of $n = 100$, $p = 300$, $d = \lfloor n/\ln(n) \rfloor$ and $n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$. Tables 5 and 6 clearly indicate that the results of SVS_1, SVS_2 based on DC significantly improve the marginal solution. QaSIS and NIS have poor performance, although ISIS and ISIRS have better performance, in general are still not comparable with DC_1 and DC_2. In the Online Supplement, we provide further simulation results for Example 2.

The difference between our two examples is that in Example 1, all the active predictors are marginally dependent of Y, while in Example 2 the active predictor X_4 is marginally independent of Y. Therefore, marginal screening methods fail, but our new approaches will detect X_4.

Table 3
Example 1: $\mathcal{P}_s$ and $\mathcal{P}_a$ for some existing methods.

Model		$r = 0.5$					$r = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
		X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
$n = 100, p = 30, d = \lfloor p/\ln(p) \rfloor$											
	Lasso	1.000	1.000	1.000	1.000	1.000	1.000	0.980	1.000	1.000	0.980
(1.*a*)	ISIS	1.000	1.000	1.000	1.000	1.000	1.000	0.880	0.995	1.000	0.875
	QaSIS	1.000	0.895	0.840	0.980	0.720	0.990	0.945	0.425	0.925	0.350
	NIS	1.000	0.975	0.885	0.995	0.855	0.995	0.995	0.435	0.985	0.425
	SIRS	1.000	0.965	0.875	1.000	0.840	1.000	0.985	0.230	0.975	0.205
	ISIRS	1.000	0.855	1.000	1.000	0.855	0.995	0.850	0.910	0.970	0.725
	DC-seq	0.960	0.135	0.365	0.800	0.075	0.885	0.295	0.065	0.500	0.005
	Lasso	0.235	0.230	0.985	0.995	0.075	0.300	0.210	0.900	0.980	0.085
(1.*b*)	ISIS	0.780	0.830	0.995	1.000	0.625	0.805	0.765	0.960	0.980	0.570
	QaSIS	0.820	0.785	0.910	0.985	0.590	0.960	0.945	0.645	0.925	0.500
	NIS	0.905	0.855	0.940	1.000	0.740	0.990	0.975	0.595	0.930	0.515
	SIRS	0.085	0.070	0.975	1.000	0.010	0.005	0.005	0.710	0.995	0.000
	ISIRS	0.315	0.270	0.990	1.000	0.115	0.115	0.105	0.990	0.995	0.050
	DC-seq	0.085	0.075	0.600	0.960	0.030	0.145	0.155	0.325	0.905	0.030
	Lasso	0.160	0.155	0.090	0.965	0.010	0.190	0.125	0.080	0.940	0.020
(1.*c*)	ISIS	0.850	0.775	0.790	0.990	0.530	0.820	0.835	0.780	0.980	0.525
	QaSIS	0.965	0.965	0.295	0.985	0.255	0.995	0.990	0.080	0.960	0.075
	NIS	0.980	0.945	0.095	0.990	0.085	0.995	0.980	0.030	0.965	0.025
	SIRS	0.140	0.115	0.740	0.995	0.015	0.035	0.010	0.265	1.000	0.000
	ISIRS	0.265	0.295	0.800	0.990	0.055	0.190	0.165	0.675	0.995	0.050
	DC-seq	0.395	0.395	0.200	0.975	0.055	0.430	0.410	0.010	0.915	0.005
