Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Test for conditional independence with application to conditional screening

Yeqing Zhou [a], Jingyuan Liu [b], Liping Zhu [c,*]

[a] School of Mathematical Sciences, Tongji University, Shanghai 200092, China
[b] MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, 422 Siming South Road, Xiamen, 361005, China
[c] Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China

ARTICLE INFO

Article history:
Received 8 January 2019
Received in revised form 27 September 2019
Accepted 27 September 2019
Available online 9 October 2019

AMS 2010 subject classifications:
62G10

Keywords:
Conditional independence
Feature screening
High dimensional data
Independence
Sure screening property

ABSTRACT

Measuring and testing conditional dependence are fundamental problems in statistics. Imposing mild conditions on Rosenblatt transformations (Rosenblatt, 1952), we establish an equivalence between the conditional and unconditional independence, which appears to be entirely irrelevant at the first glance. Such an equivalence allows us to convert the problem of testing conditional independence into the problem of testing unconditional independence. We further adopt the Blum–Kiefer–Rosenblatt correlation (Blum et al., 1961) to develop a test for conditional independence, which is powerful to capture nonlinear dependence and is robust to heavy-tailed errors. We obtain explicit forms for the asymptotic null distribution which involves no unknown tunings, rendering fast and easy implementation of our test for conditional independence. With this conditional independence test, we further propose a conditional screening method for high dimensional data to identify truly important covariates whose effects may vary with exposure variables. We use the false discovery rate to determine the screening cutoff. This screening approach possesses both the sure screening and the ranking consistency properties. We illustrate the finite sample performances through simulation studies and an application to the gene expression microarray dataset.

1. Introduction

Independence and conditional independence are two of the most important concepts in statistics and probability theory, which form the basis of limit theorems, Markov chain, sufficiency and causality [5], among others. Conditional independence also plays a central role in graphical modeling [13], causal inference [19] and artificial intelligence [32]. Denote $X \in \mathbb{R}^1$, $Y \in \mathbb{R}^1$ and $\mathbf{u} = (U_1, \ldots, U_d)^\top \in \mathbb{R}^d$. The conditional independence between X and Y given $\mathbf{u}$ (denoted by $Y \perp\!\!\!\perp X \mid \mathbf{u}$, where $\perp\!\!\!\perp$ stands for statistical independence) reflects that knowing $\mathbf{u}$, X does not provide additional information about Y. When $Y \perp\!\!\!\perp X \mid \mathbf{u}$ holds, we can drop X and merely use $\mathbf{u}$ to predict Y, achieving the goal of feature screening.

Numerous efforts have been made to test conditional independence between X and Y given $\mathbf{u}$. [6] and [16] developed nonparametric tests of conditional independence using empirical process theory. The asymptotic null distributions of their test statistics depend on data generating processes. Consequently, bootstrap procedures are required to obtain the critical values. Later on, several omnibus tests are proposed based on empirical likelihood [27], conditional densities [11,26] or

[*] Corresponding author.
E-mail address: zhu.liping@ruc.edu.cn (L. Zhu).

https://doi.org/10.1016/j.jmva.2019.104557

conditional characteristic functions [25,29,30]. However, their critical values obtained from asymptotic normal approximation are not reliable when the sample size is pretty small. To construct an appropriate rejection region, a bootstrap procedure has to be used which, however, is computationally expensive.

In this paper, our first goal is to develop a novel test for conditional independence. With Rosenblatt transformation — a multivariate version of a probability integral transformation considered by [22], we show that conditional independence between X and Y given $\mathbf{u}$ is equivalent to the unconditional independence between the respective conditional distributions $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$ under mild conditions. Thus, the problem of testing conditional independence is converted to that of testing unconditional independence, which in turn promotes cross-fertilization: "when two different areas of study are found to be isomorphic, known results in one area immediately become available for use in the other" [5]. Such an equivalence also ensures our test to have nontrivial power against alternatives. We adopt the correlation proposed by [3] to test the unconditional independence between $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$. The resulting asymptotic null distribution does not depend upon the parental distributions of X, Y or $\mathbf{u}$, making the test distribution-free. In addition, the asymptotic limiting distribution has an explicit form without tunings. The test also possesses an oracle property in the sense that the asymptotic null distribution based on $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$ is identical to that based on $\widehat{F}_1(X \mid \mathbf{u})$ and $\widehat{F}_2(Y \mid \mathbf{u})$. Our proposed test for conditional independence is methodologically surprising, numerically attractive and theoretically beautiful.

Our second goal in this paper is to introduce another important application of conditional independence: conditional screening for high dimensional regressions. In many scientific areas, a tremendous number of explanatory covariates are collected, while only a few are truly important to the response. Under some circumstances, the effects of covariates on the response may depend on certain exposure variables (sometimes also referred to as controlling or confounding variables), such as time or environmental factors, with an unknown pattern. For instance, genetic effects on the growth of plants may vary with certain environmental indices. Suppose $\mathbf{x} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ is the high dimensional covariate vector, Y is the response variable and $\mathbf{u}$ stands for the vector of exposure variables. In general, the goal of conditional feature screening is to remove the covariates that satisfy the conditional independence $Y \perp\!\!\!\perp X_k \mid \mathbf{u}$. To account for the effects of exposure variables $\mathbf{u}$, [1,8,17] studied several conditional screening methods based on partial and conditional correlation learning. [18] adopted quantile partial correlation to measure the conditional contribution of each covariate to the response at different quantile levels. These existing conditional screening methods assume that the response depends on the covariates linearly given exposure variables. In the analysis of high dimensional data, we often lack prior information on the dependence structure [15,35], and the linearity assumption is easily violated. Therefore, conditional feature screening for nonlinearly dependent data is highly desired. Following the concept of conditional independence test, we propose a conditional dependence metric to measure the possibly nonlinear effects of the exposure variables and dependence of the response upon the covariates. Inspired by [24], we suggest a bootstrap method to estimate the false discovery rate, which is a soft threshold guide for determining the cutoff of this conditional screening procedure. We show that our proposed conditional screening approach possesses both the desirable sure screening [7] and ranking consistency properties [35]. In addition, this conditional screening approach is model-free in that it can capture arbitrarily nonlinear dependence. This model-free property, together with the desirable theoretical properties, makes our proposed screening approach very appealing in high dimensional data analysis.

In Section 2, we develop a distribution-free test for conditional independence and examine its asymptotic behaviors. We demonstrate its finite sample performance through some illustrative examples. In Section 3, we propose a conditional screening procedure and study its theoretical properties carefully. We also suggest a bootstrap approach to estimate its false discovery rate. In Section 4, we conduct Monte Carlo simulations to evaluate the finite sample performances of the conditional screening approach and apply this conditional screening approach to analyze the gene expression data. We give some brief discussions in Section 5. All technical proofs are relegated to Appendix.

2. A distribution-free test for conditional independence

2.1. The rationale

In this section we show the equivalence between the conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{u}$ and the unconditional independence $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u})$ under mild conditions. We remark here that, if $\mathbf{u}$ follows uniform distribution, $\{\mathbf{u}, F_1(X \mid \mathbf{u})\}$ and $\{\mathbf{u}, F_2(Y \mid \mathbf{u})\}$ are the respective Rosenblatt transformations of $(\mathbf{u}, X)$ and $(\mathbf{u}, Y)$ [22]. In the present context, we do not require $\mathbf{u}$ to follow uniform distribution, but we still refer to the transformations, $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$, as Rosenblatt transformations.

We first illustrate how to convert the problem of testing conditional independence into the problem of testing unconditional independence.

Lemma 1. *Assume the distribution functions of X and Y, conditional on $\mathbf{u}$, are continuous for all possible values of $\mathbf{u}$.*

1. *The conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{u}$ implies the unconditional independence $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u})$;*
2. *With the assumption $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$, the unconditional independence $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u})$ implies the conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{u}$.*

If we aim to test the conditional independence $Y \perp\!\!\!\perp X \mid \mathbf{u}$ through testing the unconditional independence $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u})$, the first statement in Lemma 1 ensures that the test size will be well controlled as long as both X and Y have continuous distribution functions conditional on $\mathbf{u}$. The second statement in Lemma 1 guarantees that the test will have nontrivial power against all alternatives satisfying $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$. Both $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$ are marginally independent of $\mathbf{u}$, respectively. Although the assumption $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$ cannot be derived from $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp \mathbf{u}$ and $F_2(Y \mid \mathbf{u}) \perp\!\!\!\perp \mathbf{u}$, it holds true for a wide range of cases. For instance, X and Y are generated from the additive models, e.g., $X = g_1(\mathbf{u}) + \varepsilon$ and $Y = g_2(\mathbf{u}) + \delta$, where (ε, δ) are assumed to be independent of $\mathbf{u}$. In this case, $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$ holds exactly true. When X, Y and $\mathbf{u}$ are jointly normal, the additive model structures are satisfied and the assumption also holds. In general, we regard $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$ as a mild assumption. In fact, even though the requirement $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$ is not fulfilled, the test size can still be well controlled.

The rationale of testing conditional independence through testing unconditional independence is parallel to that of testing conditional independence using partial correlation [14]. Suppose for now X, Y and $\mathbf{u}$ are jointly normal, $X = E(X \mid \mathbf{u}) + \varepsilon$ and $Y = E(Y \mid \mathbf{u}) + \delta$, where $(\varepsilon, \delta) \perp\!\!\!\perp \mathbf{u}$. The partial correlation calculates the correlation between ε and δ, and the zero correlation is equivalent to the conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{u}$. Without normality assumption and the additive structure, $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$ can be viewed to characterize the respective variabilities of the implicit errors ε and δ. Therefore, our test can be regarded as a generalization of partial correlation for testing the conditional independence where neither the additive structure nor the normality is required.

Lemma 1 indicates that all tests for unconditional independence can be used to test conditional independence. We use the correlation introduced by [3] to test the independence between $V \overset{\text{def}}{=} F_1(X \mid \mathbf{u})$ and $W \overset{\text{def}}{=} F_2(Y \mid \mathbf{u})$. Specifically, let $F_V(v)$ and $F_W(w)$ denote the respective marginal distribution functions of V and W, and $F_{V,W}(v, w)$ refers to the joint distribution function of (V, W). The Blum–Kiefer–Rosenblatt (BKR) correlation [3] between V and W is defined as:

$$\rho^{CI} \overset{\text{def}}{=} \int_{\mathbb{R}^1} \int_{\mathbb{R}^1} \{F_{V,W}(v, w) - F_V(v)F_W(w)\}^2 dF_V(v) dF_W(w). \tag{1}$$

It is easy to verify that ρ^{CI} is nonnegative, $\rho^{CI} = 0$ if and only if V and W are independent. It also remains invariant after order-preserving transformation of V or W, because we merely use their distribution functions.

We remark here that some other metrics, such as distance correlation [28], projection correlation [36] and a modification of the BKR correlation [33], can also be used to test the independence between V and W. We use the BKR correlation because the distribution of its sample estimate is asymptotically distribution-free.

2.2. Sample estimate and its asymptotic properties

In practice, we adopt the Nadaraya–Watson estimator for each term of (1) based on the random sample $\{(X_i, Y_i, \mathbf{u}_i), i = 1, \ldots, n\}$, where n is the sample size. Let

$$\widehat{F}_1(x \mid u) \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(\mathbf{u}_i - u) I(X_i \le x) \Big/ n^{-1} \sum_{i=1}^{n} K_h(\mathbf{u}_i - u),$$

$$\widehat{F}_2(y \mid u) \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(\mathbf{u}_i - u) I(Y_i \le y) \Big/ n^{-1} \sum_{i=1}^{n} K_h(\mathbf{u}_i - u), \tag{2}$$

where $K_h(\mathbf{u}) \overset{\text{def}}{=} K(\mathbf{u}/h)/h^d$, d is the length of vector $\mathbf{u}$, $K(\cdot)$ is a product of d univariate kernel functions and h is the bandwidth. Set $\widehat{V}_i \overset{\text{def}}{=} \widehat{F}_1(X_i \mid \mathbf{u}_i)$ and $\widehat{W}_j \overset{\text{def}}{=} \widehat{F}_2(Y_j \mid \mathbf{u}_j)$. A natural estimator of ρ^{CI} is

$$\widehat{\rho}^{CI} \overset{\text{def}}{=} n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ F_{n,\widehat{V},\widehat{W}}(\widehat{V}_i, \widehat{W}_j) - F_{n,\widehat{V}}(\widehat{V}_i) F_{n,\widehat{W}}(\widehat{W}_j) \right\}^2, \tag{3}$$

where $F_{n,\widehat{V}}$, $F_{n,\widehat{W}}$ and $F_{n,\widehat{V},\widehat{W}}$ are the empirical distribution functions.

