Statistica Sinica Preprint No: SS-2014-0117R3	
Title	Sure Independence Screening Adjusted for Confounding Covariates with Ultrahigh-dimensional Data
Manuscript ID	SS-2014-0117
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202014.0117
Complete List of Authors	Xueqin Wang Canhong Wen Wenliang Pan and Mian Huang
Corresponding Author	Xueqin Wang
E-mail	wangxq88@mail.sysu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Statistica Sinica

SURE INDEPENDENCE SCREENING ADJUSTED FOR CONFOUNDING COVARIATES WITH ULTRAHIGH DIMENSIONAL DATA

Canhong Wen[1], Wenliang Pan[1], Mian Huang[2], and Xueqin Wang[1,*]

[1]*Sun Yat-Sen University and* [2]*Shanghai University of Finance and Economics*

Abstract: Detecting candidate genetic variants in genomic studies often encounters confounding problems, particularly when the data are ultrahigh dimensional. Confounding covariates, such as age and gender, not only can reduce the statistical power, but also introduce spurious genetic association. How to control for the confounders in ultrahigh dimensional data analysis is a critical and challenging issue. In this paper, we propose a novel sure independence screening method based on conditional distance correlation under the ultrahigh dimensional model setting. Our proposal accomplishes the adjustment by conditioning on the confounding variables. With the model-free feature of conditional distance correlation, our method does not need any parametric modeling assumptions and is thus quite flexible. In addition, it is applicable to data with multivariate response. We show that under some mild technical conditions, the proposed method enjoys the sure screening property even when the dimensionality is an exponential order of the sample size. The simulation studies and a data analysis demonstrate that the proposed procedure has competitive performance.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Key words and phrases: Confounding; Feature Screening; Model Free; Multivariate Response; Ultrahigh Dimension.

1. Introduction

Identifying variants associated with common complex disease in ultra-high dimensional data is a central goal of genome-wise association studies (GWAS). The genetic disease eciteffects are potentially confounded by such covariates as age, gender, or education levels, which not only could reduce the statistical power, but also cause spurious genetic associations (Glorioso and Sibille (2011); Wang et al. (2012)). As a motivating example, consider a study of the association between copy number changes and gene expression levels in breast cancer. In this study, there were a total of 88 subjects with 19672 genes and 2149 measurements of copy number changes after preprocessing. Age at diagnosis as well as other covariates have been found to be confounders of the cancer effect with significant interaction terms in some biomarkers (Stephens et al. (2012)). Genetic variants identification without proper adjustment for confounders could yield spurious associated genetic detections. In addition, the response variable is multivariate and thus traditional feature selection techniques cannot be directly utilized.

In the last two decades, variable selection plays an prominent role in

*To whom correspondence should be addressed.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

high-dimensional statistical modeling, especially for genetic variants detection, see Fan and Li (2006) and Fan and Lv (2010) for a comprehensive overview. Yet this is challenged with ultrahigh dimensional data by computational cost and estimation accuracy.

Fan and Lv (2008) introduced the concept of sure screening and proposed the Sure Independent Screening (SIS) method to select important variables in ultra-high dimensional linear models. They showed that this correlation-ranking procedure enjoys a sure screening property in linear models with Gaussian error: with probability close to 1, the SIS procedure retains all of the important variables. Later, the SIS method has been extended by relaxing the model assumptions or the error distribution assumption, see Fan and Song (2010), Hall and Miller (2009), Fan et al. (2011), and Zhu et al. (2011). In particular, Li et al. (2012) proposed a model-free screening procedure called DC-SIS by ranking the marginal utility measure based on distance correlation, which is an efficient measure of dependence. The distance correlation of two random vectors is zero if and only if they are independent (Székely et al. (2007); Székely and Rizzo (2009)), a property that is not shared by other correlations. Furthermore, due to the nature of distance correlation, DC-SIS can be directly applied in cases with multivariate responses. However, these approaches ignored effects from confounding

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

covariates, which calls for research on SIS procedures to take them into account.

In this paper, we propose a novel model-free feature screening procedure by ranking the conditional distance correlation of the response and each predictor on confounding covariates. The conditional distance correlation was proposed by Wang et al. (2015) and possesses an appealing property analogous to the distance correlation, the conditional distance correlation of two random vectors given a random vector is zero if and only if they are conditionally independence. Compared with DC-SIS, our proposal incorporates confounding covariates into the feature screening process and hence can increase the statistical power. Furthermore, it does not require one to specify the distribution or the regression model, making the procedure particularly flexible in feature screening. Our method is also applicable to multivariate response by the virtues of conditional distance correlation.

Theoretically, we establish the sure screening property for the proposed procedure under the ultrahigh dimensional model setting. The sure screening property guarantees that our screening procedure includes the true model with probability tending to 1 at an exponential rate of the sample size n. This property is valid provided that the dimensionality of the predictors p grows slightly slower than $\exp(an)$ for any fixed $a > 0$. This rate

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

is comparable to those achieved by DC-SIS and SIS. In simulation studies, we demonstrate that our method possesses the sure screening property, and has superior performance than DC-SIS and SIS under a variety of settings.

The rest of this paper is organized as follows. In Section 2, we develop our feature screening approach corrected for confounding covariates. The sure screening property of this procedure is established in Section 3. Section 4 illustrates its finite performance by Monte Carlo simulations and an analysis of breast cancer data. A brief discussion is provided in Section 5. Proofs can be found in the online Supplementary Materials.

2. Independence Screening using Conditional Distance Correlation

Let Y be a q_y-dimensional response variable, which can be either univariate or multivariate. Let $(X_1, \ldots, X_p)$ be predictor vectors, and Z be the q_z-dimensional confounding covariates of such as age and education. The predictor $X_r, r = 1, \ldots, p$, is p_r-dimensional to allow categorical or grouped variables. We allow p to grow with n and denote it by p_n whenever necessary. There is the none of active predictors and inactive predictors conditioning on Z. Without specifying a regression model, we define the

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

index sets of active and inactive predictors given Z by

$$\mathcal{A} = \{r : \text{ Some } Y \text{ depends on } X_r \text{ given } Z\},$$

$$\mathcal{I} = \{r : \text{ Any } Y \text{ does not depend on } X_r \text{ given } Z\}.$$

The intersection of $\mathcal{A}$ and $\mathcal{I}$ is empty. Our primary interest is in identify all active predictors given the confounding covariates. When there are no confounding effects, the sure independence screening procedure based on distance correlation (DC-SIS, Li et al. (2012)) is desirable for its model-free property and flexible application to grouped predictor variables and multivariate response variables. The distance correlation works by measuring a weighted $L_q (0 < q \leq 2)$ distance between the joint characteristic function and the product of the two marginal characteristic functions. A suitable weight is selected to make the distance correlation a proper and scale invariant correlation measurement. This type of weight function leads to a simple product-average form of the covariance analogous to Pearson covariance (Székely et al. (2007); Székely and Rizzo (2009)). This motivates us to extend the DC-SIS feature screening procedure to take the confounding effects into account.