	Lasso	0.490	0.215	0.290	0.350	0.075	0.555	0.255	0.255	0.350	0.070
(1.*d*)	ISIS	0.990	0.925	0.955	0.970	0.845	0.985	0.940	0.980	0.970	0.875
	QaSIS	1.000	0.875	0.845	0.755	0.535	0.980	0.965	0.655	0.665	0.405
	NIS	0.680	0.440	0.485	0.850	0.125	0.560	0.465	0.320	0.800	0.055
	SIRS	1.000	0.985	0.895	0.195	0.175	0.995	0.995	0.705	0.105	0.050
	ISIRS	1.000	0.895	0.955	0.335	0.280	0.995	0.975	0.855	0.200	0.145
	DC-seq	0.975	0.315	0.485	0.645	0.190	0.885	0.410	0.170	0.450	0.070
$n = 100, p = 300, d = \lfloor n/\ln(n) \rfloor$											
	Lasso	1.000	1.000	1.000	1.000	1.000	1.000	0.960	1.000	1.000	0.960
(1.*a*)	ISIS	1.000	0.975	1.000	1.000	0.975	1.000	0.870	1.000	1.000	0.870
	QaSIS	0.995	0.895	0.785	0.980	0.690	0.990	0.985	0.510	0.950	0.450
	NIS	1.000	0.975	0.880	0.995	0.850	1.000	1.000	0.630	0.995	0.625
	SIRS	1.000	0.990	0.860	1.000	0.850	1.000	1.000	0.455	1.000	0.455
	ISIRS	1.000	0.950	1.000	1.000	0.950	1.000	0.990	0.995	1.000	0.985
	DC-seq	0.965	0.160	0.310	0.820	0.080	0.920	0.310	0.040	0.515	0.015
	Lasso	0.145	0.140	0.945	0.990	0.020	0.130	0.130	0.855	0.970	0.030
(1.*b*)	ISIS	0.145	0.145	0.955	0.995	0.045	0.110	0.125	0.840	0.975	0.035
	QaSIS	0.725	0.700	0.940	0.990	0.515	0.980	0.975	0.745	0.955	0.655
	NIS	0.860	0.860	0.940	0.990	0.705	1.000	1.000	0.770	0.965	0.740
	SIRS	0.040	0.025	0.980	1.000	0.000	0.035	0.020	0.865	1.000	0.010
	ISIRS	0.125	0.150	0.990	1.000	0.045	0.120	0.080	0.950	0.995	0.045
	DC-seq	0.070	0.065	0.645	0.965	0.015	0.205	0.180	0.365	0.905	0.020
	Lasso	0.090	0.070	0.015	0.925	0.000	0.075	0.100	0.025	0.855	0.000
(1.*c*)	ISIS	0.165	0.125	0.040	0.965	0.005	0.120	0.150	0.070	0.915	0.010
	QaSIS	0.920	0.945	0.160	0.955	0.130	1.000	1.000	0.065	0.960	0.065
	NIS	0.910	0.930	0.040	0.980	0.040	0.995	1.000	0.035	0.945	0.035
	SIRS	0.060	0.030	0.415	1.000	0.000	0.035	0.025	0.335	1.000	0.010
	ISIRS	0.100	0.060	0.525	0.995	0.015	0.095	0.100	0.415	0.980	0.030
	DC-seq	0.280	0.295	0.120	0.985	0.040	0.510	0.445	0.010	0.870	0.005
	Lasso	0.305	0.085	0.105	0.215	0.000	0.360	0.170	0.115	0.305	0.015
(1.*d*)	ISIS	0.470	0.245	0.200	0.535	0.020	0.520	0.425	0.210	0.515	0.025
	QaSIS	0.990	0.795	0.755	0.680	0.390	0.980	0.965	0.535	0.625	0.330
	NIS	0.365	0.195	0.220	0.760	0.030	0.435	0.375	0.200	0.750	0.065
	SIRS	1.000	0.960	0.820	0.115	0.075	1.000	1.000	0.625	0.120	0.075
	ISIRS	0.995	0.915	0.835	0.165	0.130	1.000	0.990	0.825	0.135	0.105
	DC-seq	0.945	0.365	0.460	0.630	0.135	0.885	0.410	0.180	0.405	0.030