We consider the following hypothesis:

$$H_0 : Y \perp\!\!\!\perp X \mid \mathbf{u}, \text{ versus } H_1 : \text{ otherwise}. \tag{4}$$

Lemma 1 implies that the test (4) is equivalent to

$$H_0 : \rho^{CI} = 0, \text{ versus } H_1 : \rho^{CI} > 0, \tag{5}$$

under the assumption that $\{F_1(X \mid \mathbf{u}), F_2(Y \mid \mathbf{u})\} \perp\!\!\!\perp \mathbf{u}$. It is thus natural to use $\widehat{\rho}^{CI}$ in (3) to test for the hypotheses (5). We assume the following regularity conditions to derive asymptotic properties. These assumptions may not be the weakest possible though, they are assumed to facilitate technical derivations.

(A1) *(The Kernel Function)* The kernel $K(\cdot)$ is symmetric about zero and Lipschitz continuous. In addition, it satisfies

$$\int_{-1}^{1} K(t)dt = 1, \quad \int_{-1}^{1} t^{i-1} K(t)dt = 0, \ 0 \le i \le m-1, \quad 0 \ne \int t^m K(t)dt = \nu_m < \infty.$$

It is bounded uniformly such that $\sup_{\mathbf{u} \in \mathbb{U}} |K(\mathbf{u})| \le M_K < \infty$.

(A2) *(The Density)* The probability density functions of $\mathbf{u}$, V and W, denoted by $f(\mathbf{u})$, $f_V(v)$ and $f_W(w)$, respectively, are bounded away from 0 to infinity.

(A3) *(The Derivatives)* The $(m-1)$th derivatives of both $f(\mathbf{u})$ and $F(\cdot \mid \mathbf{u})f(\mathbf{u})$ are locally Lipschitz-continuous with respect to $\mathbf{u}$.

(A4) *(The Bandwidth)* The bandwidth h satisfies $h = O(n^{-\theta})$, where $(2m)^{-1} < \theta < (2d)^{-1}$, and d is the length of vector $\mathbf{u}$ and m is defined in (A1).

Under regularity conditions (A1)–(A4), Theorem 1 states the asymptotic distributions of $\widehat{\rho}^{CI}$ under the null and the alternative hypothesis.

Theorem 1. *Assume conditions (A1)–(A4) are fulfilled.*

(i) *Under the null hypothesis H_0 of* (4) and (5)*,*

$$n\,\widehat{\rho}^{CI} \xrightarrow{d} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{\chi^2_{ij}(1)}{\pi^4 i^2 j^2},\quad \text{as } n\to\infty, \tag{6}$$

where $\chi^2_{ij}(1)$, $i,j = 1,2,\ldots$ are independent and identically distributed chi-square random variables with one degree of freedom, and $\xrightarrow{d}$ denotes convergence in distribution.

(ii) *Under the alternative hypothesis H_1 of* (4) and (5)*,*

$$n^{1/2}\left(\widehat{\rho}^{CI}-\rho^{CI}\right) \xrightarrow{d} \mathcal{N}\{0, 4\,var(Z)\},\quad \text{as } n\to\infty,$$

where Z is defined in (20) in the Appendix, and $var(Z)$ denotes the variance of Z.

Theorem 1 yields a distribution-free conditional independence test, referred to as the BKR test. An appealing property of the BKR test is that the asymptotic null distribution of $\widehat{\rho}^{CI}$ does not involve any unknown tunings.

To put this test into practice, we approximate the asymptotic null distribution with

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\chi^2_{ij}(1)/\pi^4 i^2 j^2. \tag{7}$$

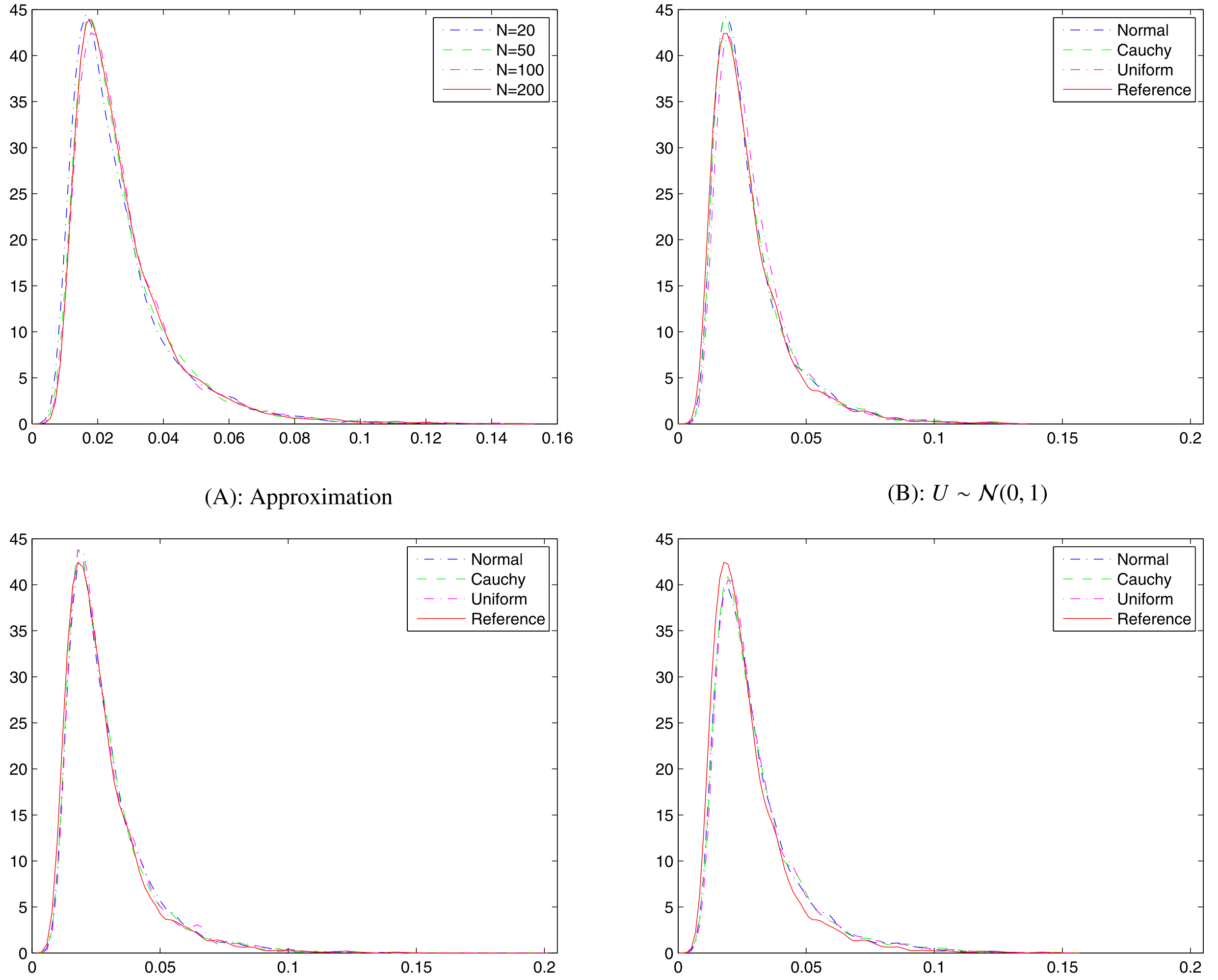
With a fixed N, we generate, say, 10,000 random variables through (7). The critical value c_α is the $\alpha \times 100\%$ quantile of these 10,000 realizations of (7) and α is a specified significance level. We reject the null hypothesis $Y \perp\!\!\!\perp X \mid \mathbf{u}$ if $n\widehat{\rho}^{CI} \geq c_\alpha$. Fig. 1(A) shows that the density function in (7) with $N = 20$ appears very close to that with $N = 100$, indicating that $N = 20$ is sufficient for a good approximation. To empirically assess its performance, we consider two examples.

Example 1. Let $Y = U + \varepsilon$ and $X = U + \delta$, where U is generated from (i) standard normal distribution $\mathcal{N}(0, 1)$, (ii) uniform distribution $\mathcal{U}(0, 1)$ and (iii) exponential distribution with mean 1/5, respectively. The error terms ε and δ are independently drawn from (i) standard normal, (ii) Cauchy and (iii) uniform distributions, respectively. In all these scenarios, X and Y are conditionally independent given U. The sample size n is set to be 100. We display the estimated density curves of $n\,\widehat{\rho}^{CI}$ based on 1000 repetitions in Fig. 1(B)–(D). All the estimated density curves are close to the reference, which is obtained by approximating the asymptotic null distribution in (6) with $N = 100$.

Example 2. In this example, we consider the nonlinearly conditional dependence between X and Y given $\mathbf{u} = (U_1, U_2)^\top$. Let $Y = U_1U_2 + \varepsilon$ and $X = U_1 + U_2 + 0.75\kappa\varepsilon^2 + c\delta$. We draw $\mathbf{u}$ from $N(0, \boldsymbol{\Sigma})$ whose covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{2\times 2}$ has entries $\sigma_{ii} = 1$, $i = 1, 2$ and $\sigma_{ij} = 0.5$ for $i \neq j$. The error ε is generated independently from standard normal distribution. The error δ is drawn independently of ε, from (i) standard normal and (ii) Cauchy distributions with $c = 1$ and $c = 1/2$, respectively. The value of κ describes the degree of nonlinear dependency of Y and X given $\mathbf{u}$, where $\kappa = 0$ corresponds to the null hypothesis. We set the sample size $n = 100$ and repeat the simulations 1000 times. The power curves in this example are illustrated in Fig. 2, in comparison with those from the standard test based on the partial correlation between X and Y given $\mathbf{u}$ [14], the test using maximal nonlinear conditional correlation [11] and the test based on conditional distance correlation [30]. All four tests have sizes near 0.05 at $\kappa = 0$. The standard partial correlation test fails to detect the nonlinear conditional dependence in both scenarios. The power of the BKR test rapidly approaches 1 as κ increases. On the other hand, the conditional distance correlation is comparative to BKR test when the error term is normal, but deteriorates sharply when the error term is heavy-tailed. In conclusion, the BKR test is distribution-free, robust to the extreme values, and is effective to detect nonlinearly conditional dependence.

3. Conditional feature screening for high dimensional data

In this section, we propose a conditional screening method for high dimensional data through the concept of conditional independence. Let $Y \in \mathbb{R}^1$ be the response variable and $\mathbf{x} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be the associated covariate



(A): Approximation

(B): $U \sim \mathcal{N}(0,1)$

(C): $U \sim \mathcal{U}(0,1)$

(D): $U \sim Exp(1/5)$

Fig. 1. (A): We approximate the limiting null distribution of $\widehat{\rho}^{CI}$ with the first $N \times N$ terms with $N = 20$, 50, 100 and 200. (B)–(D): The estimated density functions of $n\widehat{\rho}^{CI}$ in Example 1, in comparison with the limiting null distribution in (6).