Tests of conditional dependence are a widely used statistical method for controlling confounding effects. A series of conditional dependence measures have been developed using a generalization of empirical distribution

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

function (Linton and Gozalo (1996)), smoothing empirical likelihood (Su and White (2003)), a normalized conditional cross-covariance operator in reproducing kernel Hilbert space (Gretton et al. (2005)), conditional characteristic function (Su and White (2007)), weighted Hellinger distance (Su and White (2008)) and the Hibert-Schmidt norm with Copula transformation (J Reddi and Póczos (2013)). Inspired by the success of distance correlation (Székely et al. (2007)), several conditional measurements have been proposed based on the distance correlation. Póczos and Schneider (2012) replaced the characteristic function with the conditional characteristic function and derived the estimation based on the k-nearest neighbour method. Fan et al. (2015) proposed a conditional independence measure based on the distance covariance between the residuals after adjusting for the covariates. Wang et al. (2015) proposed a novel conditional dependence measure called conditional distance correlation (CDC), which was shown to satisfy the property, the conditional distance correlation of two random vectors given a random vector is zero if and only if they are conditionally independent. This guarantees that the conditional distance correlation can describe exactly the relationship between two variables given a third variable. They derived a corresponding statistic for a test of conditional dependence.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Given the confounding covariate Z, the conditional distance correlation between the response variable Y and each predictor variable $X_r, r = 1, \ldots, p$, is defined by

$$\rho_r(Z) = \mathrm{CDCor}^2(X_r, Y \mid Z) = \frac{\mathrm{CDCov}^2(X_r, Y \mid Z)}{\sqrt{\mathrm{CDCov}^2(X_r, X_r \mid Z)\mathrm{CDCov}^2(Y, Y \mid Z)}},$$

if $\mathrm{CDCov}^2(X_r, X_r \mid Z)\mathrm{CDCov}^2(Y, Y \mid Z) > 0$, and 0 otherwise. The conditional distance covariance between X_r and Y given Z is defined as

$$\begin{aligned} & \mathrm{CDCov}^2(X_r, Y \mid Z) \\ = \; & \|\phi_{X_r,Y|Z}(t,s) - \phi_{X_r|Z}(t)\phi_{Y|Z}(s)\|^2 \\ = \; & \frac{1}{c_{p_r}c_{q_y}} \int_{\mathbb{R}^{p_r+q_y}} \frac{|\phi_{X_r,Y|\mathbf{z}}(t,s) - \phi_{X_r|Z}(t)\phi_{Y|\mathbf{z}}(s)|^2}{|t|_{p_r}^{1+p_r}|s|_{q_y}^{1+q_y}} dtds, \end{aligned}$$

where $c_{p_r} = \pi^{(1+p_r)/2}/\Gamma((1+p_r)/2)$ and $c_{q_y} = \pi^{(1+q_y)/2}/\Gamma((1+q_y)/2)$ are constants related to p_r and q_y. Here $\phi_{X_r,Y|Z}(t,s)$ is the joint conditional characteristic function of X_r and Y given Z, and $\phi_{X_r|Z}(t)$ and $\phi_{Y|Z}(s)$ are the marginal conditional characteristic functions for $X_r \mid Z$ and $Y \mid Z$, respectively. The terms $\mathrm{CDCov}^2(X_r, X_r \mid Z)$ and $\mathrm{CDCov}^2(Y, Y \mid Z)$ are defined similarly.

From the definition, $\rho_r(Z)$ is a function of Z and therefore not yet ready for ranking, so we define a marginal utility to screen features as

$$\rho_r^* = E(\rho_r(Z)) = E\{\mathrm{CDCor}^2(X_r, Y \mid Z)\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

To estimate ρ_r^*, let us first proceed with the estimation of $\mathrm{CDCov}^2(X_r, Y \mid Z)$. It essentially requires the estimation of $\phi_{X_r,Y|Z}(t,s)$, $\phi_{X_r|Z}(t)$ and $\phi_{Y|Z}(s)$, which can be estimated from their empirical versions, respectively. In particular, we use the Gaussian kernel smoothing method to them.

Suppose $\mathbf{W} = (\mathbf{X}_1, \ldots, \mathbf{X}_p, \mathbf{Y}, \mathbf{Z}) = \{(X_{k1}, \ldots, X_{kp}, Y_k, Z_k) : k = 1, \ldots, n\}$ is a random sample from the joint distribution of random vectors X, Y, and Z. For the rth predictor, let $d_{ij,r}^X = d(X_{ir}, X_{jr})$ be the Euclidean distance of X_{ir} and X_{jr}, $i, j = 1, \ldots, n$. Similarly, let $d_{ij}^Y = d(Y_i, Y_j)$ denote the Euclidean distance of Y_i and Y_j, $i, j = 1, \ldots, n$. Define the distance function as

$$d_{ijkl,r}^s = (d_{ijkl,r} + d_{ijlk,r} + d_{ilkj,r})/3,$$

where $d_{ijkl,r} = (d_{ij,r}^X + d_{kl,r}^X - d_{ik,r}^X - d_{jl,r}^X)(d_{ij}^Y + d_{kl}^Y - d_{ik}^Y - d_{jl}^Y)$.

Given Z, the sample conditional distance covariance $\mathrm{CDCov}^2(X_r, Y \mid Z)$ is

$$\widehat{\mathrm{CDCov}}^2(\mathbf{X}_r, \mathbf{Y}, \mathbf{Z} \mid Z) = n^{-4} \sum_{i,j,k,l} \psi_n(\mathbf{W}_i, \mathbf{W}_j, \mathbf{W}_k, \mathbf{W}_l; Z),$$

with the symmetric random kernel of degree 4,

$$\psi_n(\mathbf{W}_i, \mathbf{W}_j, \mathbf{W}_k, \mathbf{W}_l; Z) = \frac{n^4 \omega_i(Z)\omega_j(Z)\omega_k(Z)\omega_l(Z)}{4\omega^4(Z)} d_{ijkl,r}^s,$$

where $\omega_i(Z)$ is an estimate for the density function in Z_i and $\omega(Z)$ is $n^{-1}\sum \omega_i(Z)$. Wang et al. (2015) showed that $\widehat{\mathrm{CDCov}}^2(\mathbf{X}_r, \mathbf{Y}, \mathbf{Z} \mid Z)$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

a V process that has a well-established asymptotic framework (Lee (1990)). The sample conditional distance variances $\widehat{\mathrm{CDCov}}^2(\mathbf{X}_r, \mathbf{X}_r, \mathbf{Z} \mid Z)$ and $\widehat{\mathrm{CDCov}}^2(\mathbf{Y}, \mathbf{Y}, \mathbf{Z} \mid Z)$ can be defined similarly. Thus the sample conditional distance correlation $\hat{\rho}_r(Z)$ is defined as

$$\begin{aligned}\hat{\rho}_r(Z) &= \widehat{\mathrm{CDCor}}^2(\mathbf{X}_r, \mathbf{Y}, \mathbf{Z} \mid Z) \\ &= \frac{\widehat{\mathrm{CDCov}}^2(\mathbf{X}_r, \mathbf{Y}, \mathbf{Z} \mid Z)}{\sqrt{\widehat{\mathrm{CDCov}}^2(\mathbf{X}_r, \mathbf{X}_r, \mathbf{Z} \mid Z)\widehat{\mathrm{CDCov}}^2(\mathbf{Y}, \mathbf{Y}, \mathbf{Z} \mid Z)}}.\end{aligned}$$

A plug-in estimate of ρ_r^* is

$$\hat{\rho}_r^* = n^{-1}\sum_{k=1}^{n}\hat{\rho}_r(Z_k).$$

Using $\hat{\rho}_r^*$ as a marginal utility, we propose a screening procedure for ultrahigh dimensional data with the control of confounding as:

$$\hat{\mathcal{M}}_{d_n} = \{r : \hat{\rho}_r^* \text{ is among the first } d_n \text{ largest of all, } r = 1, \ldots, p\},$$

where the submodel size d_n is predefined to be smaller than the sample n. This reduces the full model of size $p \gg n$ to a submodel with size d_n. This procedure is referred as conditional distance correlation sure independence screening (CDC-SIS for short).

3. Theoretical Properties

In this section, we establish the asymptotic properties of the CDC-SIS

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

procedure. To derive the sure screening property for CDC-SIS, we impose some regularity conditions on X, Y, and Z as follows.

(C1): The kernel function $K(\cdot)$ is bounded uniformly such that $K(u) \geq 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\int \|u\|^2 K(u)du < \infty$.

(C2): There exists a positive constant s_0 such that for all $0 < s \leq s_0$,

$$\sup_p \max_{1\leq r\leq p} E(\exp(s\|X_r\|_p^2)) < \infty, \quad E(\exp(s\|Y\|_{q_y}^2)) < \infty,$$

where p and q_y are the dimensions of the predictor X_r and the response variable Y, respectively.