Table 4
Example 1: $\mathcal{P}_s$ and $\mathcal{P}_a$ for DC's and some existing methods.

Model		$r = 0.5$					$r = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
		X_1	X_2	X_{12}	X_{22}	All	X_1	X_2	X_{12}	X_{22}	All
$n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$											
	DC	1.000	1.000	0.965	1.000	0.965	1.000	1.000	0.740	1.000	0.740
(1.*a*)	DC_1	1.000	1.000	0.965	1.000	0.965	1.000	1.000	0.730	1.000	0.730
	DC_2	1.000	1.000	0.965	1.000	0.965	1.000	1.000	0.730	1.000	0.730
	DC	0.750	0.715	0.990	1.000	0.605	0.990	0.985	0.925	1.000	0.905
(1.*b*)	DC_1	0.755	0.710	0.990	1.000	0.610	0.990	0.985	0.920	1.000	0.900
	DC_2	0.750	0.710	0.990	1.000	0.605	0.990	0.985	0.920	1.000	0.900
	DC	0.880	0.920	0.780	1.000	0.615	1.000	1.000	0.765	1.000	0.765
(1.*c*)	DC_1	0.875	0.920	0.775	1.000	0.605	1.000	1.000	0.765	1.000	0.765
	DC_2	0.875	0.920	0.775	1.000	0.605	1.000	1.000	0.765	1.000	0.765
	DC	1.000	0.945	0.845	1.000	0.805	0.995	0.975	0.660	0.995	0.640
(1.*d*)	DC_1	1.000	0.945	0.845	0.995	0.800	0.995	0.975	0.655	0.995	0.635
	DC_2	1.000	0.945	0.845	0.995	0.800	0.995	0.975	0.655	0.995	0.635
$n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$											
	Lasso	1.000	1.000	1.000	1.000	1.000	1.000	0.98	1.000	1.000	0.980
(1.*a*)	ISIS	1.000	1.000	1.000	1.000	1.000	1.000	0.905	1.000	1.000	0.905
	QaSIS	1.000	0.925	0.890	0.995	0.820	1.000	0.995	0.640	0.985	0.625
	NIS	1.000	0.995	0.955	1.000	0.950	1.000	1.000	0.805	1.000	0.805
	SIRS	1.000	1.000	0.940	1.000	0.940	1.000	1.000	0.575	1.000	0.575
	ISIRS	1.000	0.990	1.000	1.000	0.990	1.000	1.000	1.000	1.000	1.000
	DC-seq	0.940	0.070	0.360	0.800	0.040	0.870	0.185	0.050	0.520	0.000
	Lasso	0.045	0.030	0.975	1.000	0.005	0.090	0.075	0.890	0.985	0.020
(1.*b*)	ISIS	0.060	0.035	0.975	1.000	0.000	0.085	0.070	0.860	0.985	0.015
	QaSIS	0.875	0.890	0.980	1.000	0.775	1.000	1.000	0.860	0.995	0.855
	NIS	0.965	0.980	0.955	1.000	0.905	1.000	1.000	0.850	0.995	0.850
	SIRS	0.015	0.030	0.990	1.000	0.005	0.010	0.015	0.880	1.000	0.000
	ISIRS	0.045	0.050	0.990	1.000	0.000	0.060	0.060	0.965	1.000	0.020
	DC-seq	0.035	0.030	0.655	0.980	0.010	0.125	0.145	0.305	0.930	0.020
	Lasso	0.035	0.030	0.025	0.975	0.000	0.065	0.045	0.015	0.955	0.000
(1.*c*)	ISIS	0.070	0.030	0.035	0.990	0.000	0.065	0.060	0.045	0.970	0.005
	QaSIS	0.995	0.990	0.140	1.000	0.140	1.000	1.000	0.110	1.000	0.110
	NIS	0.985	0.995	0.025	0.980	0.025	1.000	1.000	0.035	0.985	0.035
	SIRS	0.010	0.010	0.575	1.000	0.000	0.005	0.000	0.555	1.000	0.000
	ISIRS	0.035	0.055	0.685	1.000	0.005	0.050	0.050	0.520	1.000	0.000
	DC-seq	0.245	0.225	0.100	0.980	0.010	0.390	0.370	0.000	0.955	0.000
	Lasso	0.205	0.055	0.060	0.170	0.000	0.245	0.100	0.045	0.185	0.000
(1.*d*)	ISIS	0.375	0.190	0.120	0.405	0.005	0.430	0.290	0.135	0.475	0.040
	QaSIS	1.000	0.925	0.855	0.630	0.485	0.995	0.985	0.710	0.560	0.415
	NIS	0.230	0.095	0.110	0.675	0.000	0.295	0.195	0.080	0.695	0.010
	SIRS	1.000	0.985	0.900	0.030	0.030	1.000	1.000	0.745	0.065	0.060
	ISIRS	1.000	0.975	0.885	0.095	0.090	1.000	1.000	0.865	0.095	0.095
	DC-seq	0.980	0.290	0.500	0.725	0.160	0.920	0.325	0.190	0.440	0.035

4.1. Cardiomyopathy data

We analyze the cardiomyopathy microarray data which was considered in [18,23,29]. In the analysis of these data, the aim is to identify the most influential genes for over-expression of a G protein-coupled receptor, designated Ro1, in mice. The response Y is the Ro1 expression level, and the predictors X_α are other genetic expression levels. Our goal is to identify the most influential genes for Ro1 with sample size $n = 30$, $p = 6319$. Because $n \ll p$, we choose the DC approach. We use the threshold value $d = \lfloor n/\ln(n) \rfloor = 8$.

