Journal of Machine Learning Research 23 (2022) 1-41 Submitted 7/20; Revised 8/21; Published 2/22

A Distribution Free Conditional Independence Test with Applications to Causal Discovery

Zhanrui Cai ZHANRUIC@ANDREW.CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Runze Li RZLI@PSU.EDU
Department of Statistics
The Pennsylvania State University
University Park, PA 16802, USA

Yaowu Zhang ZHANG.YAOWU@MAIL.SHUFE.EDU.CN
Research Institute for Interdisciplinary Sciences
School of Information Management and Engineering
Shanghai University of Finance and Economics
Shanghai, 200433, China

Editor: Garvesh Raskutti

Abstract

This paper is concerned with test of the conditional independence. We first establish an equivalence between the conditional independence and the mutual independence. Based on the equivalence, we propose an index to measure the conditional dependence by quantifying the mutual dependence among the transformed variables. The proposed index has several appealing properties. (a) It is distribution free since the limiting null distribution of the proposed index does not depend on the population distributions of the data. Hence the critical values can be tabulated by simulations. (b) The proposed index ranges from zero to one, and equals zero if and only if the conditional independence holds. Thus, it has nontrivial power under the alternative hypothesis. (c) It is robust to outliers and heavy-tailed data since it is invariant to conditional strictly monotone transformations. (d) It has low computational cost since it incorporates a simple closed-form expression and can be implemented in quadratic time. (e) It is insensitive to tuning parameters involved in the calculation of the proposed index. (f) The new index is applicable for multivariate random vectors as well as for discrete data. All these properties enable us to use the new index as statistical inference tools for various data. The effectiveness of the method is illustrated through extensive simulations and a real application on causal discovery.

Keywords: Conditional independence, mutual independence, distribution free.

1. Introduction

Conditional independence is fundamental in graphical models and causal inference (Jordan, 1998). Under multinormality assumption, conditional independence is equivalent to the corresponding partial correlation being 0. Thus, partial correlation may be used to measure conditional dependence (Lawrance, 1976). However, partial correlation has low power in detecting conditional dependence in the presence of nonlinear dependence. In addition, it cannot control Type I error when the multinormality assumption is violated. In general, testing for conditional independence is much more challenging than for unconditional independence (Zhang et al., 2011; Shah and Peters, 2020).

Recent works on test of conditional independence have focused on developing omnibus conditional independence test without assuming specific functional forms of the dependencies. Linton and Gozalo (1996) proposed a nonparametric conditional independence test based on the generalization of empirical distribution function, and proposed using bootstrap to obtain the null distribution of the proposed test. This diminishes the computational efficiency. Other approaches include measuring the difference between conditional characteristic functions (Su and White, 2007), the weighted Hellinger distance (Su and White, 2008), and the empirical likelihood (Su and White, 2014). Although these authors established the asymptotical normality of the proposed test under conditional independence, the performance of their proposed tests relies heavily on consistent estimate of the bias and variance terms, which are quite complicated in practice. The asymptotical null distribution may perform badly with a small sample. Thus, the authors recommended obtaining critical values of the proposed tests by a bootstrap. This results in heavy computation burdens. Huang (2010) proposed a test of conditional independence based on the maximum nonlinear conditional correlation. By discretizing the conditioning set into a set of bins, the author transforms the original problem into an unconditional testing problem. Zhang et al. (2011) proposed a kernel-based conditional independence test, which essentially tests for zero Hilbert-Schmidt norm of the partial cross-covariance operator in the reproducing kernel Hilbert spaces. The test also required a bootstrap to approximate the null distribution. Wang et al. (2015) introduced the energy statistics into the conditional test and developed the conditional distance correlation based on Székely et al. (2007), which can also be linked to kernel-based approaches. But the test statistics requires to compute high order U–statistics and therefore suffers heavy computation burden, which is of order $O(n^3)$ for a sample with size n. Runge (2018) proposed a non-parametric conditional independence testing based on the information theory framework, in which the conditional mutual information was estimated directly via combining the k-nearest neighbor estimator with a nearest-neighbor local permutation scheme. However, the theoretical distribution of the proposed test is unclear.

In this paper, we develop a new methodology to test conditional independence and propose conditional independence tests that are applicable for continuous or discrete random variables or vectors. Let X, Y and Z be three continuous random variables. We are interested in testing whether X and Y are statistically independent given Z:

$$H_0 : X \perp\!\!\!\perp Y \mid Z, \quad \text{versus} \quad H_1 : \text{ otherwise}.$$

Here we focus on random variables for simplicity. We will consider test of conditional independence for random vectors in Section 3. To begin with, we observe that with Rosenblatt transformation (Rosenblatt, 1952), i.e., $U \stackrel{\text{def}}{=} F_{X|Z}(X \mid Z)$, $V \stackrel{\text{def}}{=} F_{Y|Z}(Y \mid Z)$ and $W \stackrel{\text{def}}{=} F_Z(Z)$, $X \perp\!\!\!\perp Y \mid Z$ is equivalent to the mutual independence of U, V and W. Thus we convert a conditional independence test into a mutual independence test, and any technique for testing mutual independence can be readily applied. For example, Chakraborty and Zhang (2019) proposed the joint distance covariance to test mutual independence and

Drton et al. (2020) constructed a family of tests with maxima of rank correlations in high dimensions. However, these mutual independence tests do not consider the intrinsic properties of U, V and W. That is, U, V and W are all uniformly distributed, $U \perp\!\!\!\perp W$ and $V \perp\!\!\!\perp W$. This motivates us to develop a new index ρ to measure the mutual dependence. We show that the index ρ has a closed form, which is much simpler than that of Chakraborty and Zhang (2019). In addition, it is symmetric, invariant to strictly monotone transformations, and ranges from zero to one, and is equal to zero if and only if U, V and W are mutually independent. Based on the index ρ, we further proposed tests of conditional independence. We would like to further note a recent work proposed by Zhou et al. (2020), who suggested to simply test whether U and V are independent. However, this is not fully equivalent to the conditional independence test and it is unclear what kind of power loss one might have.