vector. Assume that the effect of $\mathbf{x}$ on Y may vary with the exposure variables $\mathbf{u} = (U_1, \ldots, U_d)^\top \in \mathbb{R}^d$. One example is that the influence of active genes on body mass index may rely on the age. Let $\mathcal{A} \subseteq \{1, \ldots, p\}$ be the index set of the truly active covariates, and $\mathcal{A}^c$ be the truly unimportant ones. Specifically, $\mathcal{A}$ is the complement of $\mathcal{A}^c$, which is a collection of X_ks that satisfy

$$Y \perp\!\!\!\perp X_k \mid \mathbf{u}.$$

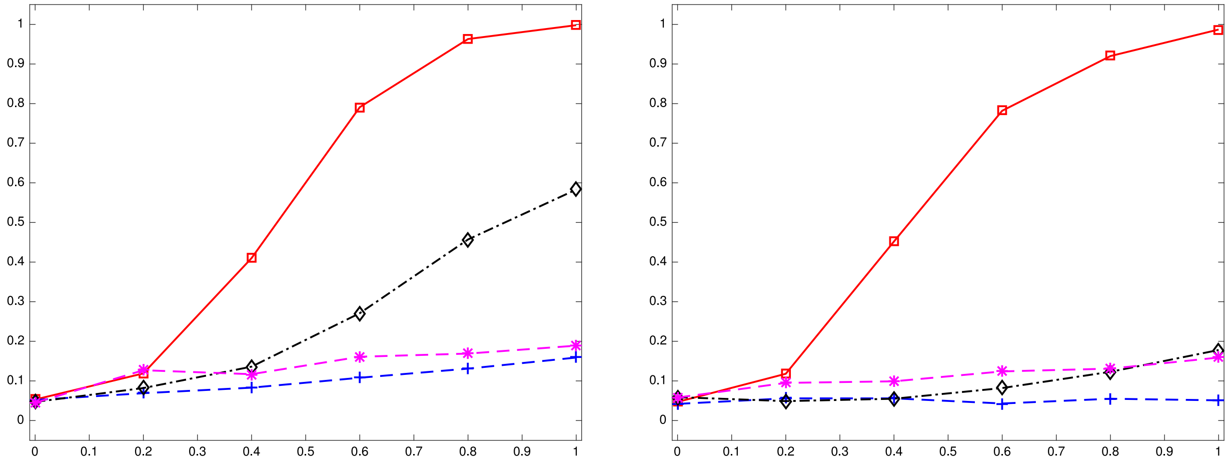
That is, given $\mathbf{u}$, X_k is not predictive for Y. Inspired by the test for conditional independence, we define the marginal utility of each covariate X_k by

$$\rho_k^{CI} = \int_{\mathbb{R}^1}\int_{\mathbb{R}^1} \{F_{V_k,W}(v_k, w) - F_{V_k}(v_k)F_W(w)\}^2 dF_{V_k}(v_k)dF_W(w), \tag{8}$$

where $V_k = F_1(X_k \mid \mathbf{u})$ and $W = F_2(Y \mid \mathbf{u})$. The sample individual BKR correlation $\widehat{\rho}_k^{CI}$ can be obtained in the same fashion as (3). We then suggest a conditional screening procedure to identify the active set by

$$\widehat{\mathcal{A}} = \{k : \widehat{\rho}_k^{CI} \text{ ranks in the top } \ell \text{ among all } \widehat{\rho}^{CI}\text{'s}, \text{ for } 1 \le k \le p\}, \tag{9}$$

where ℓ is the user-specified threshold value. We refer to our proposed sure independence conditional screening procedure based on BKR correlation as BKR-CSIS in the rest of this paper. In practice, we suggest using the false discovery



(A): $\delta \sim \mathcal{N}(0,1)$, $c=1$

(B): $\delta \sim t(1)$, $c=1/2$

Fig. 2. The empirical power curves in Example 2, based on the partial correlation test (blue dashed line marked with plus), maximal nonlinear conditional correlation test (magenta dashed line marked with star), conditional distance correlation test (black dashdot line marked with diamond) and BKR test (red solid line marked with square), respectively. The horizontal axis: values of κ varying from 0 to 1. The vertical axis: the power, increasing from 0 to 1.

rate (FDR for short) of BKR-CSIS to choose the screening threshold ℓ, which is defined as

$$\mathrm{FDR} \overset{\mathrm{def}}{=} E\left(\frac{|\widehat{\mathcal{A}} \cap \mathcal{A}^c|}{|\widehat{\mathcal{A}}|}\right). \tag{10}$$

The FDR yields the expected proportion of identified covariates that are not important. To control the FDR at a pre-specified level α^*, we define

$$\hat{t} = \inf\left[0 \le t \le 1 : \sum_{k\in\mathcal{A}^c} I(\widehat{\rho}_k^{CI} \ge t) \Bigg/ \max\left\{\sum_{1\le k\le p} I(\widehat{\rho}_k^{CI} \ge t), 1\right\} \le \alpha^*\right].$$

We retain the covariates X_k if $\widehat{\rho}_k^{CI} \ge \hat{t}$. To be precise, $\widehat{\mathcal{A}} \overset{\mathrm{def}}{=} \{k : \widehat{\rho}_k^{CI} \ge \hat{t},\ \text{for } 1 \le k \le p\}$. The estimated model size is given by

$$\ell = \sum_{1\le k\le p} I(\widehat{\rho}_k^{CI} \ge \hat{t}).$$

We establish the properties for the BKR-CSIS.

Theorem 2. *Suppose the bandwidth $h = O(n^{-\theta})$, under the conditions (A1)–(A4), for any $0 < 2\gamma + \theta d/2 \le 1$, if p satisfies $n^4 p \exp(-cn^{1-2\gamma-\theta d/2}) \to 0$ for some positive constant c, then*

$$\Pr\left(\max_{1\le k\le p} |\widehat{\rho}_k^{CI} - \rho_k^{CI}| \ge cn^{-\gamma}\right) \le O\left\{n^4 p \exp(-cn^{1-2\gamma-\theta d/2})\right\}.$$

(i) *If we further assume for some $c > 0$, such that $\min_{k\in\mathcal{A}}\ \rho_k^{CI} \ge 2cn^{-\gamma}$, then*

$$\Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \ge 1 - O\left\{n^4 |\mathcal{A}| \exp(-cn^{1-2\gamma-\theta d/2})\right\}. \tag{11}$$

(ii) *If we further assume that $\min_{k\in\mathcal{A}}\ \rho_k^{CI} - \max_{k\in\mathcal{A}^c}\ \rho_k^{CI} \ge 0$, then*

$$\liminf_{n\longrightarrow\infty}\left\{\min_{k\in\mathcal{A}} \widehat{\rho}_k^{CI} - \max_{k\in\mathcal{A}^c} \widehat{\rho}_k^{CI}\right\} > 0 \textit{ in probability}. \tag{12}$$

(iii) *If we further assume $\min_{k\in\mathcal{A}}\ \rho_k^{CI} - \max_{k\in\mathcal{A}^c}\ \rho_k^{CI} \ge 2cn^{-\gamma}$ for some $c > 0$, then*

$$\min_{k\in\mathcal{A}} \widehat{\rho}_k^{CI} > \left(\min_{k\in\mathcal{A}}\ \rho_k^{CI} + \max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2, \quad \max_{k\in\mathcal{A}^c} \widehat{\rho}_k^{CI} < \left(\min_{k\in\mathcal{A}}\ \rho_k^{CI} + \max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2 \quad \textit{in probability},$$

where c may take different values at different places.

The sure screening property [7] of the BKR-CSIS procedure is stated in (11) of Theorem 2(i), which guarantees that the probability of including all the truly active covariates goes to one at an exponential rate by BKR-CSIS as $n \to \infty$. The sufficient condition for Theorem 2(i) is typically used in the feature screening literature, indicating that the minimal signal cannot be too small, although it is allowed to converge to zero at a slower rate than $n^{-\gamma}$. Furthermore, the ranking consistency property in (12) of Theorem 2(ii) ensures that the active covariates are consistently ranked on the top with an overwhelming probability. This property is on the basis of the typical assumption of population ranking consistency. Theorem 2(iii) indicates that we can identify all significant covariates and make no false discoveries with an overwhelming probability.

In practice, $\mathcal{A}^c$ and FDR are unknown. Motivated by [24], we propose a bootstrap approach to estimate FDR. The procedure is conducted in the following steps.

S1. Estimate the conditional distribution function of $F_2(y \mid u)$ by $\widehat{F}_2(y \mid u)$ in (2) based on the sample $\{(\mathbf{x}_i, Y_i, \mathbf{u}_i),\ i = 1, \ldots, n\}$.

S2. For each $i = 1, \ldots, n$, bootstrap Y_i^* from $\widehat{F}_2(y \mid u)$. Define the bootstrapped sample $\mathbf{y}^* \overset{\text{def}}{=} (Y_1^*, \ldots, Y_n^*)$.

S3. Repeat the step **S2** B times. Based on the bth bootstrap sample, $b = 1, \ldots, B$, calculate the kth sample BKR correlation $\widehat{\rho}_{k,b}^{CI}$ using the bootstrapped data $\mathbf{y}^*$ and $(X_{1k}, \ldots, X_{nk})$, $k = 1, \ldots, p$.

S4. For any given choice of ℓ, estimate the FDR with

$$\widehat{FDR} \overset{\text{def}}{=} \sum_{b=1}^{B} \sum_{k=1}^{p} I\left(\widehat{\rho}_{k,b}^{CI} > \widehat{\rho}_{*}^{CI}\right) \Big/ (B\ell), \tag{13}$$

where $\widehat{\rho}_{*}^{CI}$ is the smallest sample BKR correlation among the ℓ chosen covariates based on the original sample.

The numerator of (13) computes the average count of spurious correlations $\widehat{\rho}_{k,b}^{CI}$ that are ranked before the cutoff $\widehat{\rho}_{*}^{CI}$ and falsely included into the selected model. Thus, $\widehat{FDR}$ in (13) may serve as a reasonable approximation to true FDR defined in (10).

4. Numerical studies

4.1. The performance of conditional screening

In this section, we illustrate the finite sample performance of our proposed conditional screening procedure through Monte Carlo simulations. We consider two scenarios. In scenario 1, we generate $\mathbf{x} = (X_1, \ldots, X_p)^\top$ and an intermediate variable U^* from the multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_{(p+1)\times(p+1)}$. We define the exposure variable U by $U \overset{\text{def}}{=} \Phi(U^*)$, where $\Phi(\cdot)$ is cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$. The error term is independently drawn from $\mathcal{N}(0, 1)$. In scenario 2, we draw $\mathbf{x}$ and U^* from multivariate t distribution with $\boldsymbol{\Sigma}_{(p+1)\times(p+1)}$ and three degrees of freedom. The exposure variable U is obtained from the cumulative distribution function of $t(3)$ distribution. The error term follows student's t distribution with three degrees of freedom. In both scenarios, we set $n = 200$ and $p = 1000$. The covariance matrix is defined as $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{(p+1)\times(p+1)}$, where $\rho = 0.5$ in scenario 1 and $\rho = 0.7$ in scenario 2. Six screening methods, BKR-CSIS, CC-SIS [17], NIS [8], the screening based on partial correlation (PC-SIS), maximal nonlinear conditional correlation [11, MNC-SIS] and conditional distance correlation [30, CDC-SIS] are compared. The BKR-CSIS, NIS [8] and PC-SIS were implemented with Matlab. The simulation studies of CDC-SIS [30] were conducted with the R package cdcsis. The CC-SIS [17] and MNC-SIS [11] were implemented by R with the codes provided by the authors. We use the following three criteria to assess their performances.

1. R_j: The median of the ranks of each important covariate out of 1000 replications.
2. $\mathcal{S}$: The minimum model size to ensure that all important covariates are retained after screening. It is the largest rank of the truly important covariates. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$.
3. FDR: The average number of false discovery rates out of 1000 replications.

Example 3. We generate the response Y from the following five models respectively:

$$\begin{aligned}
\text{Model (4.1)}: &\quad Y = \mathbf{x}^\top \boldsymbol{\beta}_c(U) + \varepsilon/3;\\
\text{Model (4.2)}: &\quad Y = \exp\{2\mathbf{x}^\top \boldsymbol{\beta}(U)\} + \varepsilon;\\
\text{Model (4.3)}: &\quad Y = \mathbf{x}^\top \boldsymbol{\beta}_{1,2,3}(U) + \exp\{2\mathbf{x}^\top \boldsymbol{\beta}_{4,5}(U)\} + \varepsilon;\\
\text{Model (4.4)}: &\quad Y = \mathbf{x}^\top \boldsymbol{\beta}_{1,2,3}(U) + \exp\{2\mathbf{x}^\top \boldsymbol{\beta}_{4,5}(U) + \varepsilon\};\\
\text{Model (4.5)}: &\quad Y = \{\mathbf{x}^\top \boldsymbol{\beta}_{1,2}(U)\}^2 + \{\mathbf{x}^\top \boldsymbol{\beta}_{3,4}(U)\}^2 + \{\beta_5(U)X_5\}^2 + \varepsilon,
\end{aligned}$$

where $\boldsymbol{\beta} = \{\beta_1(u), \ldots, \beta_5(u), 0 \ldots, 0\}^\top$, $\boldsymbol{\beta}_c = \{\beta_1(u), 0.8\beta_2(u), 0.6\beta_3(u), 0.4\beta_4(u), 0.2\beta_5(u), 0 \ldots, 0\}^\top$ and $\boldsymbol{\beta}_{i,j,k} = \{0, \ldots, 0, \beta_i(u), \beta_j(u), \beta_k(u), 0, \ldots, 0\}^\top$ for $i, j, k \in \{1, \ldots, 5\}$. We set $\beta_1(u) = 2I(u > 0.4)$, $\beta_2(u) = 1 + u$, $\beta_3(u) = (2-3u)^2$, $\beta_4(u) = 2\sin(2\pi u)$, $\beta_5(u) = \exp\{u/(u+1)\}$. Model (4.1) is a homoscedastic varying coefficient model, and Models

Table 1
The median of R_j of each truly important covariate for Example 3. The smaller R_j is, the better the screening method performs.