The DC and DC_1 procedures select the same eight genes: Msa.2134.0, Msa.2877.0, Msa.26025.0, Msa.5583.0, Msa.1590.0, Msa.1166.0, Msa.2400.0, Msa.15442.0. In contrast, the DC_2 procedure selects slightly different genes: Msa.2134.0, Msa.2877.0, Msa.26025.0, Msa.5583.0, Msa.1590.0, Msa.1166.0, Msa.2400.0, Msa.5618.0.

Note that the top two genes, viz. Msa.2134.0 and Msa.2877.0, were also selected by Li et al. [23], and two genes, viz. Msa.2877.0 and Msa.1166.0, were selected by Hall and Miller [18]. Both sets are in our final results. However, compared with DC (DC_1), the DC_2 procedure selects a different gene, viz. Msa.5618.0. To identify whether the gene Msa.5618.0 is important, following Li et al. [23] we consider an additive model which always includes the top two genes Msa.2134.0,

Table 5
Example 2: $\mathcal{P}_s$ and $\mathcal{P}_a$ for DC's and some existing methods.

Model		$r = 0.5$					$r = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
		X_1	X_2	X_3	X_4	All	X_1	X_2	X_3	X_4	All
$n = 100$, $p = 300$, $d = \lfloor n/\ln(n) \rfloor$											
	DC	0.945	0.960	0.960	0.005	0.005	0.830	0.815	0.815	0.000	0.000
(2.*a*)	DC_1	0.940	0.960	0.960	1.000	0.905	0.820	0.815	0.810	1.000	0.750
	DC_2	0.940	0.960	0.960	1.000	0.905	0.825	0.825	0.810	1.000	0.750
	DC	0.995	0.995	1.000	0.235	0.230	0.980	0.995	0.995	0.355	0.340
(2.*b*)	DC_1	0.995	0.995	1.000	0.930	0.920	0.975	0.990	0.995	0.975	0.935
	DC_2	0.995	0.995	1.000	0.990	0.980	0.975	0.990	0.995	0.980	0.940
	DC	1.000	0.995	0.995	0.175	0.175	0.970	0.960	0.940	0.275	0.195
(2.*c*)	DC_1	0.995	0.995	0.995	0.970	0.955	0.970	0.960	0.935	0.985	0.850
	DC_2	0.995	0.995	0.995	1.000	0.985	0.970	0.960	0.935	1.000	0.865
$n = 100$, $p = 300$, $d = \lfloor n/\ln(n) \rfloor$											
	Lasso	1.000	1.000	1.000	0.005	0.005	1.000	1.000	1.000	0.000	0.000
(2.*a*)	ISIS	1.000	1.000	1.000	0.995	0.995	1.000	1.000	1.000	0.995	0.995
	QaSIS	0.870	0.895	0.905	0.045	0.035	0.510	0.555	0.470	0.020	0.000
	NIS	0.975	0.975	0.980	0.030	0.030	0.800	0.780	0.805	0.015	0.000
	SIRS	0.990	0.975	1.000	0.000	0.000	0.930	0.895	0.915	0.000	0.000
	ISIRS	1.000	1.000	1.000	0.920	0.920	1.000	1.000	1.000	0.930	0.930
	DC-seq	0.835	0.765	0.760	0.000	0.000	0.510	0.485	0.490	0.000	0.000
	Lasso	1.000	1.000	1.000	0.945	0.945	0.860	0.845	0.875	0.365	0.330
(2.*b*)	ISIS	1.000	1.000	1.000	1.000	1.000	0.875	0.845	0.840	0.935	0.660
	QaSIS	0.945	0.945	0.970	0.080	0.060	0.815	0.815	0.810	0.120	0.070
	NIS	0.995	0.995	0.995	0.195	0.195	0.955	0.925	0.890	0.545	0.425
	SIRS	1.000	1.000	0.990	0.005	0.005	0.920	0.920	0.905	0.020	0.020
	ISIRS	0.980	0.975	0.985	1.000	0.945	0.850	0.835	0.830	0.930	0.560
	DC-seq	0.975	0.965	0.955	0.000	0.000	0.930	0.910	0.890	0.000	0.000
	Lasso	0.960	0.945	0.955	0.970	0.915	0.495	0.525	0.475	0.430	0.235
(2.*c*)	ISIS	0.970	0.930	0.975	1.000	0.895	0.610	0.560	0.540	0.915	0.230
	QaSIS	0.940	0.925	0.920	0.600	0.470	0.245	0.295	0.255	0.765	0.020
	NIS	0.830	0.815	0.810	0.885	0.595	0.215	0.260	0.220	0.985	0.060
	SIRS	0.995	0.985	0.995	0.000	0.000	0.940	0.905	0.920	0.000	0.000
	ISIRS	0.985	0.970	0.975	1.000	0.930	0.855	0.830	0.815	1.000	0.570
	DC-seq	0.960	0.930	0.940	0.035	0.030	0.825	0.740	0.760	0.065	0.005