The proposed tests have several appealing features. (a) The proposed test is distribution free in the sense that its limiting null distribution does not depend on unknown parameters and the population distributions of the data. The fact that both U and V are independent of W makes the test statistic n-consistent under the null hypothesis without requiring under-smoothing. In addition, even though the test statistic depends on U, V and W, which needs to be estimated nonparametrically, we show that the test statistic has the same asymptotic properties as the statistics where true U, V and W are directly available. This leads to a distribution free test statistic when further considering that U, V and W are uniformly distributed. Although some tests in the literature are also distribution free, the asymptotic distributions are either complicated to estimate (e.g., Su and White, 2007) or rely on the Gaussian process, which is not known how to simulate (e.g., Song, 2009) and would require a wild bootstrap method to determine the critical values. Compared with existing ones, the limiting null distribution of the proposed test depends on U, V and W only, and the critical values can be easily obtained by a simulation-based procedure. (b) The proposed test has nontrivial power against all fixed alternatives. The population version of the test statistic ranges from zero to one and equals zero if and only if conditional independence holds. Unlike many testing procedures that are weaker than that for conditional independence (e.g., Song, 2009), the equivalence between conditional independence and mutual independence guarantees that the newly proposed test has nontrivial power against all fixed alternatives. (c) The proposed test is robust since it is invariant to strictly monotone transformations and thus, it is robust to outliers. Furthermore, U, V and W all have bounded support, and therefore it is suitable for handling heavy-tailed data. (d) The proposed test has low computational cost. It is a V–statistic, and direct calculation requires only $O(n^2)$ computational complexity. (e) It is insensitive to tuning parameters involved in the test statistics. The test statistics are n–consistent under the null hypothesis without under-smoothing, and is hence much less sensitive to the bandwidth. The proposed index ρ is extended to continuous random vectors and discrete data in Section 3. All these properties enable us to use the new conditional independence test for various data.

The rest of this paper is organized as follows. In Section 2 we first show the equivalence between conditional independence and mutual independence. We propose a new index to measure the mutual dependence, and derive desirable properties of the proposed index in Section 2.1. We propose an estimator for the new index in Section 2.2. The asymptotic distributions of the proposed estimator under the null hypothesis, global alternative, and local alternative hypothesis are derived in Section 2.2. We extend the new index to the multivariate and discrete cases in Section 3. We conduct numerical comparisons and apply the proposed test to causal discovery in directed acyclic graphs in Section 4. Some final remarks are given in Section 5. We provide some additional simulation results as well as all the technical proofs in the appendix.

2. Methodology

To begin with, we establish an equivalence between conditional independence and mutual independence. In this section, we focus on the setting in which X, Y and Z are continuous univariate random variables, and the problem of interest is to test $X \perp\!\!\!\perp Y \mid Z$. Throughout this section, denote $U = F_{X|Z}(X \mid Z)$, $V = F_{Y|Z}(Y \mid Z)$ and $W = F_Z(Z)$. The proposed methodology is built upon the following proposition.

Proposition 1 *Suppose that X and Y are both univariate and have continuous conditional distribution functions for every given value of Z, and Z is a continuous univariate random variable. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if U, V and W are mutually independent.*

We provide a detailed proof of Proposition 1 in the appendix. Essentially, it establishes an equivalence between the conditional independence of $X \perp\!\!\!\perp Y \mid Z$ and the mutual independence among U, V and W under the conditions in Proposition 1. Therefore, we can alleviate the hardness issue of conditional independence testing (Shah and Peters, 2020) by restricting the distribution family of the data such that U, V and W can be estimated sufficiently well using samples. As shown in our theoretical analysis, we further impose certain smoothness conditions on the conditional distributions of $X, Y \mid Z = z$ as z varies in the support of Z. This distribution family is also considered in Neykov et al. (2021) to develop a minimax optimal conditional independence test. We discuss the extension of Proposition 1 to multivariate and discrete data in Section 3.

According to Proposition 1, any techniques for testing mutual independence among three random variables can be readily applied for conditional independence testing problems. For example, Chakraborty and Zhang (2019) proposed the joint distance covariance and Patra et al. (2016) developed a bootstrap procedure to test mutual independence with known marginals. However, a direct application of these metrics may not be a good choice because it ignores the fact that the variables U, V and W are all uniformly distributed, as well as $U \perp\!\!\!\perp W$ and $V \perp\!\!\!\perp W$. Next, we discuss how to develop a new mutual independence test while considering these intrinsic properties of (U, V, W).

2.1 A mutual independence test

In this section, we propose to characterize the conditional dependence of X and Y given Z through quantifying the mutual dependence among U, V and W. Although our proposed test is based on the distance between characteristics functions, our proposed test is much simpler and has different asymptotic distribution as well as different convergence rate from the conditional distance correlation proposed by Wang et al. (2015). Let $\omega(\cdot)$ be an arbitrary positive weight function and $\varphi_{U,V,W}(\cdot)$, $\varphi_U(\cdot)$, $\varphi_V(\cdot)$, and $\varphi_W(\cdot)$ be the characteristic functions of (U, V, W), U, V and W, respectively. Then

$$\begin{aligned} & U, V \text{ and } W \text{ are mutually independent} \\ \iff & \varphi_{U,V,W}(t_1,t_2,t_3) = \varphi_U(t_1)\varphi_V(t_2)\varphi_W(t_3) \text{ for all } t_1,t_2,t_3 \in \mathbb{R} \\ \iff & \iiint \left\|\varphi_{U,V,W}(t_1,t_2,t_3) - \varphi_U(t_1)\varphi_V(t_2)\varphi_W(t_3)\right\|^2 \omega(t_1,t_2,t_3)dt_1dt_2dt_3 = 0, \end{aligned}$$

where $\|\psi\|^2 = \psi^{\mathrm{T}}\overline{\psi}$ for a complex-valued function ψ and $\overline{\psi}$ is the conjugate of ψ. By choosing $\omega(t_1, t_2, t_3)$ to be the joint probability density function of three independent and identically distributed standard Cauchy random variables, the integration in the above equation has a closed form,

$$Ee^{-|U_1-U_2|-|V_1-V_2|-|W_1-W_2|} - 2Ee^{-|U_1-U_3|-|V_1-V_4|-|W_1-W_2|} \\ + Ee^{-|U_1-U_2|}Ee^{-|V_1-V_2|}Ee^{-|W_1-W_2|}, \tag{1}$$