(C3): If Z_1, Z_2, Z_3, Z_4 are independent copies of Z, then for $1 \leq r \leq p$, there exists a positive constant L, such that

$$\sup_r |E(d_{1234,r}|Z_1, Z_2, Z_3, Z_4) - E(d_{1234,r}|Z_1', Z_2, Z_3, Z_4)| \leq L|Z_1 - Z_1'|.$$

(C4): There exist some constants $c > 0$ and $0 \leq \kappa < 1/2$ such that

$$\min_{r\in\mathcal{A}} \rho_r^* \geq 2cn^{-\kappa}.$$

Condition (C1) is a mild condition on the density function $f(z)$ and kernel function $K(\cdot)$. Condition (C2) puts an exponential bound on the tails of X

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

and Y; similar condition is used in Fan and Lv (2008) and Li et al. (2012). Condition (C3) is satisfied if the first order partial derivative of $E(d_{1234,r} \mid Z_1, Z_2, Z_3, Z_4)$ is bounded. Condition (C4) requires the true conditional distance correlation between the active predictors and the response is large enough.

The proof of the following is in the Supplementary Materials.

Theorem 1. *If Conditions (C1)-(C3) hold and the bandwidth for kernel estimation of Z satisfies $h = O(n^{-\kappa/(2q_z)})$, then for any $0 < \gamma < 1/2 - \kappa$, there exists positive constants $c_1 > 0$ and $c_2 > 0$ such that*

$$\mathrm{pr}(\max_{1\le r\le p} |\hat{\rho}_r^* - \rho_r^*| \ge cn^{-\kappa}) \le p[\exp(-c_1 n^{1-2(\gamma+\kappa)}) + n^4 \exp(-c_2 n^{\gamma})] + o(1).$$

If Condition (C4) also holds, we have

$$\begin{aligned}\mathrm{pr}(\mathcal{A} \subseteq \hat{\mathcal{M}}_{d_n}) \ge & 1 - n|\mathcal{A}|[\exp(-c_1 n^{1-2(\kappa+\gamma)}) + n^4 \exp(-c_2 n^{\gamma})] \\ & - |\mathcal{A}| \exp(-c_3 n^{1-2\kappa}) + o(1).\end{aligned}$$

where $|\mathcal{A}|$ is the size of the set $\mathcal{A}$ and c_3 is a positive constant.

Theorem 1 requires that the bandwidth of kernel estimation of Z satisfies $h = O(n^{-\kappa/(2q_z)})$, where q_z is the dimension of Z, and $0 < \kappa < 1/2$. Similar conditions can be found in Liu et al. (2014). This rate is enough to ensure the density estimate to be consistent. Here this rate could be

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

faster or slower than the theoretical optimal rate of kernel density estimation, depending on the choice of κ and γ. One can see that when γ is fixed, the right side of $\text{pr}(\mathcal{A} \subseteq \hat{\mathcal{M}}_{d_n})$ increases as κ decreases, indicating that oversmoothing could benefit the screening performance in terms of the probability of true active predictors.

The convergence rate of the sure property depends only on $|\mathcal{A}|$ rather than the dimensionality p. This has the size of $\mathcal{A}$ is smaller than the sample size n, and much smaller than p. Thus, CDC-SIS is an effective and general alternative as a sure screening procedure adjusted for confounding covariates.

4. Numerical Studies

4.1 General Setting

We conducted three simulation studies and a genetic data analysis to evaluate the finite sample performance of CDC-SIS, and compared its performance with those of SIS (Fan and Lv (2008)) and DC-SIS (Li et al. (2012))cite. The bandwidth parameter of $\hat{\rho}_r^*$ in the CDC-SIS method was determined by optimizing the conditional distance correlation. If the bandwidth was too close to zero, we selected the bandwidth with the plug-in method.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

To assess the feature screening performance, we considered two criteria adopted from Li et al. (2012): the minimum model size $\mathcal{S}$ to include all active predictors for a specific method; the proportion $\mathcal{P}_d$ that all active predictors are selected by a screening procedure for a given model size d in 100 replications. We report the median of $\mathcal{S}$ and draw a boxplot of $\log \mathcal{S}$ out of 100 replications.

Here $\mathcal{S}$ is used to measure the model complexity of the resulting model for an underlying screening procedure. The closer to the size of active predictors that $\mathcal{S}$ is, the better the screening procedure performs. The sure screening property ensures that $\mathcal{P}_d$ is close to one when the estimated model size d is sufficiently large. Li et al. (2012) suggested setting $d = \gamma\lceil n/\log(n)\rceil$, where $\lceil a\rceil$ refers to the integer part of a and γ is an integer. To examine the overall performance of the choice of d, we consider a plot of $\mathcal{P}_d$ with $d = \gamma\lceil n/\log(n)\rceil$ as the y coordinate versus γ as the x coordinate.

For all simulations, we generated $U = \{Z, X\} = (Z, X_1, \ldots, X_p)^T$ from the normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{(p+1)\times(p+1)}$, characterized by ρ. The ρ was set at 0, 0.5 and 0.8 to examine the impact of correlation to the screening performance. The sample size n was fixed to be 100 and the dimensionality p varied from 1000 to 5000. All simulations were replicated for 100 times.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

4.2 Simulation Studies

EXAMPLE 1 In this example, we considered four models:

$$\text{(1.a):} \quad Y = 2.5Z + 3X_1 + 1.5X_2 + 2X_5 + \epsilon;$$

$$\text{(1.b):} \quad Y = 2.5Z + 3X_1 + 1.5X_2 + 2X_5^2 + \epsilon;$$

$$\text{(1.c):} \quad Y = 2.5Z + 3X_1 + 1.5X_2 + 2\sin(0.5\pi X_5) + \epsilon;$$

$$\text{(1.d):} \quad Y = 3X_1 + 1.5X_2 + 4ZX_5 + \epsilon,$$

with the ϵ's i.i.d. standard normal. For X_5, the regression function is linear in model (1.a), but nonlinear in all others. The regression function of X_5 is non-monotone in model (1.b), and periodic in model (1.c). In model (1.d), there is an interaction term involving the confounding covariate Z, ZX_5. Two covariance matrices Σ were considered: compound symmetric (CS); first-order autoregressive (AR). The CS covariance matrix Σ has entries $\sigma_{ii} = 1, i = 1, \ldots, p+1$, and $\sigma_{ij} = \rho, i \neq j$. The AR matrix covariance Σ has entries $\sigma_{ij} = \rho^{|i-j|}, i, j = 1, \ldots, p+1$.

To save space, we report only the summary results with compound symmetric covariance matrix of X and $p = 1000$ in Table 1 and Figures 1-2. The other summary results are in Table S1 and Figures S1-S6 of the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Table 1: Example 1: Median of the minimum model size S for the SIS, DC-SIS, and CDC-SIS methods for different values of p and ρ based on 100 replications under different models with a compound symmetric (CS) covariance matrix.

.

Model	p	ρ	SIS	DC-SIS	CDC-SIS
(1.a)	1000	0	6	8	5
		0.5	18	48	6
		0.8	40	67	16
	5000	0	16	28	9
		0.5	114	208	18
		0.8	189	292	40
(1.b)	1000	0	273	16	13
		0.5	598	404	82
		0.8	454	384	70
	5000	0	1261	58	30
		0.5	2630	1655	344
		0.8	2604	2040	370
(1.c)	1000	0	83	40	25
		0.5	94	43	12
		0.8	132	62	18
	5000	0	234	60	82
		0.5	558	324	72
		0.8	804	472	76
(1.d)	1000	0	312	128	5
		0.5	498	218	17
		0.8	661	283	33
	5000	0	1818	694	27
		0.5	2915	1331	66
		0.8	2654	1225	139