	Method	Scenario 1					Scenario 2				
		X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
Model (4.1)	CC-SIS	2	1	3	4	5	2	1	3	4	5
	NIS	2	1	3	4	27	3	1	2	4	40
	PC-SIS	2	1	3	4	6	3	1	2	4	6
	MNC-SIS	2	1	2	6	93	3	1	2	4	9
	CDC-SIS	2	1	3	4	6	2	1	3	4	5
	BKR-CSIS	2	1	3	4	6	2	1	3	4	5
Model (4.2)	CC-SIS	6	3	6	19	37	114	89	77	83	119
	NIS	116	53	76	152	188	290	253	219	280	367
	PC-SIS	47	24	41	117	121	224	213	192	307	363
	MNC-SIS	4	2	2	3	6	4	2	2	3	4
	CDC-SIS	30	12	21	87	87	162	121	152	239	264
	BKR-CSIS	2	1	3	5	4	3	1	3	4	4
Model (4.3)	CC-SIS	390	309	99	2	1	416	319	187	38	11
	NIS	475	443	243	13	27	468	390	254	116	85
	PC-SIS	455	416	187	9	8	419	355	211	86	58
	MNC-SIS	61	13	5	1	2	29	5	3	1	2
	CDC-SIS	478	402	164	3	2	552	441	279	88	52
	BKR-CSIS	4	2	3	5	1	5	2	3	4	1
Model (4.4)	CC-SIS	412	313	109	2	1	434	337	205	46	18
	NIS	489	438	233	14	30	461	401	265	122	96
	PC-SIS	484	402	201	10	8	432	369	225	97	64
	MNC-SIS	61	14	5	1	2	37	6	3	1	2
	CDC-SIS	484	382	179	4	2	537	443	286	98	57
	BKR-CSIS	4	2	3	5	1	5	2	3	4	1
Model (4.5)	CC-SIS	2	2	7	50	255	5	4	8	34	101
	NIS	3	2	12	113	284	61	25	34	136	264
	PC-SIS	22	41	248	339	410	229	246	241	327	381
	MNC-SIS	459	442	436	469	496	463	356	365	319	390
	CDC-SIS	2	1	4	5	22	4	2	2	6	23
	BKR-CSIS	7	4	19	13	32	6	3	5	4	8

(4.2)–(4.5) contain nonlinear effects of covariates depending on U. In Model (4.1), $E\{\beta_4(U)\} = 0$, leading to conditional dependence but marginally independence between X_4 and Y.

The simulation results of R_j and $\mathcal{S}$ are charted in Tables 1 and 2, respectively. We expect R_j not to exceed 5 and $\mathcal{S}$ to be close to 5. It can be seen that, the BKR-CSIS performs the best under most model settings. The active covariates rank on the top and the medians of the minimum model size $\mathcal{S}$ are very close to the number of truly important covariates across most scenarios, which illustrates the ranking consistency property of BKR-CSIS. Both CC-SIS and NIS are designed for the homoscedastic varying coefficient model while PC-SIS aims at capturing the linear dependence between the covariates and response given U. Thus, these three methods perform satisfactorily in Model (4.1). However, their stories become totally different in other models. They are not able to detect nonlinear conditional dependence and barely rank the important covariates above the unimportant ones. In particular, the medians of the minimum model size $\mathcal{S}$ of NIS in Models (4.2)–(4.5) are all above 300, which is much larger than the sample size $n = 200$. Both MNC-SIS and CDC-SIS capture conditional independence, but MNC-SIS loses power in Model (4.5), and performance of CDC-SIS is easily influenced by heavily-tailed distributions of $\mathbf{x}$ and Y.

The simulation results of FDR are tabulated in Table 3. The cardinality of the screened model ℓ is chosen to be 5, 10, 15, and 20. The corresponding oracle values for FDR are 0, $1/2$, $2/3$, $3/4$. It is obvious that BKR-CSIS outperforms the other five methods with relatively low FDR. In Models (4.2)–(4.4), the FDR values of CC-SIS, NIS, PC-SIS and CDC-SIS are close to one for $\ell = 20$, indicating that their submodels are made up of unimportant covariates. By contrast, the BKR-CSIS barely makes false identification. In Model (4.5), CDC-SIS and BKR-CSIS have similar performances since they are both able to capture nonlinear conditional dependence. However, for MNC-SIS, the FDR is close to one in this model.

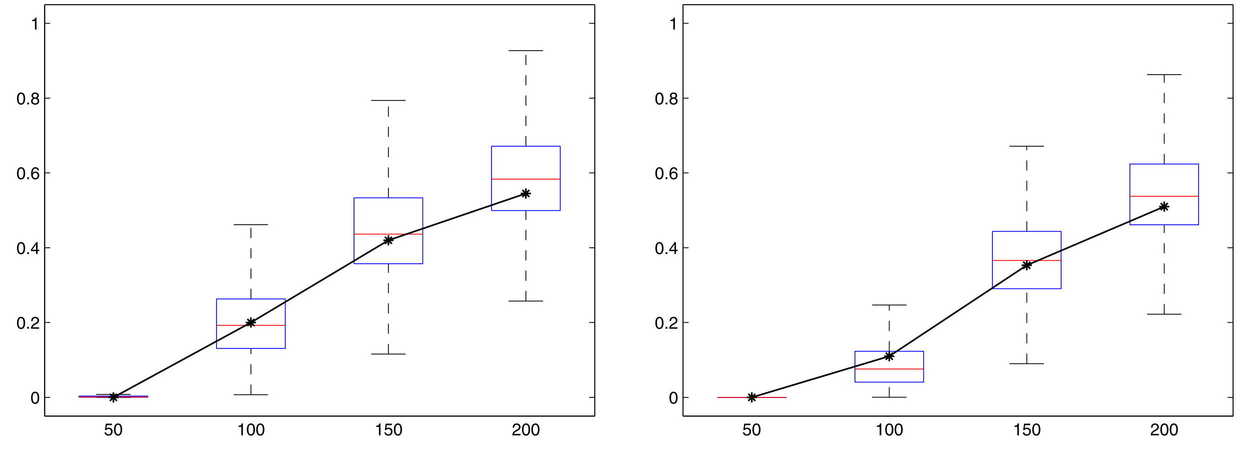
4.2. The accuracy of estimated FDR

Example 4. We employ the following model for generating the response variable:

$$\text{Model (4.6)}: \quad Y = \sum_{j=1}^{5} \sum_{i=20(j-1)+1}^{20j} \beta_j(U) X_i + \varepsilon,$$

Table 2
The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ for Example 3. The smaller $\mathcal{S}$ is, the better the screening method performs.

	Method	Scenario 1					Scenario 2				
		5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
Model (4.1)	CC-SIS	5	5	5	9	65	5	5	5	5	37
	NIS	5	8	28	109	420	5	8	49	238	686
	PC-SIS	5	5	6	12	84	5	5	7	50	628
	MNC-SIS	6	27	102	283	686	5	5	10	36	250
	CDC-SIS	5	5	6	15	138	5	5	5	5	6
	BKR-CSIS	5	5	7	18	120	5	5	5	5	6
Model (4.2)	CC-SIS	14	42	95	199	509	52	174	343	543	794
	NIS	129	320	524	746	931	140	394	602	823	966
	PC-SIS	73	233	480	740	949	117	378	619	842	967
	MNC-SIS	5	6	11	31	175	5	5	5	8	41
	CDC-SIS	44	190	354	591	849	122	363	627	837	966
	BKR-CSIS	5	5	5	5	8	5	5	5	5	6
Model (4.3)	CC-SIS	129	344	579	782	959	128	377	593	794	947
	NIS	195	513	716	861	968	164	440	684	857	966
	PC-SIS	163	482	716	866	976	115	409	677	868	976
	MNC-SIS	7	33	109	265	633	5	12	39	127	506
	CDC-SIS	200	451	630	791	943	180	546	762	895	977
	BKR-CSIS	5	5	5	6	11	5	5	6	6	7
Model (4.4)	CC-SIS	122	370	588	792	962	133	393	617	792	954
	NIS	214	511	713	856	975	185	460	699	870	972
	PC-SIS	183	475	711	876	979	125	428	701	879	980
	MNC-SIS	7	31	119	286	685	6	13	48	161	561
	CDC-SIS	199	442	659	815	953	186	538	763	898	978
	BKR-CSIS	5	5	5	6	10	5	5	5	6	7
Model (4.5)	CC-SIS	16	120	321	599	905	9	69	208	453	764
	NIS	33	170	376	650	918	52	257	500	754	921
	PC-SIS	142	492	728	873	979	135	516	763	896	981
	MNC-SIS	491	718	851	937	990	345	636	800	909	979
	CDC-SIS	6	13	27	60	185	5	13	48	161	588
	BKR-CSIS	16	35	62	109	249	6	10	18	35	94



$n = 200$ $n = 400$

Fig. 3. Boxplots of estimated FDR with true FDR (solid line marked with star) in Example 4. The horizontal axis is the threshold ℓ and the vertical is the FDR value.

where $\mathbf{x}$, U, ε and $\beta_j(u)$ for $j = 1, \ldots, 5$ have the same setup as the Scenario 1 of Example 3 with $\rho = 0.9$. The first 100 covariates are truly active in Model (4.6). Therefore, for $\ell \in \{50, 100, 150, 200\}$, the optimally oracle values are 0, 0, 1/3, 1/2 for the FDR.

To evaluate the bootstrap estimating procedure of FDR, we provide the boxplots of estimated FDR out of 1000 replications, with the true simulated FDR marked as a star in Fig. 3. We can see that the estimated FDR performs satisfactorily in each scenario, implying the rationality of using the estimated FDR as a reference for determining ℓ in practice.

Table 3
The mean of FDR values for Example 3. The cardinality of the screened model ℓ is chosen to be 5, 10, 15, and 20. The corresponding oracle values for FDR are 0, 1/2, 2/3, 3/4.

	Method	Scenario 1				Scenario 2			
		$\ell = 5$	$\ell = 10$	$\ell = 15$	$\ell = 20$	$\ell = 5$	$\ell = 10$	$\ell = 15$	$\ell = 20$
Model (4.1)	CC-SIS	0.089	0.522	0.678	0.757	0.050	0.511	0.673	0.754
	NIS	0.206	0.576	0.710	0.780	0.258	0.602	0.729	0.794
	PC-SIS	0.123	0.530	0.680	0.759	0.163	0.557	0.701	0.773
	MNC-SIS	0.293	0.620	0.738	0.800	0.174	0.555	0.697	0.769
	CDC-SIS	0.120	0.533	0.683	0.761	0.015	0.502	0.667	0.751
	BKR-CSIS	0.135	0.538	0.686	0.762	0.017	0.502	0.667	0.750
Model (4.2)	CC-SIS	0.420	0.654	0.753	0.808	0.872	0.907	0.926	0.937
	NIS	0.896	0.924	0.937	0.945	0.977	0.978	0.981	0.982
	PC-SIS	0.788	0.861	0.894	0.915	0.960	0.966	0.970	0.972
	MNC-SIS	0.221	0.562	0.697	0.769	0.111	0.521	0.676	0.755
	CDC-SIS	0.763	0.838	0.875	0.896	0.943	0.959	0.964	0.967
	BKR-CSIS	0.021	0.501	0.667	0.750	0.009	0.500	0.667	0.750
Model (4.3)	CC-SIS	0.597	0.783	0.849	0.884	0.857	0.906	0.927	0.939
	NIS	0.885	0.912	0.928	0.938	0.959	0.967	0.970	0.973
	PC-SIS	0.815	0.876	0.904	0.920	0.936	0.950	0.958	0.962
	MNC-SIS	0.414	0.667	0.766	0.817	0.294	0.606	0.726	0.790
	CDC-SIS	0.719	0.836	0.881	0.904	0.937	0.953	0.960	0.964
	BKR-CSIS	0.100	0.506	0.668	0.751	0.104	0.500	0.667	0.750
Model (4.4)	CC-SIS	0.615	0.788	0.853	0.886	0.878	0.916	0.934	0.943
	NIS	0.893	0.917	0.932	0.942	0.964	0.970	0.973	0.976
	PC-SIS	0.823	0.881	0.909	0.924	0.939	0.952	0.961	0.964
	MNC-SIS	0.422	0.668	0.766	0.818	0.317	0.614	0.731	0.794
	CDC-SIS	0.733	0.843	0.887	0.909	0.939	0.955	0.961	0.966
	BKR-CSIS	0.089	0.505	0.668	0.751	0.083	0.500	0.667	0.750
Model (4.5)	CC-SIS	0.523	0.730	0.810	0.852	0.640	0.777	0.834	0.867
	NIS	0.610	0.770	0.833	0.868	0.802	0.871	0.902	0.919
	PC-SIS	0.842	0.898	0.923	0.937	0.940	0.958	0.964	0.969
	MNC-SIS	0.990	0.991	0.992	0.993	0.983	0.985	0.987	0.987
	CDC-SIS	0.359	0.614	0.725	0.786	0.416	0.655	0.754	0.808
	BKR-CSIS	0.696	0.774	0.815	0.845	0.452	0.625	0.721	0.780

Example 5. The model setup remains identical to Model (4.6) in Example 4. We design the coefficient functions as $\beta_1(u) = 1$, $\beta_2(u) = 0.8$, $\beta_3(u) = 1.2$, $\beta_4(u) = -0.8$, $\beta_5(u) = -1.2$, corresponding to the linear model. The same phenomenon is observed, thus the results are omitted to save space. This example shows that BKR-CSIS is also valid in the context of general linear models.