where (U_k, V_k, W_k), $k = 1, \ldots, 4$, are four independent copies of (U, V, W). Here the choice of the weight function $\omega(t_1, t_2, t_3)$ is mainly for the convenient analytic form of the integration. Different from the distance correlation (Székely et al., 2007), our integration exists without any moment conditions on the data, which is more widely applicable. Furthermore, the integration has an exact upper bound. To see this, with the fact that $U \perp\!\!\!\perp W$ and $V \perp\!\!\!\perp W$, (1) boils down to

$$E\left\{S_U(U_1,U_2)S_V(V_1,V_2)e^{-|W_1-W_2|}\right\}, \tag{2}$$

where $S_U(U_1, U_2)$ and $S_V(V_1, V_2)$ are defined as

$$\begin{aligned} S_U(U_1,U_2) &= E\left\{e^{-|U_1-U_2|} + e^{-|U_3-U_4|} - e^{-|U_1-U_3|} - e^{-|U_2-U_3|} \mid (U_1,U_2)\right\}, \\ S_V(V_1,V_2) &= E\left\{e^{-|V_1-V_2|} + e^{-|V_3-V_4|} - e^{-|V_1-V_3|} - e^{-|V_2-V_3|} \mid (V_1,V_2)\right\}. \end{aligned}$$

Recall that U, V and W are uniformly distributed on $(0, 1)$. With further calculations based on (2), we obtain a normalized index and define it as ρ to measure the mutual dependence:

$$\begin{aligned} \rho(X,Y \mid Z) &= c_0E\{ \left(e^{-|U_1-U_2|} + e^{-U_1} + e^{U_1-1} + e^{-U_2} + e^{U_2-1} + 2e^{-1} - 4\right) \\ & \left(e^{-|V_1-V_2|} + e^{-V_1} + e^{V_1-1} + e^{-V_2} + e^{V_2-1} + 2e^{-1} - 4\right) e^{-|W_1-W_2|}\}, \end{aligned} \tag{3}$$

where $c_0 = (13e^{-3} - 40e^{-2} + 13e^{-1})^{-1}$. Several appealing properties of the proposed index $\rho(X, Y \mid Z)$ are summarized in Theorem 2.

Theorem 2 *Suppose that the conditions in Proposition 1 are fulfilled. The index $\rho(X,Y \mid Z)$ defined in (3) has the following properties:*

(1) $0 \le \rho(X,Y \mid Z) \le 1$, $\rho(X,Y \mid Z) = 0$ holds if and only if $X \perp\!\!\!\perp Y \mid Z$. Furthermore, if $F_{X|Z}(X \mid Z) = F_{Y|Z}(Y \mid Z)$ or $F_{X|Z}(X \mid Z) + F_{Y|Z}(Y \mid Z) = 1$, then $\rho(X,Y \mid Z) = 1$.

(2) The index ρ is symmetric conditioning on Z. That is, $\rho(X,Y \mid Z) = \rho(Y,X \mid Z)$.

(3) For any strictly monotone transformations $m_1(\cdot)$, $m_2(\cdot)$ and $m_3(\cdot)$, $\rho(X,Y \mid Z) = \rho\{m_1(X), m_2(Y) \mid m_3(Z)\}$.

The step-by-step derivation of $\rho(X,Y \mid Z)$ and proof of Theorem 2 are presented in the appendix. Property (1) indicates that the index ρ ranges from zero to one, equals zero when the conditional independence holds, and is equal to one if Y is a strictly monotone transformation of X conditional on Z. Property (2) shows that the index ρ is a symmetric measure of conditional dependence. Property (3) illustrates that the index ρ is invariant to any strictly monotone transformation. In fact, ρ is not only invariant to marginal strictly monotone transformations, but also invariant to strictly monotone transformations conditional on Z. For example, it can be verified that $\rho(X,Y \mid Z) = \rho[m_1\{X - E(X \mid Z)\}, Y \mid Z]$.

2.2 Asymptotic properties

In this section, we establish the asymptotic properties of the sample version of the proposed index under the null and alternative hypothesis. Consider independent and identically distributed samples $\{X_i, Y_i, Z_i\}$, $i = 1, \dots, n$. To estimate the proposed index $\rho(X,Y \mid Z)$, we apply the nonparametric kernel estimator for the conditional cumulative distribution function (Li and Racine, 2007). Specifically, define

$$\begin{aligned}
\widehat{f}_Z(z) &= n^{-1}\sum_{i=1}^{n} K_h(z - Z_i),\\
\widehat{U} = \widehat{F}_{X|Z}(x \mid z) &= n^{-1}\sum_{i=1}^{n} K_h(z - Z_i)\mathbb{1}(X_i \le x)/\widehat{f}_Z(z),\\
\widehat{V} = \widehat{F}_{Y|Z}(y \mid z) &= n^{-1}\sum_{i=1}^{n} K_h(z - Z_i)\mathbb{1}(Y_i \le y)/\widehat{f}_Z(z),
\end{aligned}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is the bandwidth. Besides, we use empirical distribution function to estimate the cumulative distribution function, i.e., $\widehat{W} = \widehat{F}_Z(z) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(Z_i \le z)$. The sample version of the index, denoted by $\widehat{\rho}(X,Y \mid Z)$, is thus given by

$$\begin{aligned}
\widehat{\rho}(X,Y \mid Z) \;=\; & c_0 n^{-2}\sum_{i,j}\Big\{\Big(e^{-|\widehat{U}_i-\widehat{U}_j|} + e^{-\widehat{U}_i} + e^{\widehat{U}_i-1} + e^{-\widehat{U}_j} + e^{\widehat{U}_j-1} + 2e^{-1} - 4\Big)\\
& \Big(e^{-|\widehat{V}_i-\widehat{V}_j|} + e^{-\widehat{V}_i} + e^{\widehat{V}_i-1} + e^{-\widehat{V}_j} + e^{\widehat{V}_j-1} + 2e^{-1} - 4\Big)e^{-|\widehat{W}_i-\widehat{W}_j|}\Big\}.
\end{aligned}$$

One can also obtain a normalized index ρ_0, which is a direct normalization based on (1) without considering $U \perp\!\!\!\perp W$ and $V \perp\!\!\!\perp W$:

$$\begin{aligned}
\rho_0(X,Y \mid Z) \;=\; & c_0\Big\{Ee^{-|U_1-U_2|-|V_1-V_2|-|W_1-W_2|} + 8e^{-3}\\
& -2E\left(2 - e^{-U_1} - e^{U_1-1}\right)\left(2 - e^{-V_1} - e^{V_1-1}\right)\left(2 - e^{-W_1} - e^{W_1-1}\right)\Big\}.
\end{aligned}$$