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

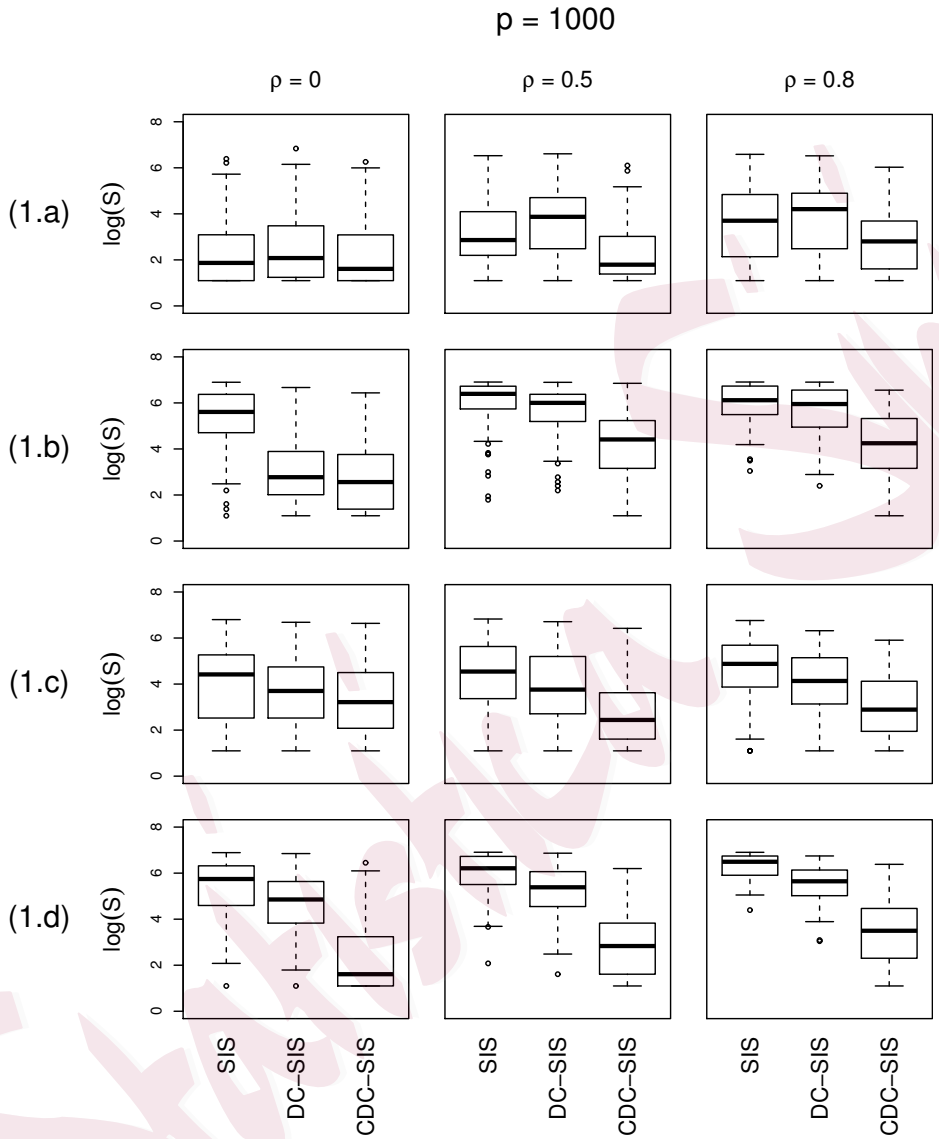

Figure 1: Example 1: Boxplots of $\log(\mathcal{S})$ for the SIS, DC-SIS, and CDC-SIS methods for different values of ρ based on 100 replications under different models with $p = 1000$ and a compound symmetric (CS) covariance matrix.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

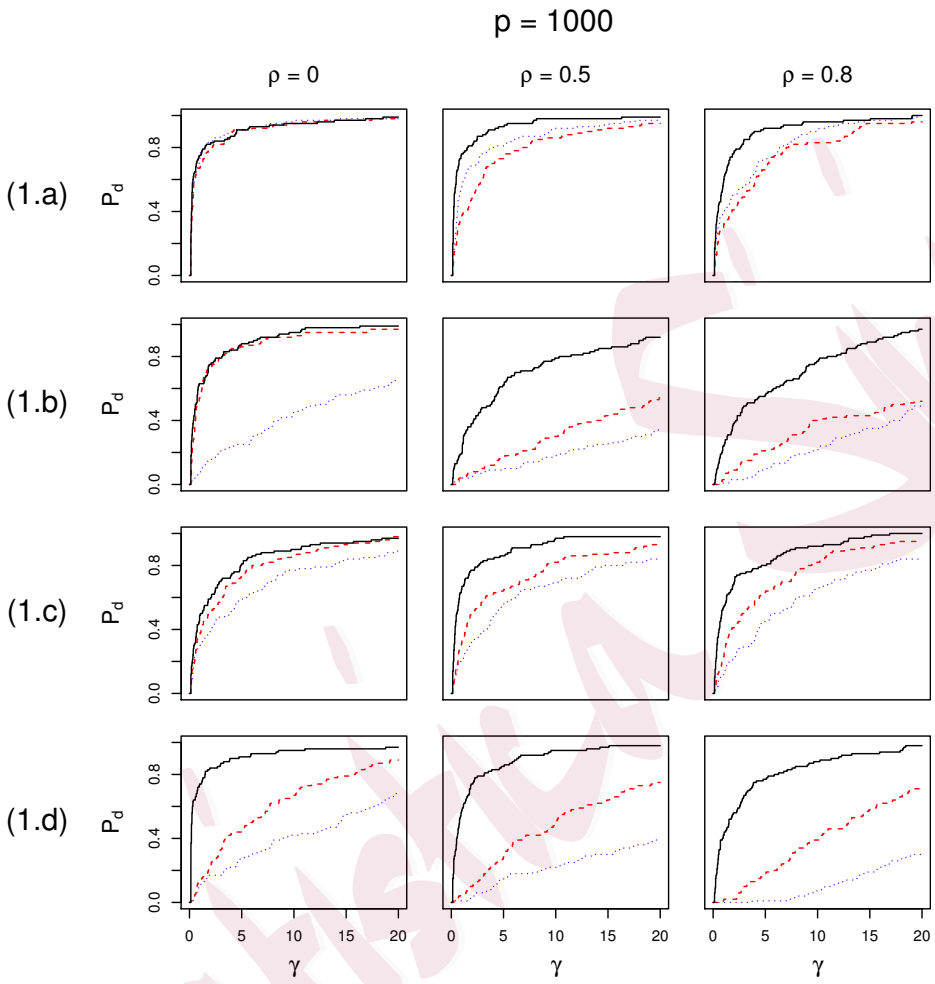

Figure 2: Examples 1: Summary results of the proportion of $\mathcal{P}_d^\gamma$ for the SIS (dotted line), DC-SIS (dashed line), and CDC-SIS (solid line) methods for different values of ρ based on 100 replications under different models with $p = 1000$ and a compound symmetric (CS) covariance matrix.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Obviously, DC-SIS and CDC-SIS outperform SIS in all models except (1.a) where a linear regression model is assumed; This was also found by Li et al. (2012). The performance of CDC-SIS is slightly better than that of SIS and much better than those of DC-SIS in model (1.a), indicating that CDC-SIS has a robust performance if the working models is linear.

Compared with DC-SIS, CDC-SIS shrinks the full model to a much smaller scale by taking account of the confounding covariate Z in models (1.b)-(1.d). In particular, in model (1.d) with the CS covariance matrix, while the other two screening procedures fail to identify the active predictor even with the threshold d being n (= 100), CDC-SIS ranks all the active predictors at the top positions for $p = 1000$ and $\rho = 0.5$. In addition, the proportions $\mathcal{P}_d$ of our proposed method are close to one, which supports the assertion that it possesses the sure screening property. Compared to other screening methods, CDC-SIS has significant better performance with the curves of $\mathcal{P}_d^{\gamma}$ being the one at the upper left, especially in models (1.b)-(1.d). The effect is more pronounced for higher values of ρ and higher dimensionality. It suggests that adjusting for the confounding covariate(s) helps reduce false selection and might subsequently improve the prediction accuracy subsequently.

Furthermore, CDC-SIS is less sensitive to the change in the correlation

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

ρ than SIS. While the median $\mathcal{S}$ of SIS at least doubles with ρ increasing from 0.5 to 0.8, those of CDC-SIS remain almost the same. This phenomenon is exemplified in Fan and Lv (2008) for the SIS with the Pearson correlation. When the confounding covariates are highly correlated with the predictors, the dimensionality and thus the correlation increases. Since the conditional distance correlation can remove the confounding effect, it can be less influenced by the correlation between the confounding covariates and the predictors variables.