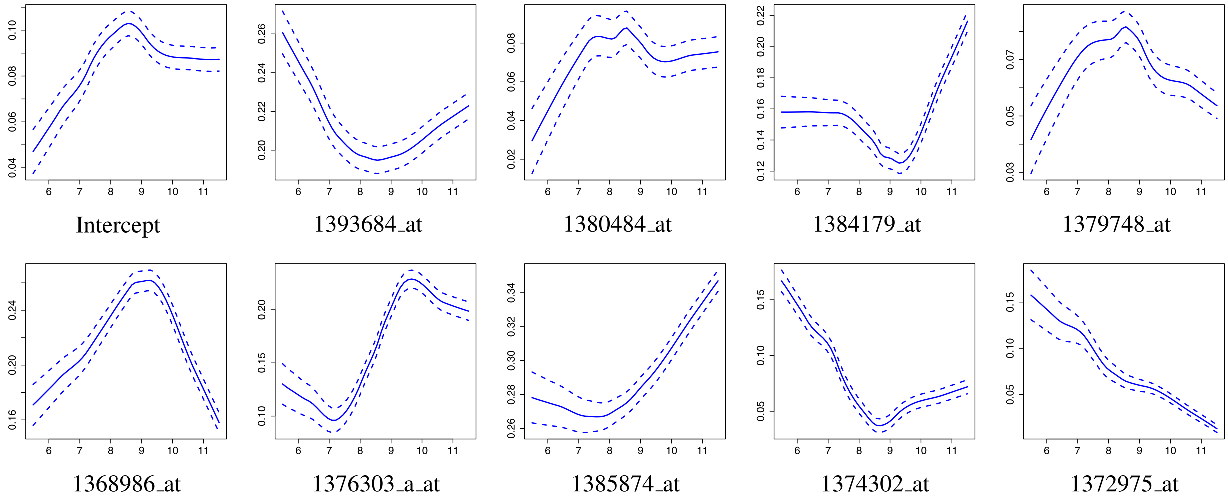
4.3. Real data analysis

We apply the proposed method to analyze the gene expression micro-array dataset collected by [23]. The dataset recorded 31,042 gene probes and 399 genetic markers from the eyes of 120 twelve-week-old male F_2 rats, selected from the offspring of SR/JrHsd males and SHRSP females. [23] removed a set of gene probes with a small expression signal and insufficient variation, leaving 18,976 probes. Among 399 genetic markers, a total of 64 markers without missing values are taken into consideration. [4] discovered that the gene TRIM32 at probe 1389163_at was associated with genetical multisystem human disease including the retina. Our interest is to identify the active explanatory gene probes for TRIM32 given the fully informative markers. The exposure variable U is defined as the first principal component of 64 genetic markers.

To select the significant gene probes that provide additional information for predicting TRIM32 when the 64 genetic markers are considered, we first apply CC-SIS, NIS, PC-SIS, MNC-SIS, CDC-SIS and BKR-CSIS to this dataset. For the BKR-CSIS procedure, the submodel size is determined by the estimated FDR. At an estimated FDR of 0.0001, 10 active gene probes are included in the screened model. For a fair comparison, other five methods are also performed to reduce the dimensionality of the involved genes to $\ell = 10$. We further conduct a post-screening variable selection procedure to fit an interpretable model. Following [17], we consider the varying coefficient model with SCAD penalty. The tuning parameter is chosen by the BIC criterion. Table 4 gives the information of the gene probes selected at least twice by different methods. The probes detected by only one method are omitted. Our BKR-CSIS procedure ranks the probe 1393684_at at the top, which is also identified by CC-SIS, PC-SIS, MNC-SIS and CDC-SIS methods. In fact, this probe is named "TBC1 domain family, member 12-predicted", and symbolized as "TBC1D12_predicted" by [10]. LASSO and adaptive LASSO also regard this gene as important [12]. The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/gene/23232) reports that this gene probe belongs to homo sapiens group and has already been known as an GTPase activator.

Table 4
Results for analyzing the rat eye data. The first twelve rows are the gene probes selected at least twice by different methods. The last two rows are the model size and the median squared prediction error (MSPE) of each method.

Selected gene probes	CC-SIS	NIS	PC-SIS	MNC-SIS	CDC-SIS	BKR-CSIS
1393684_at	✓		✓	✓	✓	✓
1380484_at	✓		✓		✓	✓
1384179_at	✓		✓		✓	✓
1368986_at					✓	✓
1376303_a_at					✓	✓
1385874_at	✓					✓
1381556_at	✓		✓			
1387347_at	✓		✓			
1390627_a_at	✓		✓			
1395542_at	✓				✓	
1379748_at				✓		✓
1374302_at				✓		✓
Model size	9	9	10	10	8	9
MSPE	0.0298	0.0206	0.0250	0.0171	0.0325	0.0125



Intercept 1393684_at 1380484_at 1384179_at 1379748_at

1368986_at 1376303_a_at 1385874_at 1374302_at 1372975_at

Fig. 4. The estimated coefficient functions of significant gene probes (the solid lines) and their 95% simultaneous confidence bands (the dashed lines).

We also report the model size and the median squared prediction error (MSPE) of each method in Table 4. The BKR-CSIS outperforms other five methods in terms of MSPE. It is also worth noting that the gene probes selected by NIS do not overlap with the other five methods while NIS outperforms PC-SIS, CC-SIS and CDC-SIS in terms of MSPE. NIS may select some genes that are predictive for the expression level of TRIM32, and need further biological research to confirm such a selection. For example, NIS selects the gene at probe 1370355_at, named "stearoyl-Coenzyme A desaturase 1" according to [10]. More details about this gene can also be found at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/gene/20249).

Fig. 4 shows that the selected probes by BKR-CSIS have marker-dependent impacts on the expression level of TRIM32. The estimated coefficient functions and their 95% simultaneous confidence bands [31] suggest that the selected probes are most likely to be truly relevant ones, since their penalized coefficient functions are unlikely to constantly be zero, and none of their simultaneous confidence bands completely well cover the horizontal zero line.

5. A brief discussion

In this paper, we propose a new metric based on Blum–Kiefer–Rosenblatt (BKR) correlation to test for conditional independence between two random variables given certain exposure variables. The asymptotic null and alternative distributions of the test are systematically studied. We then develop a model-free conditional sure independence screening procedure, BKR-CSIS, based on the BKR correlation, with a discussion of the desirable sure screening and the ranking consistency properties. Furthermore, we propose a bootstrap method to estimate the false discovery rate, which can be used as a reference for choosing the cutoff of the screening procedure. Simulation results and rat eye micro-array analysis demonstrate the performances of both the conditional independence test and the corresponding screening procedure.

6. Supplementary material

The Matlab codes to reproduce the numerical results in this paper are available from the authors upon request.

Acknowledgments

The authors would like to thank the Editor-in-Chief, Dietrich von Rosen, the Associate Editor and referees for their valuable comments and suggestions, and Biqing Liu and Xiang Li from School of Economics, Xiamen University for their technical support. The first two authors contributed equally to this paper. This work is supported by National Key Research and Development Program of China (2018YFC0830301), Beijing Natural Science Foundation, China (Z19J00009), National Natural Science Foundation of China (11731011, 11771361, 11671334, 11871409 and 11931014) and Ministry of Education, China of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002).

Appendix A. Lemmas

Lemma 2 (*Hoeffding's Inequality* [9]). *Let* $X_1, \ldots, X_n$ *be independent random variables, and* $\Pr(a_i \le X_i \le b_i) = 1$ *for* $i = 1, \ldots, n$. *Then*

$$\Pr\left\{|\overline{X} - E(\overline{X})| \ge t\right\} \le 2\exp\left\{-2n^2t^2 \Big/ \sum_{i=1}^{n}(b_i - a_i)^2\right\},\ t > 0.$$

Let $\mathcal{F}_1 \overset{\text{def}}{=} \left\{\widehat{V} : |\widehat{V} - V| \le \delta_{n,1}\right\}$ and $\mathcal{F}_2 \overset{\text{def}}{=} \left\{\widehat{W} : |\widehat{W} - W| \le \delta_{n,2}\right\}$, where $\delta_{n,k} = c_{1,k}(nh^d)^{-1/2}\ln n + c_{2,k}h^m$ for $k \in \{1, 2\}$. The following lemma is a direct application of Theorem 3.1 of [34].

Lemma 3. *For* $\varepsilon_n > 0$,

$$\begin{aligned}
&\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} n^{1/2}\left|n^{-1}\sum_{i=1}^{n}\left\{I\left(\widehat{V}_i \le v, \widehat{W}_i \le w\right) - I\left(V_i \le v, W_i \le w\right)\right\} - E\left\{I\left(\widehat{V} \le v, \widehat{W} \le w\right) - I\left(V \le v, W \le w\right)\right\}\right|\right. \\
&\qquad \left. \ge 8\varepsilon_n \right\} \\
&\le 3A\left(\varepsilon_n n^{-1/2}\right)^{-8}\exp\left(-\varepsilon_n^2/128\delta_n^2\right) + 4A\delta_n^{-16}\exp\left(-n\delta_n^2\right), \qquad (14) \\
&\Pr\left\{\sup_{\mathcal{F}_1} n^{1/2}\left|n^{-1}\sum_{i=1}^{n}\{I(\widehat{V}_i \le v) - I(V_i \le v)\} - E\left\{I(\widehat{V}_i \le v) - I(V_i \le v)\right\} \ge 8\varepsilon_n\right.\right\} \\
&\le 3A\left(\varepsilon_n n^{-1/2}\right)^{-4}\exp\left(-\varepsilon_n^2/128\delta_n^2\right) + 4A\delta_n^{-8}\exp\left(-n\delta_n^2\right), \qquad (15)
\end{aligned}$$

where $\delta_n^2 \ge c\max\{\delta_{n,1}, \delta_{n,2}\}$. Similar results in (15) hold for $W \overset{\text{def}}{=} F_2(Y \mid \mathbf{u})$.

Appendix B. Proofs related to the test

Proof of Lemma 1. Suppose X and Y have continuous conditional distribution functions for any given vector $\mathbf{u}$. Then $F_1(X \mid \mathbf{u})$ and $F_2(Y \mid \mathbf{u})$ are uniformly distributed for any given $\mathbf{u}$ and

$$\begin{aligned}
\Pr\{F_2(Y \mid \mathbf{u}) \le y, \mathbf{u} \le \widetilde{\mathbf{u}}\} &= E\left\{E\left[I\{F_2(Y \mid \mathbf{u}) \le y, \mathbf{u} \le \widetilde{\mathbf{u}}\} \mid \mathbf{u}\right]\right\} = E\left\{I(\mathbf{u} \le \widetilde{\mathbf{u}})E\left[I\{F_2(Y \mid \mathbf{u}) \le y\} \mid \mathbf{u}\right]\right\} \\
&= E\{I(\mathbf{u} \le \widetilde{\mathbf{u}})\}y = \Pr\{\mathbf{u} \le \widetilde{\mathbf{u}}\}\Pr\{F_2(Y \mid \mathbf{u}) \le y\},
\end{aligned}$$

which implies $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp \mathbf{u}$ and $F_2(Y \mid \mathbf{u}) \perp\!\!\!\perp \mathbf{u}$.

(i) If $X \perp\!\!\!\perp Y \mid \mathbf{u}$, then $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u}) \mid \mathbf{u}$, that is, $V \perp\!\!\!\perp W \mid \mathbf{u}$. [2] proved that if $V \perp\!\!\!\perp \mathbf{u}$ and $W \perp\!\!\!\perp \mathbf{u}$, the $V \perp\!\!\!\perp W \mid \mathbf{u}$ implies $V \perp\!\!\!\perp W$.