The corresponding moment estimator is

$$\begin{aligned}
\widehat{\rho}_0(X,Y \mid Z) \;=\; & c_0\left\{n^{-2}\sum_{i,j} e^{-|\widehat{U}_i-\widehat{U}_j|-|\widehat{V}_i-\widehat{V}_j|-|\widehat{W}_i-\widehat{W}_j|} + 8e^{-3}\right.\\
& \left. -2n^{-1}\sum_{i=1}^{n}\left(2 - e^{-\widehat{U}_i} - e^{\widehat{U}_i-1}\right)\left(2 - e^{-\widehat{V}_i} - e^{\widehat{V}_i-1}\right)\left(2 - e^{-\widehat{W}_i} - e^{\widehat{W}_i-1}\right)\right\}.
\end{aligned}$$

Although $\rho(X, Y \mid Z) = \rho_0(X, Y \mid Z)$ at the population level, those two statistics $\widehat{\rho}(X, Y \mid Z)$ and $\widehat{\rho}_0(X, Y \mid Z)$ exhibit different properties at the sample level. This is because $\widehat{\rho}(X, Y \mid Z)$ considers the fact that $U \perp\!\!\!\perp W$ and $V \perp\!\!\!\perp W$. But on the other hand, $\widehat{\rho}_0(X, Y \mid Z)$ is only a regular mutual independence test statistic, where $\widehat{U}_i$, $\widehat{V}_i$, and $\widehat{W}_i$ are exchangeable. When $X \perp\!\!\!\perp Y \mid Z$, under Conditions 1-4 listed below, $\widehat{\rho}(X, Y \mid Z)$ is of order $(n^{-1} + h^{4m})$, while $\widehat{\rho}_0(X, Y \mid Z)$ is of order $(n^{-1} + h^{2m})$ because of the bias caused by nonparametric estimation. Note that m is the order of kernel functions and equal to 2 when using regular kernel functions such as Gaussian and epanechnikov kernels. This indicates that $\widehat{\rho}(X, Y \mid Z)$ is essentially n consistent without under-smoothing while $\widehat{\rho}_0(X, Y \mid Z)$ typically requires under-smoothing. In addition, our statistic $\widehat{\rho}(X, Y \mid Z)$ has the same asymptotic properties as if U, V and W are observed, but $\widehat{\rho}_0(X, Y \mid Z)$ does not. See Figure 1 for a numerical comparison between the empirical null distributions of the two statistics.

We next study the asymptotical behaviors of the estimated index, $\widehat{\rho}(X, Y \mid Z)$, under both the null and the alternative hypotheses. The following regularity conditions are imposed to facilitate our subsequent theoretical analyses. In what follows, we derive the limiting distribution of $\widehat{\rho}(X, Y \mid Z)$ under the null hypothesis in Theorem 3.

Condition 1. The univariate kernel function $K(\cdot)$ is symmetric about zero and Lipschitz continuous. In addition, it satisfies

$$\int K(v)dv = 1, \quad \int v^i K(v)dv = 0, 1 \le i \le m-1, \quad 0 \ne \int v^m K(v)dv < \infty.$$

Condition 2. The bandwidth h satisfies $nh^2/\log^2(n) \to \infty$, and $nh^{4m} \to 0$.

Condition 3. The probability density function of Z, denoted by $f_Z(z)$ is bounded away from 0 to infinity.

Condition 4. The $(m-1)$th derivatives of $F_{X|Z}(x \mid z)f(z)$, $F_{Y|Z}(y \mid z)f(z)$ and $f_Z(z)$ with respect to z are locally Lipschitz-continuous.

Theorem 3 *Suppose that Conditions 1-4 hold and the conditions in Proposition 1 are fulfilled. Under the null hypothesis,*

$$n\widehat{\rho}(X, Y \mid Z) \to c_0 \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

in distribution, where $\chi_j^2(1)$, $j = 1, 2, \dots$ are independent chi-square random variables with one degree of freedom, and $\lambda_j s$, $j = 1, 2, \dots$ are eigenvalues of

$$\begin{aligned} h(u, v, w; u', v', w') = \; & (e^{-|u-u'|} + e^{-u} + e^{u-1} + e^{-u'} + e^{u'-1} + 2e^{-1} - 4)(e^{-|v-v'|} \\ & + e^{-v} + e^{v-1} + e^{-v'} + e^{v'-1} + 2e^{-1} - 4)e^{-|w-w'|}. \end{aligned}$$

That is, there exists orthonormal eigenfunction $\Phi_j(u, v, w)$ such that

$$\int_0^1 \int_0^1 \int_0^1 h(u, v, w; u', v', w')\Phi_j(u', v', w')du'dv'dw' = \lambda_j \Phi_j(u, v, w).$$

The proof of Theorem 3 is given in the appendix. To understand the asymptotic distributions intuitively, we showed in the proof that $n\widehat{\rho}(X, Y \mid Z)$ can be approximated by degenerate V-statistics, i.e.,

$$n\widehat{\rho}(X, Y \mid Z) = c_0 n^{-1} \sum_{i,j} h(U_i, V_i, W_i; U_j, V_j, W_j) + o_p(n^{-1}),$$

and $E\{h(U_i, V_i, W_i; U_j, V_j, W_j) \mid (U_i, V_i, W_i)\} = 0$. By the spectral decomposition,

$$h(u, v, w; u', v', w') = \sum_{j=1}^{\infty} \lambda_j \Phi_j(u, v, w) \Phi_j(u', v', w').$$

Therefore, $n\widehat{\rho}(X, Y \mid Z) = c_0 \sum_{j=1}^{\infty} \lambda_j \{n^{-1/2} \sum_{i=1}^{n} \Phi_j(U_i, V_i, W_i)\}^2 + o_p(n^{-1})$, which converges in distribution to the weighted sum of independent chi squared distributions provided in Theorem 3 because $n^{-1/2} \sum_{i=1}^{n} \Phi_j(U_i, V_i, W_i)$ is asymptotically standard normal (Korolyuk and Borovskich, 2013). Moreover, the λ_js, $j = 1, 2, \ldots$ are real numbers associated with the distribution of U, V and W, all of which follow uniform distributions on $[0, 1]$. In addition, U, V and W are mutually independent under the null hypothesis. This indicates that the proposed test statistic is essentially distribution free under the null hypothesis. However, the critical value may be difficult to calculate because of the complicated form of limiting distribution. Therefore, we suggest a simulation procedure to approximate the null distribution and decide the critical value, which is commonly used in the literature (e.g., Székely et al., 2007; Zhu et al., 2018). The simulation procedure can be independent of the original data and hence greatly improved the computation efficiency. In what follows, we describe the simulation-based procedure in detail to decide the critical value c_α.