We include comparison results for ISIS, an iterative version of SIS for variable selection. The threshold was set at $d = \lceil n/\log(n) \rceil$ for all methods. For ISIS, the SCAD variable selection method with regularization parameter tuning by Bayesian information criterion was used after the SIS screening step and we kept collecting variables until we obtained $\lceil n/\log(n) \rceil$ of them. Table 2 and Table S2 in the online Supplementary Material report the percentages of SIS, ISIS, DC-SIS, and the prosed CDC-SIS that include the true model $\{X_1, X_2, X_5\}$, an index also used in Fan and Lv (2008). From Table 2 and Table S2, ISIS always improves the performance of SIS, especially for larger value of ρ, which confirms the findings in Fan and Lv (2008). Yet compared with SIS and ISIS, CDC-SIS always has the best performance except for Model (1.a), where a linear model holds.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Table 2: Example 1: Accuracy of SIS, ISIS, DC-SIS, and CDC-SIS in including the true model $\{X_1, X_2, X_5\}$ for different values of ρ and p with a compound symmetric (CS) covariance matrix.

.

Model	p	ρ	SIS	ISIS	DC-SIS	CDC-SIS
(1.a)	1000	0	0.73	0.97	0.67	0.75
		0.5	0.55	0.91	0.37	0.76
		0.8	0.39	0.88	0.30	0.58
	5000	0	0.53	0.78	0.45	0.61
		0.5	0.23	0.67	0.17	0.52
		0.8	0.16	0.54	0.08	0.43
(1.b)	1000	0	0.07	0.12	0.54	0.63
		0.5	0.04	0.04	0.04	0.19
		0.8	0.01	0	0.03	0.24
	5000	0	0.03	0.03	0.33	0.45
		0.5	0.01	0.01	0.01	0.14
		0.8	0	0	0	0.08
(1.c)	1000	0	0.3	0.49	0.37	0.49
		0.5	0.22	0.44	0.31	0.65
		0.8	0.11	0.19	0.23	0.54
	5000	0	0.15	0.21	0.27	0.31
		0.5	0.01	0.09	0.08	0.31
		0.8	0.03	0.05	0.08	0.16
(1.d)	1000	0	0.11	0.09	0.13	0.74
		0.5	0.01	0.03	0.05	0.57
		0.8	0	0	0.01	0.42
	5000	0	0.01	0	0.01	0.46
		0.5	0.01	0	0.03	0.34
		0.8	0	0	0	0.20

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

EXAMPLE 2 As suggested by one of the reviewers, we examined the effect of confounding covariate on the performance of CDC-SIS. Three models were considered: (2.a) the univariate covariate Z is not directly related to the response; (2.b) the confounding covariate $Z = (Z_1, Z_2)$ is two-dimensional and only Z_1 is related to the response; (2.c) the confounding covariate $Z = (Z_1, Z_2)$ is two-dimensional and both of Z_1 and Z_2 are related to the response. The response was generated, respectively, by

$$\text{(2.a):} \quad Y = 3X_1 + 1.5X_2 + 2X_5^2 + \epsilon,$$

$$\text{(2.b):} \quad Y = 2.5Z_1 + 3X_1 + 1.5X_2 + 2X_5^2 + \epsilon,$$

$$\text{(2.c):} \quad Y = 2.5Z_1 + 2.5Z_2 + 3X_1 + 1.5X_2 + 2X_5^2 + \epsilon,$$

where the ϵ's were i.i.d. standard normal. For Model (2.a), the covariance matrix of $U = \{Z, X\}$ is a block diagonal matrix with the first block having $\sigma_{ii} = 1, i = 1, \ldots, 3$, and $\sigma_{ij} = \rho, i \neq j$ and the second block with $\sigma_{ii} = 1, i = 4, \ldots, p+1$, and $\sigma_{ij} = \rho, i \neq j$. The confounding covariate Z is strongly correlated with X_1 and X_2. However, X_5 belongs to a group of strongly correlated variables $\{X_3, \cdots, X_p\}$ that are independent of Z, X_1 and X_2. For Models (2.b) and (2.c), the covariance matrix of $U = \{Z, X\}$ has entries $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho, i \neq j$. The results for $p = 1000$ are summarized in Table 3 and Figures 3-4. The results for $p = 5000$ are

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

summarized in Table S3 and Figures S7-S8 of the online Supplementary Material.

It is no surprise that the DC-SIS and CDC-SIS methods significantly outperform SIS since the model is non-linear in X_5, which again support the findings in Li et al. (2012). The results in Table 3 present interesting patterns. For $\rho = 0$, the performance of the CDC-SIS is competitive with the DSIS in terms of $\mathcal{S}$. When $\rho = 0.8$, the medians $\mathcal{S}$ of the CDC-SIS are half of those of the DC-SIS. As implied by results in Figure 4, the CDC-SIS has much better performance in terms of $\mathcal{P}_d^\gamma$. Overall, taking account for the confounding covariates in the feature screening process may help reduce false selection even when the confounding variable is not directly related to the response, or only partially correlated with the response.

Next, we consider data with multivariate response. Since the SIS cannot deal with this kind of data directly, we focus on the results from the DC-SIS and CDC-SIS methods.

EXAMPLE 3 We explore the performance of the proposal in dealing with data with multivariate response. We first considered a two-dimensional response, then a high-dimensional response. To make the simulation mimic the motivating data, the dimension in the second scenario was the same as

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

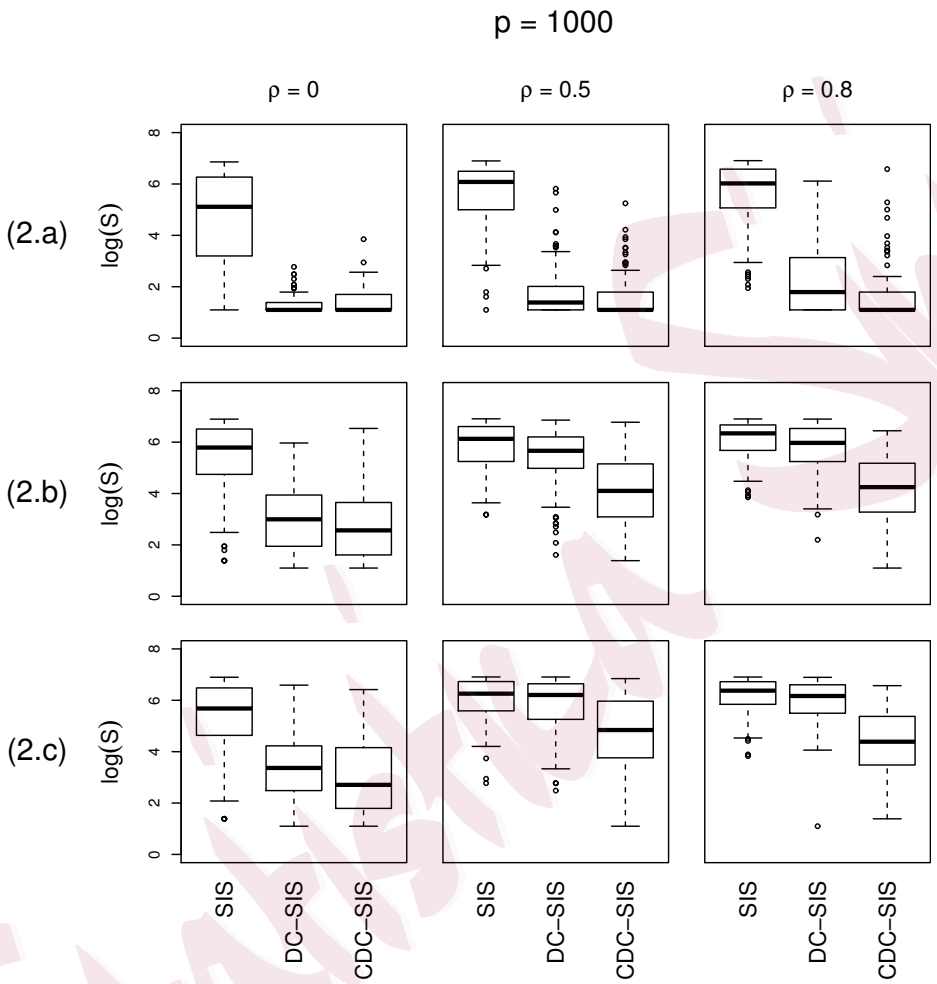

Figure 3: Example 2: Boxplots of $\log(\mathcal{S})$ for the SIS, DC-SIS, and CDC-SIS methods for different values of ρ based on 100 replications under different models with $p = 1000$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