(ii) If $V \perp\!\!\!\perp W$ and the joint distribution $(V, W) \perp\!\!\!\perp \mathbf{u}$, then the conditional cumulative distribution functions of (V, W), V, W given $\mathbf{u}$ are

$$\begin{aligned}
F_{V,W|\mathbf{u}}(v, w \mid u) &= F_{V,W}(v, w) = F_V(v)F_W(w); \\
F_{V|\mathbf{u}}(v \mid u) &= F_{V,W|\mathbf{u}}(v, \infty \mid u) = F_{V,W}(v, \infty) = F_V(v); \\
F_{W|\mathbf{u}}(w \mid u) &= F_{V,W|\mathbf{u}}(\infty, w \mid u) = F_{V,W}(\infty, w) = F_W(w).
\end{aligned}$$

Thus, $F_{V,W|\mathbf{u}}(v, w \mid u) = F_{V|\mathbf{u}}(v \mid u)F_{W|\mathbf{u}}(w \mid u)$, i.e., $F_1(X \mid \mathbf{u}) \perp\!\!\!\perp F_2(Y \mid \mathbf{u}) \mid \mathbf{u}$. Take their inverse functions to obtain $X \perp\!\!\!\perp Y \mid \mathbf{u}$.

Proof of Theorem 1. We divide the proof into two steps.
Step 1. Under the null hypothesis H_0 in (4) and (5), we prove that

$$n\,\widehat{\rho}^{CI} \xrightarrow{d} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{\chi_{ij}^2(1)}{\pi^4 i^2 j^2},\ \text{as } n\to\infty.$$

The estimated ρ^{CI} can be written as

$$\widehat{\rho}^{CI}=\int_{\mathbb{R}^1}\int_{\mathbb{R}^1}\left\{F_{n,\widehat{V},\widehat{W}}(v,w)-F_{n,\widehat{V}}(v)F_{n,\widehat{W}}(w)\right\}^2 dF_{n,\widehat{V}}(v)dF_{n,\widehat{W}}(w),$$

where $F_{n,\widehat{V}}(v,w)$, $F_{n,\widehat{W}}(v,w)$ and $F_{n,\widehat{V},\widehat{W}}(v,w)$ are empirical distribution functions. We decompose $T_n(v,w)\overset{\text{def}}{=}F_{n,\widehat{V},\widehat{W}}(v,w)$ $-\,F_{n,\widehat{V}}(v)F_{n,\widehat{W}}(w)$ into $T_{n,1}(v,w)-T_{n,2}(v,w)$. The notations $T_{n,1}(v,w)$ and $T_{n,2}(v,w)$ will be defined shortly.

$$T_{n,1}(v,w)\overset{\text{def}}{=}F_{n,\widehat{V},\widehat{W}}(v,w)-F_{n,V,W}(v,w)-w\{F_{n,\widehat{V}}(v)-F_{n,V}(v)\}-v\{F_{n,\widehat{W}}(w)-F_{n,W}(w)\}+Q_n(v,w),$$

where $Q_n(v,w)\overset{\text{def}}{=}F_{n,V,W}(v,w)-wF_{n,V}(v)-vF_{n,W}(w)+vw$. We first deal with the term $F_{n,\widehat{V},\widehat{W}}(v,w)-F_{n,V,W}(v,w)$. Let $q_{1|\mathbf{u}}(v)$ and $q_{2|\mathbf{u}}(w)$ denote the vth and wth quantiles of X and Y conditional on $\mathbf{u}$, respectively.

Define $\widehat{q}_{1|\mathbf{u}}(v)\overset{\text{def}}{=}\inf\{x:\widehat{F}_1(x\mid u)\ge v\}$ and $\widehat{q}_{2|\mathbf{u}}(w)\overset{\text{def}}{=}\inf\{y:\widehat{F}_2(y\mid u)\ge w\}$. Applying (14) in Lemma 3 and the law of iterated expectations, we have

$$\begin{aligned}F_{n,\widehat{V},\widehat{W}}(v,w)-F_{n,V,W}(v,w)=E\Big[&\Pr\left\{\widehat{F}_1(X\mid\mathbf{u})\le v,\widehat{F}_2(Y\mid\mathbf{u})\le w\mid\mathbf{u}\right\}\\&-\Pr\left\{F_1(X\mid\mathbf{u})\le v,F_2(Y\mid\mathbf{u})\le w\mid\mathbf{u}\right\}\Big]+o_p(n^{-1/2}).\end{aligned}$$

We focus on the term $Pr\left\{\widehat{F}_1(X\mid\mathbf{u})\le v,\widehat{F}_2(Y\mid\mathbf{u})\le w\mid\mathbf{u}\right\}-Pr\left\{F_1(X\mid\mathbf{u})\le v,F_2(Y\mid\mathbf{u})\le w\mid\mathbf{u}\right\}$ first. With Taylor's expansion, as $nh^{2d}\to\infty$ and $nh^{4m}\to 0$, we obtain that

$$\begin{aligned}&\Pr\left\{X\le\widehat{q}_{1|\mathbf{u}}(v),Y\le\widehat{q}_{2|\mathbf{u}}(w)\mid u\right\}-\Pr\left\{X\le q_{1|\mathbf{u}}(v),Y\le q_{2|\mathbf{u}}(w)\mid u\right\}\\&=\Pr\{X\le q_{1|\mathbf{u}}(v)\mid Y=q_{2|\mathbf{u}}(w),u\}f_{2|\mathbf{u}}\{q_{2|\mathbf{u}}(w)\}\{\widehat{q}_{2|\mathbf{u}}(w)-q_{2|\mathbf{u}}(w)\}\\&+\ \Pr\{Y\le q_{2|\mathbf{u}}(w)\mid X=q_{1|\mathbf{u}}(v),u\}f_{1|\mathbf{u}}\{q_{1|\mathbf{u}}(v)\}\{\widehat{q}_{1|\mathbf{u}}(v)-q_{1|\mathbf{u}}(v)\}+o_p(n^{-1/2}),\end{aligned}$$

where $f_{1|\mathbf{u}}$ and $f_{2|\mathbf{u}}$ stand for the density functions of X and Y conditional on $\mathbf{u}$. Define $\mathcal{F}_{q,1}\overset{\text{def}}{=}\{\widehat{q}_{1|\mathbf{u}}(v):|\widehat{q}_{1|\mathbf{u}}(v)-q_{1|\mathbf{u}}(v)|\le\delta_{n,1}\}$ and $\mathcal{F}_{q,2}\overset{\text{def}}{=}\{\widehat{q}_{2|\mathbf{u}}(w):|\widehat{q}_{2|\mathbf{u}}(w)-q_{2|\mathbf{u}}(w)|\le\delta_{n,2}\}$. Then we can verify that $var\left[I\left\{X\le\widehat{q}_{1|\mathbf{u}}(v)\right\}-I\left\{X\le q_{1|\mathbf{u}}(v)\right\}\right]\le\delta_{n,1}$ and $var\left[I\left\{Y\le\widehat{q}_{2|\mathbf{u}}(w)\right\}-I\left\{Y\le q_{2|\mathbf{u}}(w)\right\}\right]\le\delta_{n,2}$. When $nh^d\to\infty$, we combine this with Chebyshev's inequality and Theorem 37 in [20], then have

$$\begin{aligned}&\sup_{\mathcal{F}_{q,1}}\left|n^{-1}\sum_{i=1}^{n}K_h(\mathbf{u}_i-u)\Big[I\left\{X_i\le\widehat{q}_{1|\mathbf{u}}(v)\right\}-I\left\{X_i\le q_{1|\mathbf{u}}(v)\right\}\Big]-EK_h(\mathbf{u}_i-u)\Big[I\left\{X\le\widehat{q}_{1|\mathbf{u}}(v)\right\}-I\left\{X\le q_{1|\mathbf{u}}(v)\right\}\Big]\right|\\&\quad=o_p(n^{-1/2}).\end{aligned}$$

With the fact $\left|\widehat{F}_1\left\{\widehat{q}_{1|\mathbf{x}}(v)\right\}-v\right|\le\{nh^d\widehat{f}(u)\}^{-1}\sup_{u\in\mathbb{U}}|K(u)|=o_p(n^{-1/2})$, then

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n}K_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le q_{1|\mathbf{u}}(v)\right\}-EK_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le q_{1|\mathbf{u}}(v)\right\}\\&=n^{-1}\sum_{i=1}^{n}K_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le\widehat{q}_{1|\mathbf{u}}(v)\right\}-EK_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le\widehat{q}_{1|\mathbf{u}}(v)\right\}+o_p(n^{-1/2})\\&=n^{-1}\sum_{i=1}^{n}K_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le\widehat{q}_{1|\mathbf{u}}(v)\right\}-EK_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le q_{1|\mathbf{u}}(v)\right\}+EK_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le q_{1|\mathbf{u}}(v)\right\}\\&-EK_h\left(\mathbf{u}_i-u\right)I\left\{X_i\le\widehat{q}_{1|\mathbf{u}}(v)\right\}+o_p(n^{-1/2})=v\left\{\widehat{f}(u)-EK_h\left(\mathbf{u}_i-u\right)\right\}-\left\{\widehat{q}_{1|\mathbf{x}}(v)-q_{1|\mathbf{x}}(v)\right\}f\left\{q_{1|\mathbf{u}}(v),u\right\}+o_p(n^{-1/2}),\end{aligned}$$

where $\left\{\widehat{q}_{1|\mathbf{x}}(v)-q_{1|\mathbf{x}}(v)\right\}f\left\{q_{1|\mathbf{u}}(v),u\right\}$ in the last equality is obtained by the law of iterated expectations and Taylor's expansion, and $f\left\{q_{1|\mathbf{u}}(v),u\right\}$ denotes the joint density function of $\mathbf{u}$ and X given $\mathbf{u}$.

With $nh^{2d} \to \infty$ and $nh^{2m} \to 0$, $F_{n,\widehat{V},\widehat{W}}(v,w) - F_{n,V,W}(v,w)$ is simplified as $I_1 + I_2 + o_p(n^{-1/2})$, where I_1 and I_2 are defined shortly.

$$I_1 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} Pr\{X \le q_{1|\mathbf{u}_i}(v) \mid Y = q_{2|\mathbf{u}_i}(w), \mathbf{u}_i\}\big[w - I\{F_2(Y_i \mid \mathbf{u}_i) \le w\}\big], \tag{16}$$

$$I_2 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} Pr\{Y \le q_{2|\mathbf{u}_i}(w) \mid X = q_{1|\mathbf{u}_i}(v), \mathbf{u}_i\}\big[v - I\{F_1(X_i \mid \mathbf{u}_i) \le v\}\big]. \tag{17}$$

Following similar arguments,

$$F_{n,\widehat{V}}(v) - F_{n,V}(v) = n^{-1}\sum_{i=1}^{n}\big[v - I\{F_1(X_i \mid \mathbf{u}_i) \le v\}\big] + o_p(n^{-1/2}), \tag{18}$$

$$F_{n,\widehat{W}}(w) - F_{n,W}(w) = n^{-1}\sum_{i=1}^{n}\big[w - I\{F_2(Y_i \mid \mathbf{u}_i) \le w\}\big] + o_p(n^{-1/2}). \tag{19}$$

Then $T_{n,1}(v,w)$ becomes $Q_n(v,w) + I_3 + I_4 + o_p(n^{-1/2})$, where

$$I_3 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\big[\Pr\{X \le q_{1|\mathbf{u}_i}(v) \mid Y = q_{2|\mathbf{u}_i}(w), \mathbf{u}_i\} - v\big]\big[w - I\{F_2(Y_i \mid \mathbf{u}_i) \le w\}\big]$$

$$I_4 \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\big[\Pr\{Y \le q_{2|\mathbf{u}_i}(w) \mid X = q_{1|\mathbf{u}_i}(v), \mathbf{u}_i\} - w\big]\big[v - I\{F_1(X_i \mid \mathbf{u}_i) \le v\}\big]$$

Clearly, $\Pr\{X \le q_{1|\mathbf{u}}(v) \mid Y = q_{2|\mathbf{u}}(w), \mathbf{u}\} - v = 0$ and $\Pr\{Y \le q_{2|\mathbf{u}}(w) \mid X = q_{1|\mathbf{u}}(v), \mathbf{u}\} - w = 0$ hold if $X \perp\!\!\!\perp Y \mid \mathbf{u}$. Thus, $T_{n,1}(v,w)$ is degenerated to be $Q_n(v,w) + o_p(n^{-1/2})$ under null hypothesis.