1. Generate $\{U_i^*, V_i^*, W_i^*\}$, $i = 1, \ldots, n$ independently from mutually independent standard uniform distributions;

2. Compute the statistic $\widehat{\rho^*}$ based on $\{U_i^*, V_i^*, W_i^*\}$, $i = 1, \ldots, n$, i.e.,

$$\begin{aligned} \widehat{\rho^*} \;=\; & c_0 n^{-2} \sum_{i,j} \Big\{ \Big(e^{-|U_i^* - U_j^*|} + e^{-U_i^*} + e^{U_i^* - 1} + e^{-U_j^*} + e^{U_j^* - 1} + 2e^{-1} - 4 \Big) \\ & \Big(e^{-|V_i^* - V_j^*|} + e^{-V_i^*} + e^{V_i^* - 1} + e^{-V_j^*} + e^{V_j^* - 1} + 2e^{-1} - 4 \Big) e^{-|W_i^* - W_j^*|} \Big\}. \end{aligned} \tag{4}$$

3. Repeat Steps 1-2 for B times and set c_α to be the upper α quantile of the estimated $\widehat{\rho^*}$ obtained from the randomly simulated samples.

Because (U^*, V^*, W^*) has the same distribution as that of (U, V, W) under the null hypothesis, it is straightforward that this simulation-based procedure can provide a valid approximation of the asymptotic null distribution of $\widehat{\rho}(X, Y \mid Z)$ when B is large. The consistency of this procedure is guaranteed by Theorem 4.

Theorem 4 *Under Conditions 1-4, it follows that*

$$n\widehat{\rho^*} \to c_0 \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1)$$

in distribution, where $\chi_j^2(1)$*,* $j = 1, 2, \dots$ *are independent* $\chi^2(1)$ *variables, and* λ_j*,* $j = 1, 2, \dots$ *are the same as that of Theorem 3.*

Theorem 4 shows that the simulation based approximation is valid. It also serves as the practical tool we use in the numerical analysis. Different from other conditional independence tests, our test only require one round of simulation to approximate the null distribution, which saves huge computation costs in practice.

Next, we study the power performance of the proposed test under two kinds of alternative hypotheses, under which the conditional independence no longer holds. We first consider the global alternative, denoted by H_{1g}, we have

$$H_{1g}: \quad X \not\perp\!\!\!\perp Y \mid Z.$$

We then consider a sequence of local alternatives, denoted by H_{1l},

$$H_{1l}: \quad F_{X|Z}(x \mid Z = z) - F_{X|(Y,Z)}\{x \mid (Y = y, Z = z)\} = n^{-1/2}\ell(x, y, z).$$

The asymptotical properties of the test statistics $\widehat{\rho}(X, Y \mid Z)$ under the global alternative and local alternatives are given in Theorem 5, whose proof is in the appendix. Theorem 5 shows that the proposed test is consistent against all fixed alternatives, and can detect local alternatives at rate $O(n^{-1/2})$.

Theorem 5 *Suppose that Conditions 1-4 hold and the conditions in Proposition 1 are fulfilled. Under* H_{1g}*, when* $nh^{2m} \to 0$*,*

$$n^{1/2}\{\widehat{\rho}(X, Y \mid Z) - \rho(X, Y \mid Z)\} \to \mathcal{N}(0, \sigma_0^2),$$

in distribution, where $\sigma_0^2 \overset{\text{def}}{=} 4c_0^2 \text{var}(P_{1,1} + P_{2,1} + P_{3,1} + P_{4,1})$*, and* $(P_{1,1}, P_{2,1}, P_{3,1}, P_{4,1})$ *are defined in (7)-(10) in the appendix, respectively.*

Under H_{1l}*,*

$$n\widehat{\rho}(X, Y \mid Z) \to \iiint \|\zeta(t_1, t_2, t_3)\|^2 \omega(t_1, t_2, t_3) dt_1 dt_2 dt_3,$$

in distribution, where $\zeta(t_1, t_2, t_3)$ *stands for a complex-valued Gaussian random process with mean function* $E\left[it_1\ell(X, Y, Z)e^{it_1U}\left\{e^{it_2V} - \varphi_V(t_2)\right\}e^{it_3W}\right]$ *and covariance function defined in (11) in the appendix, and* $\omega(t_1, t_2, t_3)$ *is the joint probability density function of three independent and identically distributed standard Cauchy random variables.*

Theorems 3 and 5 reveal that, $\widehat{\rho}$ is n-consistent under the null hypothesis while root-n consistent under the fixed alternative, indicating that the proposed test is consistent to detect the fixed alternative when ρ is bounded away from zero. Under the local alternative hypothesis, $\widehat{\rho}$ is n-consistent and the limiting distribution is different from that of the limiting null distribution, implying that the proposed test would have nontrivial power when the distance between the alternative and the null is $O(n^{-1/2})$.