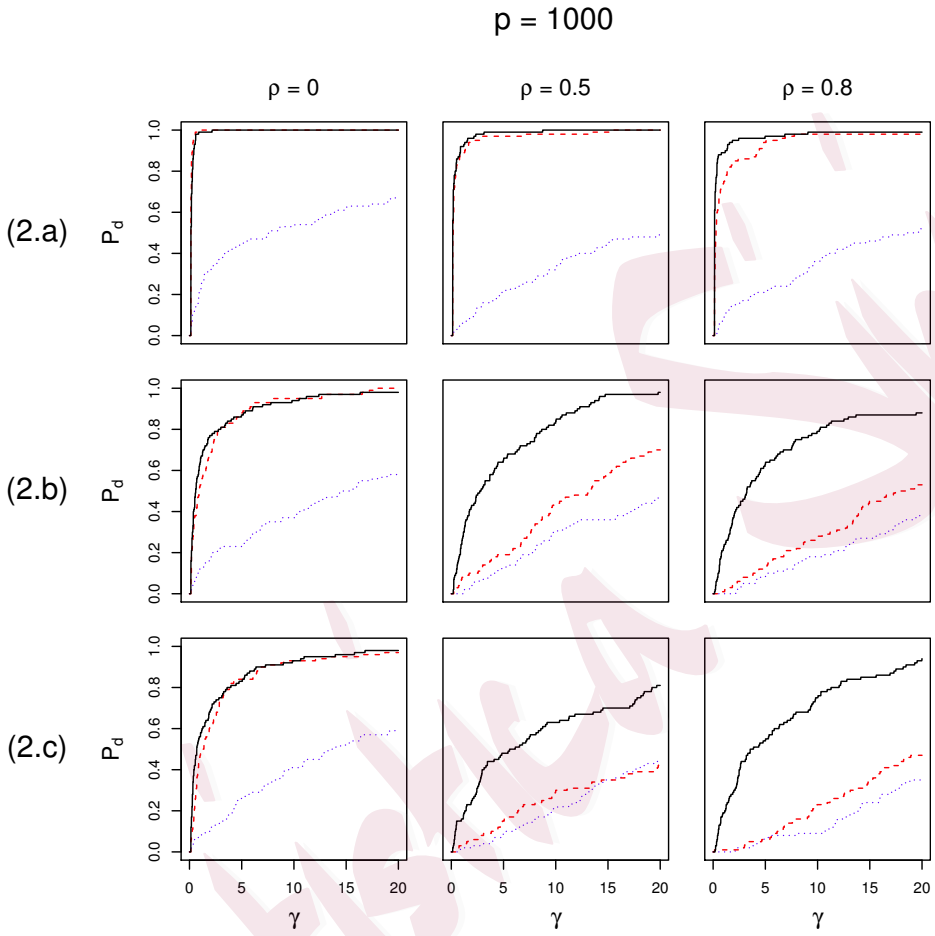

Figure 4: Examples 2: Summary results of the proportion of $\mathcal{P}_d^\gamma$ for the SIS (dotted line), DC-SIS (dashed line), and CDC-SIS (solid line) methods for different values of ρ based on 100 replications under different models with $p = 1000$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

that in the motivating data. Here, the covariance matrix Σ of $U = \{Z, X\}$ had entries $\sigma_{ii} = 1, i = 1, \ldots, p+1$, and $\sigma_{ij} = \rho, i \neq j$.

(3.a): The two-dimensional response $Y = (Y_1, Y_2)^\top$ was generated by

$$Y = Z\beta_Z + X\beta_X + E,$$

where E was generated as normal with mean zero and covariance matrix $\Sigma_{y|x} = I_{2\times 2}$. We chose a pair of (β_Z, β_X) such that Y_1 and Y_2 shared the same association with X_1, $\beta_Z = (-1.6\sqrt{\rho},\, 1.6,\, 0, \ldots, 0)^\top$ and $\beta_X = (0,\, 1.6,\, -1.6\sqrt{\rho}, 0, \ldots, 0)^\top$.

(3.b): The 136-dimensional response Y was generated by

$$Y = Z\beta_Z + X_1\beta_1 + X_2\beta_2 + E,$$

where $E^{(2)}$ was generated as normal with mean zero and covariance matrix $\Sigma_{y|x} = I_{136\times 136}$. Here the response Y is related to the first three predictors $\{X_1, X_2\}$ and confounding covariate Z. The nonzero regression coefficients, the first four column of $B^{(2)}$, were drawn as standard normal, independently.

Results are summarized in Table 4 and Figures 5-6. There it can be seen that the benefit of adjusting for the confounding effect is significant. Compared to DC-SIS, CDC-SIS needs a much smaller model to include

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

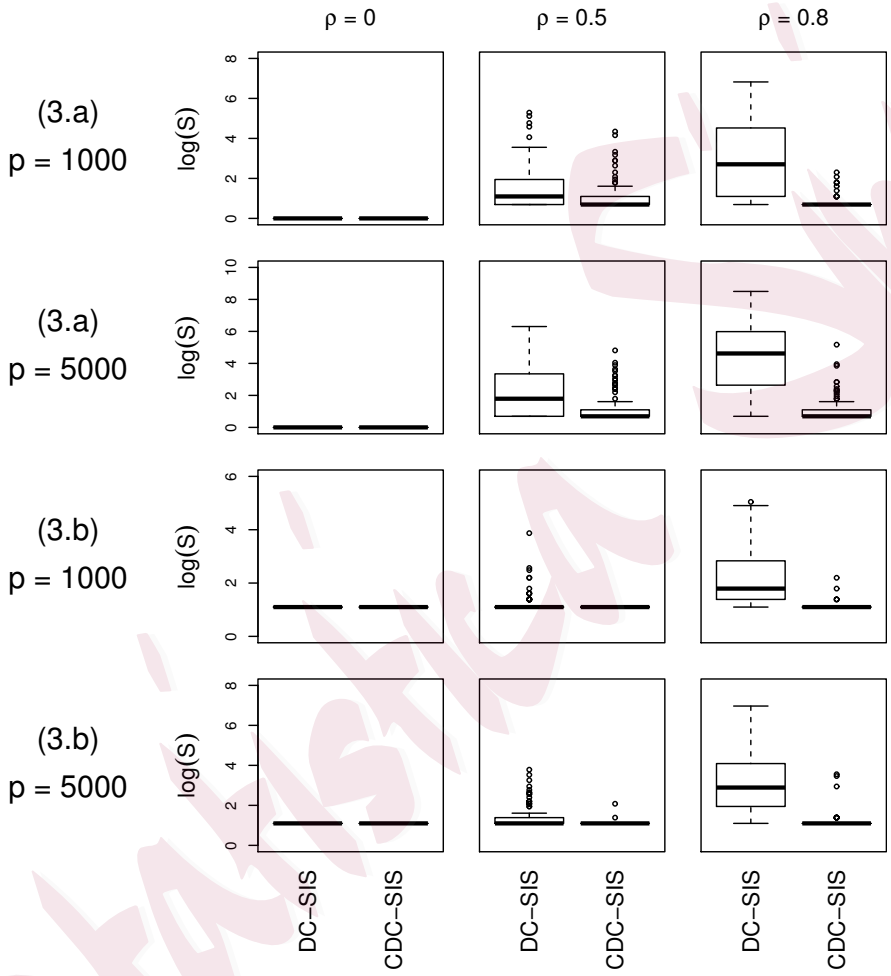

Figure 5: Example 3: Boxplots of $\log(\mathcal{S})$ for the DC-SIS and CDC-SIS for different values of p and ρ based on 100 replications under different models.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