$$\begin{aligned} T_{n,2}(v,w) \overset{\text{def}}{=} & \big\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\big\}\big\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\big\} + \big\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\big\}\big\{F_{n,W}(w) - w\big\} \\ & + \big\{F_{n,V}(v) - v\big\}\big\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\big\} + \big\{F_{n,V}(v) - v\big\}\big\{F_{n,W}(w) - w\big\} \end{aligned}$$

Invoking the fact $F_{n,T}(t) - t = O_p(n^{-1/2})$, Eqs. (18) and (19), we obtain that $T_{n,2}(v,w) = o_p(n^{-1/2})$. Based on above results, $\widehat{\rho}^{CI}$ can be simplified as

$$\widehat{\rho}^{CI} = \int_{\mathbb{R}^1}\int_{\mathbb{R}^1} Q_n^2(v,w)\, dF_{n,\widehat{V}}(v) dF_{n,\widehat{W}}(w) + o_p(n^{-1}).$$

We remark here that $Q_n = O_p(n^{-1/2})$ under the independence of V and W. Define

$$\widetilde{\rho}^{CI} \overset{\text{def}}{=} \int_{\mathbb{R}^1}\int_{\mathbb{R}^1} Q_n^2(v,w)\, dF_V(v) dF_W(w).$$

It is easy to verify that $\widehat{\rho}^{CI} - \widetilde{\rho}^{CI} = o_p(n^{-1})$. Thus, it is sufficient to prove under H_0,

$$n\widetilde{\rho}^{CI} \xrightarrow{d} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{\chi_{ij}^2(1)}{\pi^4 i^2 j^2}, \text{ as } n \to \infty.$$

For $0 \le v, w \le 1$, we define the empirical process $L_n(v,w) = n^{1/2}Q_n(v,w)$. When V and W are independent, $E\{L_n(v,w)\} = 0$. The covariance of the process is

$$E\{L_n(v,w)L_n(v',w')\} = r_n(v,w;v',w') = \{\min(v,v') - vv'\}\{\min(w,w') - ww'\}$$

Clearly, the covariance function $r_n(v,w;v',w')$ is independent of n and symmetric in (v,w) and v', w'. Regard covariance function $r_n(v,w;v',w') = r(v,w;v',w')$ as the kernel in the following eigenvalue problem

$$\int_{\mathbb{R}^1}\int_{\mathbb{R}^1} r(v,w;v',w')\phi(v',w')dv'dw' = \zeta\phi(v,w),$$

where the integral is over all components of (v',w'). The kernel is positive definite. Denote the eigenvalue by ζ_1, ζ_2, ... and the corresponding orthonormal eigenfunction by $\phi_1(v,w)$, $\phi_2(v,w)$, ... Then we obtain that $r(v,w;v',w') = \sum_{j=1}^{\infty}\zeta_j\phi_j(v,w)\phi_j(v',w')$ with uniform convergence according to Mercer's theorem.

It follows from analogue arguments of [21] that as $n \to \infty$, the limiting distribution of $n\,\widetilde{\rho}^{CI}$ is

$$\int_{\mathbb{R}^1}\int_{\mathbb{R}^1} L_n(v,w)dvdw = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\zeta_{ij}Z_{ij}^2,$$

where Z_{ij}s are independent normal random variables with mean zero and variance 1. Under the independence of V and W, [3] proved that $\zeta_{ij} = 1/\pi^4 i^2 j^2$.
Step 2. Under the alternative hypothesis H_1 in (4) and (5), we show that

$$n^{1/2}\left(\widehat{\rho}^{CI} - \rho^{CI}\right) \xrightarrow{d} \mathcal{N}\{0, 4var(Z)\}, \text{ as } n \to \infty.$$

Let $\Delta(v, w) \overset{\text{def}}{=} F_{V,W}(v, w) - vw$. If V and W are not independent, $Q_n(v, w) - \Delta(v, w) = O_p(n^{-1/2})$. Invoking conclusions in Step 1, we have

$$T_n^2(v, w) = 2\Delta(v, w)\{Q_n(v, w) - \Delta(v, w) + I_3 + I_4\} + \Delta^2(v, w) + o_p(n^{-1/2}).$$

Then we represent $\widehat{\rho}^{CI} - \rho^{CI}$ as an average of averages of independent and identically distributed random variables, that is,

$$\widehat{\rho}^{CI} - \rho^{CI} = 2n^{-1}\sum_{i=1}^{n} Z_i + o_p(n^{-1/2}),$$

where Z_i is defined as

$$\begin{aligned} Z_i \overset{\text{def}}{=} & \int_{\mathbb{R}^1}\int_{\mathbb{R}^1} \left\{F_{V,W}(v, w) - vw\right\} \Big[I(V_i \le v, W_i \le v) - F_{V,W}(v, w) - vI(W_i \le w) \\ & - wI(V_i \le v) + 2vw + \left[\Pr\{X \le q_{1|\mathbf{u}_i}(v) \mid Y = q_{2|\mathbf{u}_i}(w), \mathbf{u}_i\} - v\right]\left\{w - I(W_i \le w)\right\} \\ & + \left[\Pr\{Y \le q_{2|\mathbf{u}_i}(w) \mid X = q_{1|\mathbf{u}_i}(v), \mathbf{u}_i\} - w\right]\left\{v - I(V_i \le v)\right\} \Big] \, dvdw. \end{aligned} \tag{20}$$

All Z_is are independent and identically distributed. The second part of Theorem 1 can be completed with the classical central limit theorem. □

Appendix C. Proofs related to screening

To enhance the readability, we divide the proof into four steps.
Step 1. We first prove that, under conditions (A1)–(A4), for any $0 < 2\gamma + \theta d/2 \le 1$, there exists some positive constant c such that

$$\Pr\left\{\left|\widehat{\rho}^{CI} - \rho^{CI}\right| \ge cn^{-\gamma}\right\} \le O\left\{n^4 \exp(-cn^{1-2\gamma-\theta d/2})\right\}.$$

Recall the definitions of $T_n(v, w)$ and $\Delta(v, w)$ in the proof of Theorem 1, we have

$$\begin{aligned} \Pr\left\{\left|\widehat{\rho}^{CI} - \rho^{CI}\right| \ge cn^{-\gamma}\right\} & \le \Pr\left[\sup_{\mathcal{F}_1,\mathcal{F}_2} \left| \left\{F_{n,\widehat{V},\widehat{W}}(v, w) - F_{n,\widehat{V}}(v)F_{n,\widehat{W}}(w)\right\}^2 - \left\{F_{V,W}(v, w) - vw\right\}^2 \right| \ge cn^{-\gamma}\right] \\ & \le \Pr\left[\sup_{\mathcal{F}_1,\mathcal{F}_2} \left|T_n(v, w) - \Delta(v, w)\right| \ge cn^{-\gamma}\right] \le \Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} \left|T_{n,1}(v, w) - \Delta(v, w)\right| \ge cn^{-\gamma}/2\right\} \\ & + \Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} \left|T_{n,2}(v, w)\right| \ge cn^{-\gamma}/2\right\}, \end{aligned}$$

where the second inequality holds because the covariance between two binary random variables is smaller than 1/4.

With the Bonferroni inequality, we have that

$$\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} \left|T_{n,1}(v, w) - \Delta(v, w)\right| \ge cn^{-\gamma}/2\right\} \le \sum_{k=1}^{5} \Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,k}(v, w)| \ge cn^{-\gamma}/10\right\},$$

where $G_{1,k}(v, w)$s will be defined shortly.

$$G_{1,1}(v, w) \overset{\text{def}}{=} F_{n,\widehat{V},\widehat{W}}(v, w) - F_{n,V,W}(v, w) - E\left\{F_{n,\widehat{V},\widehat{W}}(v, w) - F_{n,V,W}(v, w)\right\}.$$

Since $0 < 2\gamma + \theta d/2 \le 1$, Lemma 3 (14) ensures that

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,1}(v, w)| \ge cn^{-\gamma}/10\right) \le O\left\{n^{8\gamma} \exp(-cn^{1-2\gamma-\theta d/2})\right\}.$$

Then we define the terms $G_{1,2}(v, w)$ and $G_{1,3}(v, w)$ as follows,

$$\begin{aligned} G_{1,2}(v, w) & \overset{\text{def}}{=} w\left[F_{n,\widehat{V}}(v) - F_{n,V}(v) - E\left\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\right\}\right], \\ G_{1,3}(v, w) & \overset{\text{def}}{=} v\left[F_{n,\widehat{W}}(w) - F_{n,W}(w) - E\left\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\right\}\right]. \end{aligned}$$

Lemma 3 (15) yields that

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,2}(v,w)| \geq cn^{-\gamma}/10\right) \leq O\left\{n^{4\gamma}\exp(-cn^{1-2\gamma-\theta d/2})\right\},$$

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,3}(v,w)| \geq cn^{-\gamma}/10\right) \leq O\left\{n^{4\gamma}\exp(-cn^{1-2\gamma-\theta d/2})\right\}.$$

Define $G_{1,4}(v,w) \overset{\text{def}}{=} Q_n(v,w) - \Delta(v,w)$. Apparently, $|I(V_i \leq v, W_i \leq w) - F_{V,W}(v,w)| \leq 1$ for any v and w. We apply Hoeffding's inequality and empirical process theory [20, page 15] to obtain that, for any $t \in (0,1)$,

$$\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} |F_{n,V,W}(v,w) - F_{V,W}(v,w)| > t\right\} \leq 2(n+1)\exp(-2nt^2).$$

This inequality and Lemma 2 allow us to show that

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,4}(v,w)| \geq cn^{-\gamma}/10\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\}.$$

The last term $G_{1,5}$ is

$$G_{1,5}(v,w) \overset{\text{def}}{=} E\left\{F_{n,\widehat{V},\widehat{W}}(v,w) - F_{n,V,W}(v,w)\right\} - vE\left\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\right\} - wE\left\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\right\}.$$

With (16)–(19) in the proof of Theorem 1 and Hoeffding's inequality in Lemma 2, as $nh^{2d} \to \infty$ and $nh^{2m} \to 0$,

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{1,5}(v,w)| \geq cn^{-\gamma}/10\right) \leq \Pr\left(\sup_{\mathcal{F}_1}\left|n^{-1}\sum_{i=1}^{n} I(V_i \leq v) - v\right| \geq cn^{-\gamma}/20\right)$$
$$+ \Pr\left(\sup_{\mathcal{F}_2}\left|n^{-1}\sum_{i=1}^{n} I(W_i \leq w) - w\right| \geq cn^{-\gamma}/20\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\}.$$

Based on above results, we obtain that, for sufficiently large n,

$$\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} |T_{n,1}(v,w) - \Delta(v,w)| \geq cn^{-\gamma}/2\right\} \leq O\left\{n^{8\gamma}\exp(-cn^{1-2\gamma-\theta d/2})\right\}. \tag{21}$$

Invoking Bonferroni inequality to deal with the term $T_{2,n}(v,w)$, we have that

$$\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} |T_{2,n}(v,w)| \geq cn^{-\gamma}/2\right\} \leq \sum_{k=1}^{4}\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{2,k}(v,w)| \geq cn^{-\gamma}/8\right\},$$

where the definitions of $G_{2,k}(v,w)$ are as follows.

$$G_{2,1} \overset{\text{def}}{=} \left\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\right\}\left\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\right\}.$$

Similar to dealing with $G_{1,5}(v,w)$, then

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{2,1}(v,w)| \geq cn^{-\gamma}/8\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\}.$$

Next, we let $G_{2,2}(v,w) \overset{\text{def}}{=} \left\{F_{n,V}(v) - v\right\}\left\{F_{n,W}(w) - w\right\}$. Note that $|I(V_i \leq v) - v| \leq 1$ for any v and $|I(W_j \leq w) - w| \leq 1$ for any w. Applying Lemma 2, we have

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{2,2}(v,w)| \geq cn^{-\gamma}/8\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\}.$$

We turn to the last two terms in what follows.

$$G_{2,3}(v,w) \overset{\text{def}}{=} \left\{F_{n,W}(w) - w\right\}\left\{F_{n,\widehat{V}}(v) - F_{n,V}(v)\right\},$$

$$G_{2,4}(v,w) \overset{\text{def}}{=} \left\{F_{n,V}(v) - v\right\}\left\{F_{n,\widehat{W}}(w) - F_{n,W}(w)\right\}.$$

Using analogue arguments applied to $G_{2,3}$ and $G_{2,4}$, we obtain that

$$\Pr\left(\sup_{\mathcal{F}_1,\in\mathcal{F}_2} |G_{2,3}(v,w)| \geq cn^{-\gamma}/8\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\},$$

$$\Pr\left(\sup_{\mathcal{F}_1,\mathcal{F}_2} |G_{2,4}(v,w)| \geq cn^{-\gamma}/8\right) \leq O\left\{n\exp(-cn^{1-2\gamma})\right\}.$$