Msa.2877.0 and an additional gene, viz.

$$Y = f_1(X_1) + f_2(X_2) + f_3(X_k) + \epsilon,$$

where X_1 = Msa.2134.0 and X_2 = Msa.2877.0, and X_k is another gene which is selected by DC (DC_1) and DC_2. The p-values of X_1 and X_2 are much smaller than the 5% significance level; thus these two variables are significant. Table 7 reports the related p-values for other variables, which indicates that gene Msa.5618.0 is significant, and is worth further consideration, while all other genes are not significant.

5. Discussion

In this paper, we proposed two sufficient variable selection procedures, illustrated using DC and HR. The generality of these procedures is that any marginal screening approach may be adapted into the two sufficient procedures as to improve the marginal screening approach. Our new procedures SVS_1 and SVS_2 provide sufficient variable selection for the response that can provide additional information from predictors which may be missed by just using marginal screening methods, as demonstrated in Example 2. In the presence of multiple utilities, different strategies may be used to combine them. The idea of Balasubramanian et al. [1], which is based on the sup-HSIC method and takes the supremum of HSIC over a family of kernels via marginal screening, can certainly be incorporated into our sufficient variable selection procedures.

Acknowledgments

The authors would like to thank the Editor-in-Chief, Christian Genest, an Associate Editor and two referees for their valuable comments, which greatly improved the paper, and colleague Katherine Thompson for reading the final version. Yang's work is supported in part by National Natural Science Foundation of China (NSFC, No. 11501472). Yin's work is supported in part by an National Science Foundation (NSF, USA) CIF-1813330.

Table 6
Example 2: $\mathcal{P}_s$ and $\mathcal{P}_a$ for DC's and some existing methods.