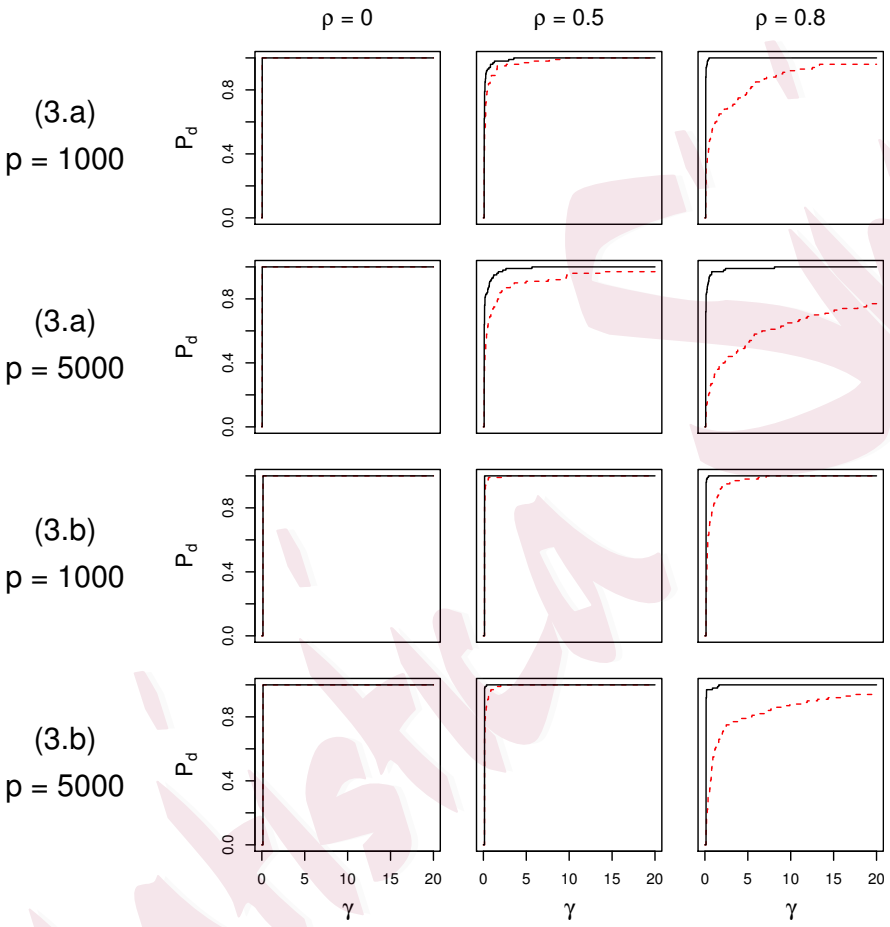

Figure 6: Examples 3: Summary results of the proportion of $\mathcal{P}_d^\gamma$ for the DC-SIS (dashed line) and CDC-SIS (solid line) methods for different values of p and ρ based on 100 replications under different models.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

all the true active predictors. CDC-SIS performs very well in terms of model complexity, while DC-SIS does not especially when $\rho = 0.8$. In addition, CDC-SIS is robust to the correlation between predictors, while the proportions for DC-SIS drop dramatically as ρ increases.

4.3. Analysis of Breast Cancer Data

We illustrate our proposed method using the public breast cancer data set reported by Chin et al. (2006) and re-analyzed by Witten et al. (2009) and Ma and Sun (2014). The data set includes the gene expression, comparative genomic hybridization (CGH) measurements, and clinical characteristics for a set of breast cancer patient samples. Here CGH measures genome copy number variation along each chromosome in cancer samples; this can be helpful in characterizing certain types of cancers and understanding how the genome aberrations influence cancer pathophysiologies (Chin et al. (2006)). Our goal is to identify a set of genes that are related to the copy number changes, with or without adjustment, to potential confounding covariates. In the literature, age at diagnosis (AGE for short) and other covariates have been found to be confounders of the disease effect with significant interaction term in some biomarkers (Stephens et al. (2012)). Here we consider AGE as a potential confounding covariate.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

We extracted both the gene expression and CGH measurements data from the R package PMA (Witten et al. (2009)) and downloaded the clinical data from `http://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-158/`. After removing missing data in AGE, the data set consisted of $n = 88$ samples, $p = 19672$ gene expression measurements and 2149 CGH measurements on 23 chromosomes. Following Witten et al. (2009), we performed the screening methods for chromosome 1 using CGH measurements on chromosome 1 and all the available gene expression data. Chromosome 1 included 136 CGH measurements in total. To assess the stability of the screening results, we adopted the idea of stability selection (Meinshausen and Bühlmann (2010)). For each fixed threshold value d, we computed the selection probability of each gene over the 500 sub-samples of size $\lceil n/2 \rceil$.

Table 5 lists the top $d = \lceil n/\log(n) \rceil = 20$ genes that were identified by the DC-SIS and CDC-SIS procedures. It also includes the selection probability for these genes. As can be seen from Table 5, 16 genes are identified by both the screening methods. It is interesting to see that the first ten ranking of CDC-SIS is almost the same as that of DC-SIS, which shows the competitiveness of the proposed method.

To gain further insight, we fit generalized additive models(GAM) using the first sparse principle component (Witten et al. (2009)) of Y with or

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

without AGE. Here we take gene "HSPC003" and "B4GALT3" for example. Gene "HSPC003" is ranked the first by both two methods, while gene "B4GALT3" is only identified by DC-SIS. The two models we considered were $V = Cont_1 + f_1(gene)$ and $V = Cont_2 + f_2(gene) + g(gene, \log(age))$, where V is the first principle component of Y, $Cont_i, i = 1, 2$ are the intercept terms, and $f_i, i = 1, 2, g$ are unknown functions.

The fitted curve plots are displayed in Figure 7 and Figure 8. Leaving out AGE , both "HSPC003" and "B4GALT3" are highly correlated with V, as can be seen from the left panels of Figures 7-8. However, when including AGE as an interaction effect with the gene, they have quite different performances. As for "HSPC003", the interaction term is insignificant with p-value equal to 0.205 and the fitted curve looks like a cylindrical surface. The interaction term for "B4GALT3" is significant with p-value equal to 0.0282 and the fitted curve is non-smooth as shown in Figure 8. This means "B4GALT3" has different correlation with V at different values of AGE.

5. Discussion

Some issues deserve further study. The threshold used in the proposed method is adopted from those in Fan and Lv (2008) and Li et al. (2012). It is interest to develop a criterion to determine the threshold for finite

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

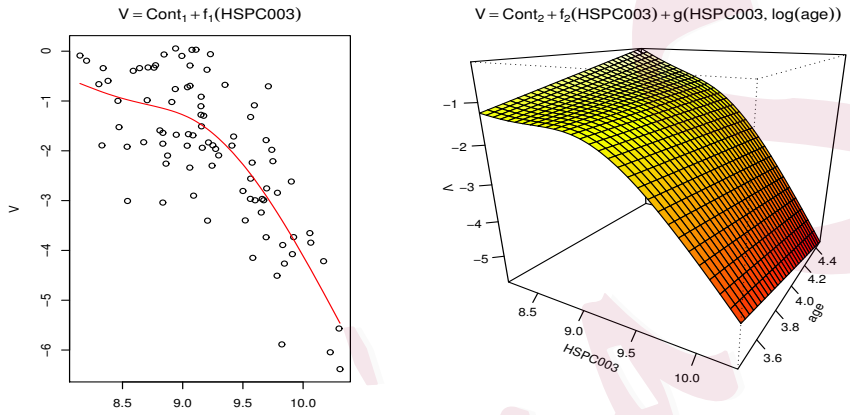

Figure 7: Data Analysis: The left panel displays the scatterplot of Y versus the expression of gene "HSPC003" well as the GAM fitting $V = Cont_1 + f_1(HSPC003)$. The right panel displays the perspective view of the GAM model $V = Cont_2 + f_2(HSPC003) + g(HSPC003, \log(age))$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