Thus, we have

$$\Pr\left\{\sup_{\mathcal{F}_1,\mathcal{F}_2}|T_{2,n}(v,w)\geq cn^{-\gamma}/2\right\}\leq O\left\{n\exp(-cn^{1-2\gamma})\right\},$$

for sufficiently large n. This, together with (21), yields that

$$\Pr\left(\max_{1\leq k\leq p}|\widehat{\rho}_k^{CI}-\rho_k^{CI}|\geq cn^{-\gamma}\right)\leq p\max_{1\leq k\leq p}\Pr(|\widehat{\rho}_k^{CI}-\rho_k^{CI}|\geq cn^{-\gamma})\leq p\,n^4\exp(-cn^{1-2\gamma-\theta d/2}),$$

where c is a positive constant.
Step 2. Assume the condition $\min_{k\in\mathcal{A}}\ \rho_k^{CI}\geq 2cn^{-\gamma}$. We prove that

$$\Pr(\mathcal{A}\subseteq\widehat{\mathcal{A}})\geq 1-O\left\{n^4|\mathcal{A}|\exp(-cn^{1-2\gamma-\theta d/2})\right\}.$$

If $\mathcal{A}\nsubseteq\widehat{\mathcal{A}}$, there must exist some $j\in\mathcal{A}$ such that $\widehat{\rho}_j^{CI}<cn^{-\gamma}$. Under the condition $\min_{k\in\mathcal{A}}\ \rho_k^{CI}\geq 2cn^{-\gamma}$, we have $|\widehat{\rho}_j^{CI}-\rho_j^{CI}|\geq cn^{-\gamma}$ for this particular j, which implies

$$\{\mathcal{A}\nsubseteq\widehat{\mathcal{A}}\}\subseteq\{|\widehat{\rho}_j^{CI}-\rho_j^{CI}|\geq cn^{-\gamma},\ \text{for some } j\in\mathcal{A}\}.$$

Then, it is clear that

$$\begin{aligned}\Pr(\mathcal{A}\subseteq\widehat{\mathcal{A}})&\geq 1-\Pr\left\{\left|\widehat{\rho}_j^{CI}-\rho_j^{CI}\right|\geq cn^{-\gamma},\ \text{for some } j\in\mathcal{A}\right\}\geq 1-|\mathcal{A}|\max_{j\in\mathcal{A}}\Pr\{|\widehat{\rho}_j^{CI}-\rho_j^{CI}|\geq cn^{-\gamma}\}\\&\geq 1-O\left\{n^4|\mathcal{A}|\exp(-cn^{1-2\gamma-\theta d/2})\right\}.\end{aligned}$$

Step 3. Assume the condition $\min_{k\in\mathcal{A}}\ \rho_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\geq 0$. We show that

$$\liminf_{n\longrightarrow\infty}\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}\right\}>0 \text{ in probability.}$$

Under the condition $\min_{k\in\mathcal{A}}\ \rho_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\geq 0$, there exists some $\delta>0$ such that $\min_{k\in\mathcal{A}}\ \rho_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}=\delta$. Then we have

$$\begin{aligned}&\Pr\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}\leq\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}\right\}\leq\Pr\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\min_{k\in\mathcal{A}}\rho_k^{CI}+\delta\leq\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\rho_k^{CI}\right\}\\&\leq\Pr\left\{\left|(\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI})-(\min_{k\in\mathcal{A}}\rho_k^{CI}-\max_{k\in\mathcal{A}^c}\rho_k^{CI})\right|\geq\delta\right\}\leq\Pr\left\{2\max_{1\leq k\leq p}\left|\widehat{\rho}_k^{CI}-\rho_k^{CI}\right|\geq\delta\right\}\\&\leq O\left\{p\,n^4\exp\left(-cn^{1-\theta d/2}\right)\right\}.\end{aligned}$$

By Fatou's Lemma,

$$\Pr\left\{\liminf_{n\longrightarrow\infty}\left(\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}\right)\leq 0\right\}\leq\lim_{n\to\infty}\Pr\left(\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}\leq 0\right)=0.$$

Thus $\Pr\left\{\liminf_{n\longrightarrow\infty}\left(\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}\right)>0\right\}=1$.
Step 4. Assume the condition $\min_{k\in\mathcal{A}}\ \rho_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\geq 2cn^{-\gamma}$. We prove that, in probability that,

$$\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}>\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2,\quad\max_{k\in\mathcal{A}^c}\widehat{\rho}_k^{CI}<\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2.$$

We first write

$$\begin{aligned}&\Pr\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}\leq\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2\right\}=\Pr\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}-\min_{k\in\mathcal{A}}\ \rho_k^{CI}\leq\left(\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}-\min_{k\in\mathcal{A}}\ \rho_k^{CI}\right)/2\right\}\\&\leq\Pr\left\{\max_{1\leq k\leq p}\left|\widehat{\rho}_k^{CI}-\rho_k^{CI}\right|\geq cn^{-\gamma}\right\}\leq O\{p\,n^4\exp(-cn^{1-2\gamma-\theta d/2})\}.\end{aligned}$$

Thus, $\lim_{n\to\infty}\Pr\left\{\min_{k\in\mathcal{A}}\widehat{\rho}_k^{CI}>\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2\right\}=1$. Similarly, it can be proven that

$$\begin{aligned}&\Pr\left\{\max_{k\in\mathcal{A}^c}\ \widehat{\rho}_k^{CI}\geq\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2\right\}=\Pr\left\{\max_{k\in\mathcal{A}^c}\ \widehat{\rho}_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\geq\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}-\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2\right\}\\&\leq\Pr\left\{\max_{1\leq k\leq p}\left|\widehat{\rho}_k^{CI}-\rho_k^{CI}\right|\geq cn^{-\gamma}\right\}\leq O\{p\,n^4\exp(-cn^{1-2\gamma-\theta d/2})\}.\end{aligned}$$

Therefore, $\lim_{n\to\infty}\Pr\left\{\max_{k\in\mathcal{A}^c}\ \widehat{\rho}_k^{CI}<\left(\min_{k\in\mathcal{A}}\ \rho_k^{CI}+\max_{k\in\mathcal{A}^c}\ \rho_k^{CI}\right)/2\right\}=1$.

References

[1] E. Barut, J. Fan, A. Verhasselt, Conditional sure independence screening, J. Amer. Statist. Assoc. 111 (515) (2016) 1266–1277.
[2] W. Bergsma, Nonparametric testing of conditional independence by means of the partial copula, 2011, Available at SSRN 1702981.
[3] J.R. Blum, J. Kiefer, M. Rosenblatt, Distribution free tests of independence based on the sample distribution function, Ann. Math. Stat. 32 (2) (1961) 485–498.
[4] A.P. Chiang, J.S. Beck, H.-J. Yen, M.K. Tayeh, T.E. Scheetz, R.E. Swiderski, D.Y. Nishimura, T.A. Braun, K.-Y.A. Kim, J. Huang, K. Elbedour, R. Carmi, D.C. Slusarski, T.L. Casavant, E.M. Stone, V.C. Sheffield, Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (bbs11), Proc. Natl. Acad. Sci. 103 (16) (2006) 6287–6292.
[5] A. Dawid, Conditional independence in statistical theory, J. R. Stat. Soc. Ser. B Stat. Methodol. (1979) 1–31.
[6] M.A. Delgado, W.G. Manteiga, Significance testing in nonparametric regression based on the bootstrap, Ann. Statist. 29 (5) (2001) 1469–1507.
[7] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5) (2008) 849–911.
[8] J. Fan, Y. Ma, W. Dai, Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models, J. Amer. Statist. Assoc. 109 (507) (2014) 1270–1284.
[9] N.I. Fisher, P.K. Sen, Probability Inequalities for Sums of Bounded Random Variables, 58, (301) Publications of the American Statistical Association, 1963, pp. 13–30.
[10] T.L. Horn, K.E. Torres, J.M. Naylor, M.J. Cwik, C.J. Detrisac, I.M. Kapetanovic, R.A. Lubet, J.A. Crowell, D.L. McCormick, Subchronic toxicity and toxicogenomic evaluation of tamoxifen citrate+ bexarotene in female rats, Toxicol. Sci. 99 (2) (2007) 612–627.
[11] T.M. Huang, Testing conditional independence using maximal nonlinear conditional correlation, Ann. Statist. 38 (4) (2010) 2047–2091.
[12] J. Huang, S. Ma, C.H. Zhang, Adaptive lasso for sparse high-dimensional regression models, Statist. Sinica 18 (4) (2008) 1603–1618.
[13] D. Koller, N. Friedman, ProbabiliStic Graphical Models: Principles and Techniques, MIT press, 2009.
[14] A. Lawrance, On conditional and partial correlation, Amer. Statist. 30 (3) (1976) 146–149.
[15] R. Li, W. Zhong, L.P. Zhu, Feature screening via distance correlation learning, J. Amer. Statist. Assoc. 107 (499) (2012) 1129–1139.
[16] O. Linton, P. Gozalo, Conditional independence restrictions: Testing and estimation, 1140, COWLES FOUNDATION, 1997, Cowles Foundation Discussion Paper.
[17] J. Liu, R. Li, R. Wu, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, J. Amer. Statist. Assoc. 109 (505) (2014) 266–274.
[18] S. Ma, R. Li, C.-L. Tsai, Variable screening via quantile partial correlation, J. Amer. Statist. Assoc. 112 (518) (2017) 650–663.
[19] J. Pearl, CauSality: Models, Reasoning and Inference, Vol.29, Springer, 2000.
[20] D. Pollard, Convergence of Stochastic Processes, Springer Science & Business Media, 2012.
[21] M. Rosenblatt, Limit theorems associated with variants of the von mises statistic, Ann. Math. Stat. 23 (4) (1952) 617–623.
[22] M. Rosenblatt, Remarks on a multivariate transformation, Ann. Math. Stat. 23 (3) (1952) 470–472.
[23] T.E. Scheetz, K.Y.A. Kim, R.E. Swiderski, A.R. Philp, T.A. Braun, K.L. Knudtson, A.M. Dorrance, G.F. DiBona, J. Huang, T.L. Casavant, et al., Regulation of gene expression in the mammalian eye and its relevance to eye disease, Proc. Natl. Acad. Sci. 103 (39) (2006) 14429–14434.
[24] N. Simon, R. Tibshirani, A permutation approach to testing interactions for binary response by comparing correlations between classes, J. Amer. Statist. Assoc. 110 (512) (2015) 1707–1716.
[25] L. Su, H. White, A consistent characteristic function-based test for conditional independence, J. Econometrics 141 (2) (2007) 807–834.
[26] L. Su, H. White, A nonparametric hellinger metric test for conditional independence, Econ. Theory 24 (4) (2008) 829–864.
[27] L. Su, H. White, Testing conditional independence via empirical likelihood, J. Econometrics 182 (1) (2014) 27–44.
[28] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. 35 (6) (2007) 2769–2794.
[29] X. Wang, Y. Hong, Characteristic function based testing for conditional independence: a nonparametric regression approach, Econ. Theory 34 (4) (2018) 815–849.
[30] X. Wang, W. Pan, W. Hu, Y. Tian, H. Zhang, Conditional distance correlation, J. Amer. Statist. Assoc. 110 (512) (2015) 1726–1734.
[31] H. Wang, Y. Xia, Shrinkage estimation of the varying coefficient model, J. Amer. Statist. Assoc. 104 (486) (2009) 747–757.
[32] K. Zhang, J. Peters, D. Janzing, B. Schölkopf, Kernel-based conditional independence test and application in causal discovery, Comput. Sci. 6 (8) (2012) 895–907.
[33] Y. Zhou, L.P. Zhu, Model-free feature screening for ultrahigh dimensional datathrough a modified Blum–Kiefer–Rosenblatt correlation, Statist. Sinica 28 (2018) 1351–1370.
[34] L.X. Zhu, Convergence rates of the empirical processes indexed by the classes of functions with applications, J. Syst. Sci. Math. Sci. 13 (1) (1993) 33–41.
[35] L.P. Zhu, L. Li, R. Li, L.X. Zhu, Model-free feature screening for ultrahigh-dimensional data, J. Amer. Statist. Assoc. 106 (496) (2011) 1464–1475.
[36] L.P. Zhu, K. Xu, R. Li, W. Zhong, Projection correlation between two random vectors, Biometrika 104 (4) (2017) 829–843.