Model		$r = 0.5$					$r = 0.8$				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
		X_1	X_2	X_3	X_4	All	X_1	X_2	X_3	X_4	All
$n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$											
	DC	0.995	1.000	1.000	0.000	0.000	0.935	0.965	0.960	0.000	0.000
(2.*a*)	DC_1	0.995	1.000	1.000	1.000	0.995	0.935	0.965	0.955	1.000	0.915
	DC_2	0.995	1.000	1.000	1.000	0.995	0.935	0.965	0.955	1.000	0.915
	DC	1.000	1.000	1.000	0.110	0.110	1.000	1.000	1.000	0.345	0.345
(2.*b*)	DC_1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DC_2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DC	1.000	1.000	1.000	0.080	0.080	0.995	0.985	1.000	0.180	0.170
(2.*c*)	DC_1	1.000	1.000	1.000	1.000	1.000	0.995	0.985	1.000	1.000	0.980
	DC_2	1.000	1.000	1.000	1.000	1.000	0.995	0.985	1.000	1.000	0.980
$n = 200$, $p = 2000$, $d = \lfloor n/\ln(n) \rfloor$											
	Lasso	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000
(2.*a*)	ISIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.525	0.525
	QaSIS	0.980	0.990	0.990	0.005	0.005	0.735	0.740	0.690	0.000	0.000
	NIS	1.000	1.000	1.000	0.005	0.005	0.950	0.940	0.955	0.000	0.000
	SIRS	1.000	1.000	1.000	0.000	0.000	0.985	0.990	0.985	0.000	0.000
	ISIRS	1.000	1.000	1.000	0.990	0.990	1.000	1.000	1.000	0.940	0.940
	DC-seq	0.945	0.910	0.915	0.000	0.000	0.615	0.625	0.655	0.000	0.000
	Lasso	1.000	1.000	1.000	1.000	1.000	0.965	0.935	0.935	0.475	0.475
(2.*b*)	ISIS	1.000	1.000	1.000	1.000	1.000	0.955	0.935	0.935	0.950	0.825
	QaSIS	1.000	1.000	1.000	0.055	0.055	0.915	0.915	0.925	0.030	0.020
	NIS	0.995	0.995	1.000	0.275	0.275	0.960	0.920	0.960	0.805	0.705
	SIRS	1.000	1.000	1.000	0.005	0.005	0.995	0.980	0.960	0.005	0.000
	ISIRS	1.000	1.000	1.000	1.000	1.000	0.975	0.940	0.940	0.995	0.855
	DC-seq	1.000	0.995	0.990	0.015	0.015	0.995	0.990	0.985	0.005	0.005
	Lasso	0.975	0.980	0.985	0.980	0.965	0.555	0.525	0.525	0.540	0.335
(2.*c*)	ISIS	0.975	0.985	0.990	0.995	0.965	0.625	0.600	0.630	0.945	0.300
	QaSIS	1.000	1.000	0.985	0.785	0.775	0.250	0.225	0.285	0.935	0.035
	NIS	0.835	0.825	0.810	0.985	0.740	0.140	0.110	0.100	0.995	0.015
	SIRS	1.000	1.000	1.000	0.000	0.000	0.990	0.980	0.970	0.000	0.000
	ISIRS	1.000	0.995	1.000	1.000	0.995	0.965	0.955	0.945	1.000	0.865
	DC-seq	1.000	1.000	1.000	0.000	0.000	0.930	0.960	0.915	0.025	0.010

Table 7
The selected genes in Cardiomyopathy data with $d = \lfloor n/\ln(n) \rfloor = 8$.

By both $DC(DC_1)$ and DC_2	X_α	Msa.26025.0	Msa.5583.0	Msa.1590.0	Msa.1166.0	Msa.2400.0
	p-values	0.093	0.071	0.121	0.591	0.121
By $DC(DC_1)$ only	X_α	Msa.15442.0				
	p-values	0.395				
By DC_2 only	X_α	Msa.5618.0				
	p-values	0.046				

References

[1] K. Balasubramanian, B.K. Sriperumbudur, G. Lebanon, Ultrahigh-dimensional feature screening via RKHS embeddings, in: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, 2013, pp. 126–134.
[2] E. Candès, T. Tao, The dantzig selector: Statistical estimation when p is much larger than n (with discussion), Ann. Statist. 35 (2007) 2313–2404.
[3] J. Chang, C.Y. Tang, Y. Wu, Marginal empirical likelihood and sure independence feature screening, Ann. Statist. 41 (2013) 2123–2148.
[4] J. Chang, C.Y. Tang, Y. Wu, Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood, Ann. Statist. 44 (2016) 515–539.
[5] R.D. Cook, Graphics for regression with binary response, J. Amer. Statist. Assoc. 91 (1996) 983–992.
[6] R.D. Cook, L. Ni, Sufficient dimension reduction via inverse regression: A minimum discrepancy approach, J. Amer. Statist. Assoc. 100 (2005) 410–428.
[7] R.D. Cook, S. Weisberg, Comment on sliced inverse regression for dimension reduction, J. Amer. Statist. Assoc. 86 (1991) 328–332.
[8] H. Cui, R. Li, W. Zhong, Model-free feature screening for ultrahigh-dimensional discriminant analysis, J. Amer. Statist. Assoc. 110 (2015) 630–641.
[9] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, J. Amer. Statist. Assoc. 106 (2011) 544–557.
[10] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[11] J. Fan, J. Lv, Sure independence screening for ultrahigh-dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008) 849–911.
[12] J. Fan, Y. Ma, W. Dai, Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models, J. Amer. Statist. Assoc. 109 (2014) 1270–1284.