3. Extensions

3.1 Multivariate continuous data

The methodology developed in Section 2 assumes that all the variables are univariate. In this section, we generalize the proposed index ρ to the multivariate case. Let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}} \in \mathbb{R}^q$ and $\mathbf{z} = (Z_1, \ldots, Z_r)^{\mathrm{T}} \in \mathbb{R}^r$ be continuous random vectors. More specifically, all elements of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are continuous random variables. Define $F_{X_1|\mathbf{z}}(X_1 \mid \mathbf{z})$ for the cumulative distribution function of X_1 given $\mathbf{z}$ and $F_{X_k|\mathbf{z},X_1,\ldots,X_{k-1}}(X_k \mid \mathbf{z}, X_1, \ldots, X_{k-1})$ for the cumulative distribution function of X_k given $\mathbf{z}, X_1, \ldots, X_{k-1}$ for $k = 2, \cdots, p$. Similar notation apply for $\mathbf{y}$ and $\mathbf{z}$. Denote $\widetilde{U}_1 = F_{X_1|\mathbf{z}}(X_1 \mid \mathbf{z})$, $\widetilde{V}_1 = F_{Y_1|\mathbf{z}}(Y_1 \mid \mathbf{z})$, $\widetilde{W}_1 = F_{Z_1}(Z_1)$,

$$
\begin{aligned}
\widetilde{U}_k &= F_{X_k|\mathbf{z},X_1,\ldots,X_{k-1}}(X_k \mid \mathbf{z}, X_1, \ldots, X_{k-1}),\ k = 2, \ldots, p,\\
\widetilde{V}_k &= F_{Y_k|\mathbf{z},Y_1,\ldots,Y_{k-1}}(Y_k \mid \mathbf{z}, Y_1, \ldots, Y_{k-1}),\ k = 2, \ldots, q,\\
\widetilde{W}_k &= F_{Z_k|Z_1,\ldots,Z_{k-1}}(Z_k \mid Z_1, \ldots, Z_{k-1}),\ k = 2, \ldots, r.
\end{aligned}
$$

Further denote $\mathbf{u} = (\widetilde{U}_1, \ldots, \widetilde{U}_p)^{\mathrm{T}}$, $\mathbf{v} = (\widetilde{V}_1, \ldots, \widetilde{V}_q)^{\mathrm{T}}$, $\mathbf{w} = (\widetilde{W}_1, \ldots, \widetilde{W}_r)^{\mathrm{T}}$. Similar to test of conditional independence for random variables, we first establish an equivalence between the conditional independence $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ and the mutual independence of $\mathbf{u}$,$\mathbf{v}$ and $\mathbf{w}$, which is stated in Theorem 6.

Theorem 6 *Assume that all the conditional cumulative distribution functions used in constructing* $\mathbf{u}$,$\mathbf{v}$ *and* $\mathbf{w}$ *are continuous for every given values, then* $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ *if and only if* $\mathbf{u}$,$\mathbf{v}$ *and* $\mathbf{w}$ *are mutually independent.*

The proof of Theorem 6 is illustrated in the appendix. Theorem 6 established an equivalence between the conditional independence $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ and the mutual independence among $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$. It is notable that when p, q, r are relatively large, $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ may be difficult to estimate because of the curse of dimensionality. In this paper, we mainly focus on the low dimensional case. Next, we develop the mutual independence test among $\mathbf{u}$,$\mathbf{v}$ and $\mathbf{w}$. Similar as the univariate case, we set the weight function to be the joint density of $(p+q+r)$ independent and identically distributed standard Cauchy random variables. We may further derive the closed form expression of $\rho(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$

$$
\rho(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = E\left\{S_{\mathbf{u}}(\mathbf{u}_1, \mathbf{u}_2) S_{\mathbf{v}}(\mathbf{v}_1, \mathbf{v}_2) e^{-\|\mathbf{w}_1 - \mathbf{w}_2\|_1}\right\},
$$

where $(\mathbf{u}_1, \mathbf{v}_1, \mathbf{w}_1)$ and $(\mathbf{u}_2, \mathbf{v}_2, \mathbf{w}_2)$ are two independent copies of $(\mathbf{u}, \mathbf{v}, \mathbf{w})$. Moreover,

$$
\begin{aligned}
S_{\mathbf{u}}(\mathbf{u}_1, \mathbf{u}_2) &= E\left\{e^{-\|\mathbf{u}_1 - \mathbf{u}_2\|_1} + e^{-\|\mathbf{u}_3 - \mathbf{u}_4\|_1} - e^{-\|\mathbf{u}_1 - \mathbf{u}_3\|_1} - e^{-\|\mathbf{u}_2 - \mathbf{u}_3\|_1} \mid (\mathbf{u}_1, \mathbf{u}_2)\right\},\\
S_{\mathbf{v}}(\mathbf{v}_1, \mathbf{v}_2) &= E\left\{e^{-\|\mathbf{v}_1 - \mathbf{v}_2\|_1} + e^{-\|\mathbf{v}_3 - \mathbf{v}_4\|_1} - e^{-\|\mathbf{v}_1 - \mathbf{v}_3\|_1} - e^{-\|\mathbf{v}_2 - \mathbf{v}_3\|_1} \mid (\mathbf{v}_1, \mathbf{v}_2)\right\}.
\end{aligned}
$$

and $\|\cdot\|_1$ is the ℓ_1 norm. Then $\rho(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ is nonnegative and equals zero if and only if $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. By estimating $\rho(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ consistently at the sample level, the resulting test is clearly consistent. To implement the test, it is still required to study the asymptotic distributions under the conditional independence using independent and identically distributed samples $\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i\}$, $i = 1, \ldots, n$. We also apply kernel estimator for the conditional cumulative distribution functions when estimating $\mathbf{u}_i, \mathbf{v}_i$ and $\mathbf{w}_i$. Specifically, we estimate $F_{A|B_1,\ldots,B_\ell}(a \mid b_1, \ldots, b_\ell)$ with

$$\widehat{F}_{A|B_1,\ldots,B_\ell}(a \mid b_1, \ldots, b_\ell) = \frac{\sum_{i=1}^n \mathbb{1}(A_i \leq a) \prod_{k=1}^\ell K_h(B_{ik} - b_k)}{\sum_{i=1}^n \prod_{k=1}^\ell K_h(B_{ik} - b_k)}.$$

where $(A, B_1, \ldots, B_\ell)^{\mathrm{T}}$ are $(Z_k, Z_1, \ldots, Z_{k-1})^{\mathrm{T}}$, $k = 2, \ldots, r$, $(X_\ell, X_1, \ldots, X_{\ell-1}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$, $\ell = 1, \ldots, p$, or $(Y_j, Y_1, \ldots, Y_{j-1}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$, $j = 1, \ldots, q$ when estimating $\mathbf{w}$, $\mathbf{u}$ and $\mathbf{v}$, respectively. The sample version of $\rho(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ is given by