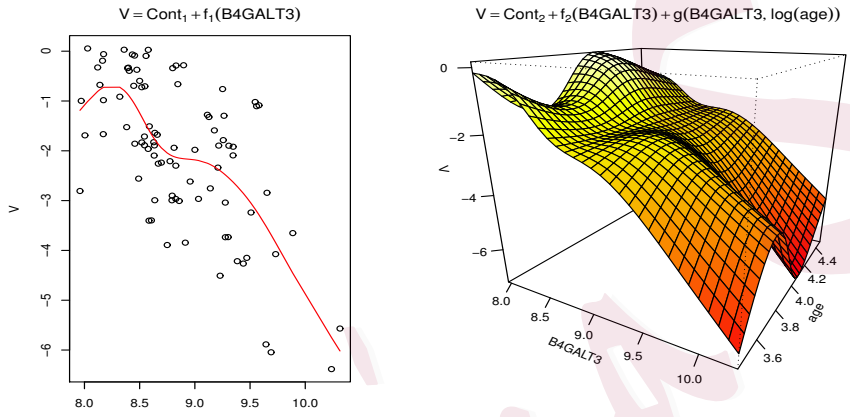

Figure 8: Data Analysis: The left panel displays the scatterplot of Y versus the expression of gene "B4GALT3" as well as the GAM fitting $V = Cont_1 + f_1(B4GALT3)$. The right panel displays the perspective view of the GAM model $V = Cont_2 + f_2(B4GALT3) + g(B4GALT3, \log(age))$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

samples and we leave it as a topic for future research. In addition, more refined model building and selection methods could be employed after feature screening, while the model-free nature of our screening method grants full flexibility in subsequent modeling.

Similar to other existing feature screening methods, the CDC-SIS procedure can fail to identify some important predictors that are marginally unrelated with the response. Thus it is an interesting problem to develop an iterative version of CDC-SIS to address such an issue. The essence of iterative procedure is to apply iteratively a large-scale variable screening followed by a moderate-scale careful variable selection. The proposed CDC-SIS procedure is model-free and thus a model-free variable selection procedure is prefered after screening. Most of the existing variable selection methods are based on a parametric regression model. In the case of multivariate response, the variable selection method should be able to handle multivariate responses. It is quite challenging to simultaneously fix both problems in theory and computation. We have an ongoing research project on this and some preliminary Monte Carlo simulations show that the iterative CDC-SIS can improve performance over the CDC-SIS procedure under univariate response setting. Similar ideas on the iterative version of DC-SIS can be found in Zhong and Zhu (2014). But without the necessary, theo-

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

retical analysis of the iterative DC-SIS and the iterative CDC-SIS deserve further study. We leave it for future investigation.

A R package implementing the CDC-SIS method, called **cdcsis**, is publicly available on CRAN.

Supplementary Materials

This includes proofs of the theoretical results and additional simulation results.

Acknowledgements

Huang's research is supported by National Natural Science Foundation of China (NNSFC), grant 11301324; and Shanghai Chenguang Program. Wang's research is partially supported by NNSFC for Excellent Young Scholar 11322108, Program for New Century Excellent Talents (NCET) 12-0559, NNSFC 11001280. The authors thank the Editor, an associate editor, and two referees for their constructive comments, which have led to a significant improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NNSFC.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

References

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

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

Fan, J., Feng, Y., and Xia, L. (2015). A conditional dependence measure with applications to undirected graphical models. *arXiv preprint arXiv:1501.01617*.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, pages 595–622.

Fan, J. and 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. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

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

Glorioso, C. and Sibille, E. (2011). Between destiny and disease: genetics and molecular path-

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

ways of human central nervous system aging. *Progress in neurobiology*, 93(2):165–181.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. *Proceedings Algorithmic Learning Theory*, pages 63–77.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3).

J Reddi, S. and Póczos, B. (2013). Scale invariant conditional dependence measures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1355–1363.

Lee, A. (1990). *U-Statistics: Theory and Practice*. Statistics: Textbooks and Monographs. M. Dekker.

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

Linton, O. and Gozalo, P. (1996). Conditional independence restrictions: Testing and estimation. *Cowles Foundation Discussion Papers*, pages 1140–1186.

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

Ma, Z. and Sun, T. (2014). Adaptive sparse reduced-rank regression. *arXiv preprint arXiv:1403.1922*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

REFERENCES

Póczos, B. and Schneider, J. (2012). Conditional distance variance and correlation. *Technique report*.

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404.

Su, L. and White, H. (2003). Testing conditional independence via empirical likelihood. *University of California at San Diego, Economics Working Paper Series*.

Su, L. and White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834.

Su, L. and White, H. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265.

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

Wang, X., Lin, Y., Song, C., Sibille, E., and Tseng, G. C. (2012). Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder. *BMC bioinformatics*, 13(1):52.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Journal of the American Statistical Association, 110(512):1726–1734.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008.

Zhong, W. and Zhu, L. (2014). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, 85(11).

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):2331–2345.

Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China

E-mail: wench6@mail.sysu.edu.cn

Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China

E-mail: panwliang@mail.sysu.edu.cn

School of Statistics and Management, and Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai, 200433, China

E-mail: huang.mian@mail.shufe.edu.cn

Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China;

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

REFERENCES

Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, 510080, China;

Xinhua College, Sun Yat-Sen University, Guangzhou, 510520, China

E-mail: wangxq88@mail.sysu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Table 3: Example 2: Median of the minimum model size S for the SIS, DC-SIS, and CDC-SIS methods for different values of p and ρ, based on 100 replications under different models.

.

Model	p	ρ	SIS	DC-SIS	CDC-SIS
(2.a)	1000	0	166	3	3
		0.5	437	4	3
		0.8	413	6	3
	5000	0	1843	5	6
		0.5	1536	5	4
		0.8	1570	14	3
(2.b)	1000	0	327	20	13
		0.5	459	288	60
		0.8	568	392	70
	5000	0	2147	130	75
		0.5	2582	1340	474
		0.8	2570	1882	330
(2.c)	1000	0	294	29	15
		0.5	522	496	127
		0.8	587	478	80
	5000	0	2154	184	84
		0.5	2292	1870	510
		0.8	3010	2774	884

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Table 4: Example 3: Median of the minimum model size S for DC-SIS and CDC-SIS methods for different values of p and ρ, based on 100 replications under different models.

.

		$p = 1000$		$p = 5000$	
Model	ρ	DC-SIS	CDC-SIS	DC-SIS	CDC-SIS
(3.a)	0*	1	1	1	1
	0.5	3	2	6	2
	0.8	15	2	102	2
(3.b)	0	3	3	3	3
	0.5	3	3	3	3
	0.8	6	3	18	3

*: When $\rho = 0$, only X_1 is correlated with Y. So the true active predictors set is $\{X_1\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0117

Table 5: Breast cancer data: The top d ($= \lceil n/\log(n) \rceil = 20$) genes identified by DC-SIS and CDC-SIS using the CGH spots on chromosome 1 and gene expression measurements on all chromosomes. Ranks are shown in the second and third columns. Their corresponding selection probability over the 500 sub-samples of size $\lceil n/2 \rceil$ with $d = 20$ are in the fourth and fifth columns.

			Selection Probability	
Gene	DC-SIS	CDC-SIS	DC-SIS	CDC-SIS
TPR	13		.382	.144
GNPAT	3	3	.894	.784
NDUFS2	12	13	.494	.354
NUP133		19	.248	.240
GGPS1	14	8	.408	.552
RAB3-GAP150	16	20	.346	.198
PEX11B	8	10	.574	.322
PIGC	7	5	.786	.826
TBCE	6	6	.674	.636
RABIF		15	.116	.244
PPOX	17	18	.318	.294
SF3B4	4	4	.790	.688
DEGS		17	.184	.336
VPS45A	20	16	.304	.310
B4GALT3	15		.408	.220
FLJ12671	5	7	.728	.522
HSPC155	10	11	.522	.386
LGTN		14	.136	.254
MRPL24	2	2	.830	.784
HSPC003	1	1	.966	.868
FLJ10876	19		.240	.164
LOC51107	18		.244	.158
C1orf27	9	9	.474	.468
MY014	11	12	.482	.324