[13] J. Fan, R. Song, Sure independence screening in generalized linear models with NP-dimensionality, Ann. Statist. 38 (2010) 3567–3604.
[14] A. Gannoun, J. Saracco, An asymptotic theory for SIR_{α} method, Statist. Sinica 13 (2003) 297–310.
[15] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with Hilbert-Schmidt norms, in: ALT, Springer-Verlag, Heidelberg, 2005, pp. 63–77.
[16] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, A.J. Smola, A kernel statistical test of independence, NeurIPS 20 (2008) 585–592.
[17] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, B. Schölkopf, Kernel methods for measuring independence, J. Mach. Learn. Res. 6 (2005) 2075–2129.
[18] P. Hall, H. Miller, Using generalized correlation to effect variable selection in very high dimensional problems, J. Comput. Graph. Statist. 18 (2009) 533–550.
[19] X. He, L. Wang, H.G. Hong, Quantile- adaptive model-free variable screening for high-dimensional heterogeneous data, Ann. Statist. 41 (2013) 342–369.
[20] A. Kim, S. Shin, The cumulative Kolmogorov filter for model-free screening in ultrahigh-dimensional data, Statist. Probab. Lett. 126 (2017) 238–243.
[21] J. Kong, S. Wang, G.Wahba, Using distance covariance for improved variable selection with application to learning genetic risk models, Stat. Med. 34 (2015) 1708–1720.
[22] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. 86 (1991) 316–342.
[23] R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, J. Amer. Statist. Assoc. 107 (2012) 1129–1139.
[24] J. Liu, R. Li, R. Wu, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, J. Amer. Statist. Assoc. 109 (2014) 266–274.
[25] Q. Mai, H. Zou, The Kolmogorov filter for variable screening in high dimensional binary classification, Biometrika 100 (2013) 229–234.
[26] Q. Mai, H. Zou, The fused Kolmogorov filter: A nonparametric model-free screening method, Ann. Statist. 43 (2015) 1471–1497.
[27] J. Schafer, K. Strimmer, A shrinkage approach to large-scale covariance estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol. 4 (2005) http://dx.doi.org/10.2202/1544-6115.1175.
[28] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, New York, 1992.
[29] M.R. Segal, K.D. Dahlquist, B.R. Conklin, Regression approach for microarray data analysis, J. Comput. Biol. 10 (2003) 961–980.
[30] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing, Ann. Statist. 41 (2013) 2263–2291.
[31] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.
[32] L. Song, A. Smola, A. Gretton, J. Bedo, K. Borgwardt, Feature selection via dependent maximization, J. Mach. Learn. Res. 13 (2012) 1393–1434.
[33] R. Song, F. Yi, H. Zou, On varying-coefficient independence screening for high-dimensional varying-coefficient models, Statist. Sinica 24 (2014) 1735–1752.
[34] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, G.R.G. Lanckriet, Hilbert Space embeddings and metrics on probability measures, J. Mach. Learn. Res. 11 (2010) 1517–1561.
[35] G.J. Székely, M.L. Rizzo, Brownian distance covariance, Ann. Appl. Stat. 3 (2009) 1236–1265.
[36] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing independence by correlation of distances, Ann. Statist. 35 (2007) 2769–2794.
[37] R.J. Tibshirani, Regression shirnkage and selection via lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288.
[38] X. Yin, H. Hilafu, Sequential sufficient dimension reduction for large p, small n problems, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (2015) 879–892.
[39] X. Yin, B. Li, R.D. Cook, Successive direction extraction for estimating the central subspace in a multiple-index regression, J. Multivariate Anal. 99 (2008) 1733–1757.
[40] N. Zhang, X. Yin, Direction estimation in single-index regressions via Hilbert-Schmidt independence criterion, Statist. Sinica 25 (2015) 743–758.
[41] L.P. Zhu, L. Li, R. Li, L.X. Zhu, Model-free feature screening for ultrahigh dimensional data, J. Amer. Statist. Assoc. 106 (2011) 1464–1475.