$$\begin{aligned}\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) \quad = \quad & n^{-2} \sum_{i,j} E\Big[\Big\{ e^{-\|\widehat{\mathbf{u}}_i - \widehat{\mathbf{u}}_j\|_1} + e^{-\|\mathbf{u} - \mathbf{u}'\|_1} - e^{-\|\widehat{\mathbf{u}}_i - \mathbf{u}\|_1} - e^{-\|\mathbf{u} - \widehat{\mathbf{u}}_j\|_1} \mid (\widehat{\mathbf{u}}_i, \widehat{\mathbf{u}}_j) \Big\} \\ & E\Big\{ e^{-\|\widehat{\mathbf{v}}_i - \widehat{\mathbf{v}}_j\|_1} + e^{-\|\mathbf{v} - \mathbf{v}'\|_1} - e^{-\|\widehat{\mathbf{v}}_i - \mathbf{v}\|_1} - e^{-\|\mathbf{v} - \widehat{\mathbf{v}}_j\|_1} \mid (\widehat{\mathbf{v}}_i, \widehat{\mathbf{v}}_j) \Big\} e^{-\|\widehat{\mathbf{w}}_i - \widehat{\mathbf{w}}_j\|_1} \Big],\end{aligned}$$

where $(\mathbf{u}', \mathbf{v}')$ is an independent copy of $(\mathbf{u}, \mathbf{v})$, and further calculations yield that $\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ is equal to

$$\begin{aligned} & n^{-2} \sum_{i,j} \Bigg[\Bigg\{ e^{-\|\widehat{\mathbf{u}}_i - \widehat{\mathbf{u}}_j\|_1} + (\frac{2}{e})^p - \prod_{k=1}^p (2 - e^{-\widehat{\widehat{U}}_{ik}-1} - e^{-\widehat{\widehat{U}}_{ik}}) - \prod_{k=1}^p (2 - e^{-\widehat{\widehat{U}}_{jk}-1} - e^{-\widehat{\widehat{U}}_{jk}}) \Bigg\} \\ & \Bigg\{ e^{-\|\widehat{\mathbf{v}}_i - \widehat{\mathbf{v}}_j\|_1} + (\frac{2}{e})^q - \prod_{k=1}^q (2 - e^{-\widehat{\widehat{V}}_{ik}-1} - e^{-\widehat{\widehat{V}}_{ik}}) - \prod_{k=1}^q (2 - e^{-\widehat{\widehat{V}}_{jk}-1} - e^{-\widehat{\widehat{V}}_{jk}}) \Bigg\} e^{-\|\widehat{\mathbf{w}}_i - \widehat{\mathbf{w}}_j\|_1} \Bigg].\end{aligned}$$

We next study the asymptotical behaviors of $\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ under the null hypothesis in Theorem 7, whose proof is given in the appendix. We begin by providing some regularity conditions for the multivariate data.

Condition $2'$. The bandwidth h satisfies $nh^{2(r+p-1)}/\log^2(n) \to \infty$, $nh^{2(r+q-1)}/\log^2(n) \to \infty$, and $nh^{4m} \to 0$.

Condition $3'$. The probability density function of the random vector $(Z_1, \ldots, Z_k)^{\mathrm{T}}$, $k = 1, \ldots, r$, $(\mathbf{z}^{\mathrm{T}}, X_1, \ldots, X_\ell)^{\mathrm{T}}$, $\ell = 1, \ldots, p-1$, and $(\mathbf{z}^{\mathrm{T}}, Y_1, \ldots, Y_j)^{\mathrm{T}}$, $j = 1, \ldots, q-1$, are all bounded away from 0 to infinity.

Condition $4'$. The $(m-1)$th derivatives of $F_{A|\mathbf{B}}(a \mid \mathbf{b}) f_{\mathbf{B}}(\mathbf{b})$, and $f_{\mathbf{B}}(\mathbf{b})$ with respect to $\mathbf{b}$ are locally Lipschitz-continuous, where $(A, \mathbf{B}^{\mathrm{T}})^{\mathrm{T}}$ can be any one of $(Z_k, Z_1, \ldots, Z_{k-1})^{\mathrm{T}}$, $k = 2, \ldots, r$, $(X_\ell, X_1, \ldots, X_{\ell-1}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$, $\ell = 1, \ldots, p$, or $(Y_j, Y_1, \ldots, Y_{j-1}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$, $j = 1, \ldots, q$.

Theorem 7 *Suppose that Conditions 1 and $2'$-$4'$ hold and the conditions in Theorem 6 are fulfilled. Under the null hypothesis,*

$$n\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) \to \sum_{j=1}^\infty \lambda_j \chi_j^2(1),$$

in distribution, where $\chi_j^2(1)$, $j = 1, 2, \ldots$ are independent chi-square random variables with one degree of freedom, and λ_js, $j = 1, 2, \ldots$ are eigenvalues of

$$\widetilde{h}(\mathbf{u}, \mathbf{v}, \mathbf{w}; \mathbf{u}', \mathbf{v}', \mathbf{w}') = S_{\mathbf{u}}(\mathbf{u}, \mathbf{u}') S_{\mathbf{v}}(\mathbf{v}, \mathbf{v}') e^{-\|\mathbf{w} - \mathbf{w}'\|_1}.$$

That is, there exists orthonormal eigenfunction $\Phi_j(\mathbf{u}, \mathbf{v}, \mathbf{w})$ *such that*

$$\iiint_{[0,1]^{p+q+r}} \widetilde{h}(\mathbf{u}, \mathbf{v}, \mathbf{w}; \mathbf{u}', \mathbf{v}', \mathbf{w}')\Phi_j(\mathbf{u}', \mathbf{v}', \mathbf{w}')d\mathbf{u}'d\mathbf{v}'d\mathbf{w}' = \lambda_j\Phi_j(\mathbf{u}, \mathbf{v}, \mathbf{w}).$$

3.2 Discrete data

In this section, we discuss the setting in which X, Y and Z are univariate discrete random variables. Specifically, we apply transformations in Brockwell (2007) to obtain U and V. Define $F_{X|Z}(x \mid z) = \Pr(X \le x \mid Z = z)$, $F_{X|Z}(x- \mid z) = \Pr(X < x \mid Z = z)$, $F_{Y|Z}(y \mid z) = \Pr(Y \le y \mid Z = z)$, and $F_{Y|Z}(y- \mid z) = \Pr(Y < y \mid Z = z)$. We further let U_X and U_Y be two independent and identically distributed $U(0, 1)$ random variables, and apply the transformations

$$\begin{aligned} U &= (1 - U_X)F_{X|Z}(X- \mid Z) + U_X F_{X|Z}(X \mid Z), \\ V &= (1 - U_Y)F_{Y|Z}(Y- \mid Z) + U_Y F_{Y|Z}(Y \mid Z). \end{aligned}$$

According to Brockwell (2007), both U and V are uniformly distributed on $(0, 1)$. In addition, $U \perp\!\!\!\perp Z$ and $V \perp\!\!\!\perp Z$. In the following proposition, we establish the equivalence between the conditional independence and the mutual independence.

Theorem 8 *For discrete random variables* X*,* Y *and* Z*,* $X \perp\!\!\!\perp Y \mid Z$ *if and only if* U, V *and* Z *are mutually independent.*

The proof of Theorem 8 is presented in the appendix. With Theorem 8, we turn a discrete conditional independence problem into a mutual independence one. Hence similar techniques can be readily applied for the mutual independence test and we omit them to avoid verbosity.

4. Numerical Validations

4.1 Conditional independence test

In this section, we investigate the finite sample performance of the proposed methods. To begin with, we illustrate that the null distribution of $\widehat{\rho}(X, Y \mid Z)$ is indeed distribution free as if U, V and W can be observed, and is insensitive to the bandwidth of the nonparametric kernel. In comparison, the null distribution of $\widehat{\rho}_0(X, Y \mid Z)$ does not enjoy such properties. To facilitate the analysis, let $X = Z + \varepsilon_1$ and $Y = Z + \varepsilon_2$, where Z, ε_1 and ε_2 are independent and identically distributed. We consider three scenarios where Z, ε_1, ε_2 are independently drawn from normal distribution $N(0, 1)$, uniform distribution $U(0, 1)$, and exponential distribution $Exp(1)$, respectively. It is clear that X and Y are conditionally independent given Z. The sample size n is set to be 100.

The simulated null distributions based on $n\widehat{\rho}(X, Y \mid Z)$ and $n\widehat{\rho}_0(X, Y \mid Z)$ are depicted in Figure 1. The estimated kernel density curves of $n\widehat{\rho}(X, Y \mid Z)$ based on 1000 repetitions are shown in Figure 1(a), where the reference curve is generated by the simulation-based statistic $n\widehat{\rho^*}(X, Y \mid Z)$ defined in (4). Clearly all the estimated density curves are close to the reference, indicating that limiting null distribution of the estimated index is indeed distribution free as if no kernel estimation is involved. In comparison, we apply the same

Table 1: Empirical power of tests of conditional independence for Models M2 - M6 for different bandwidth ch_0, where c increase from 0.5 to 1.5. The significance level $\alpha = 0.05$, $n = 100$.

	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
M2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
M3	0.957	0.962	0.98	0.977	0.975	0.971	0.972	0.968	0.955	0.960	0.956
M4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
M5	0.999	1.000	0.997	0.998	0.999	0.997	0.995	0.997	0.999	0.997	0.997
M6	0.999	0.999	0.999	1.000	0.999	0.999	0.996	1.000	0.999	1.000	1.000

simulation settings for $\widehat{\rho}_0(X, Y \mid Z)$ and plot the null distributions in Figure 1(b), from which it can be seen that the involved estimation of U and V significantly influences the null distribution of $\widehat{\rho}_0(X, Y \mid Z)$. To show the insensitivity of the choice of the bandwidth, we set the bandwidths to be ch_0, where $c = 0.5$, 1, and 2, respectively, and h_0 is the bandwidth obtained by the rule of thumb. The estimated kernel density curves of $n\widehat{\rho}(X, Y \mid Z)$ and $n\widehat{\rho}_0(X, Y \mid Z)$ based normal distributions with 1000 repetitions, together with the reference curve, are shown in Figure 1(c) and (d), from which we can see that the null distributions of $n\widehat{\rho}(X, Y \mid Z)$ almost remain the same for all choices of the bandwidths, implying that our test is insensitive to the bandwidth of the nonparametric kernel. However, the null distributions of $n\widehat{\rho}_0(X, Y \mid Z)$ can be dramatically influenced when the bandwidth changes.

Next, we perform the sensitivity analysis under the alternative hypothesis using Models M2 - M6, which will be listed shortly. We fix the sample size $n = 100$ and set the significance level $\alpha = 0.05$. To inspect how the power performance is varied with the choice of the bandwidth, we set the bandwidths to be ch_0, where h_0 is the bandwidth obtained by the rule of thumb and increase c from 0.5 to 1.5 step by 0.1. The respective empirical powers are charted in Table 1, from which we can see that the test is the most powerful when c is around 1. Therefore, we advocate using the rule of thumb (i.e., $c = 1$) to decide the bandwidth in practice.

We compare our proposed conditional independence test (denoted by "CIT") with some popular nonlinear conditional dependence measure. They are, respectively, the conditional distance correlation (Wang et al., 2015, denoted by "CDC"), conditional mutual information (Scutari, 2010, denoted by "CMI"), and the KCI.test (Zhang et al., 2011, denoted by "KCI"). We conduct 500 replications for each scenario. The critical values of the CIT are obtained by conducting 1000 simulations. We first consider the following models with random variable Z. In (M1), $X \perp\!\!\!\perp Y \mid Z$. This model is designed for examining the empirical Type I error rate. While (M2)$-$(M6) are designed for examining the power of the proposed test of conditional independence. Moreover, for M1 - M3, we generate $\widetilde{X}_1$, $\widetilde{X}_2$ and Z independently from $N(0, 1)$. For M4 - M6, we let $Z \sim N(0, 1)$, and generate $\widetilde{X}_1$, $\widetilde{X}_2 \sim t_1$ independently to investigate the power of the methods under heavy tailed distributions.

M1: $X = \widetilde{X}_1 + Z$, $Y = \widetilde{X}_2 + Z$.

M2: $X = \widetilde{X}_1 + Z$, $Y = \widetilde{X}_1^2 + Z$.

M3: $X = \widetilde{X}_1 + Z$, $Y = 0.5 \sin(\pi \widetilde{X}_1) + Z$.

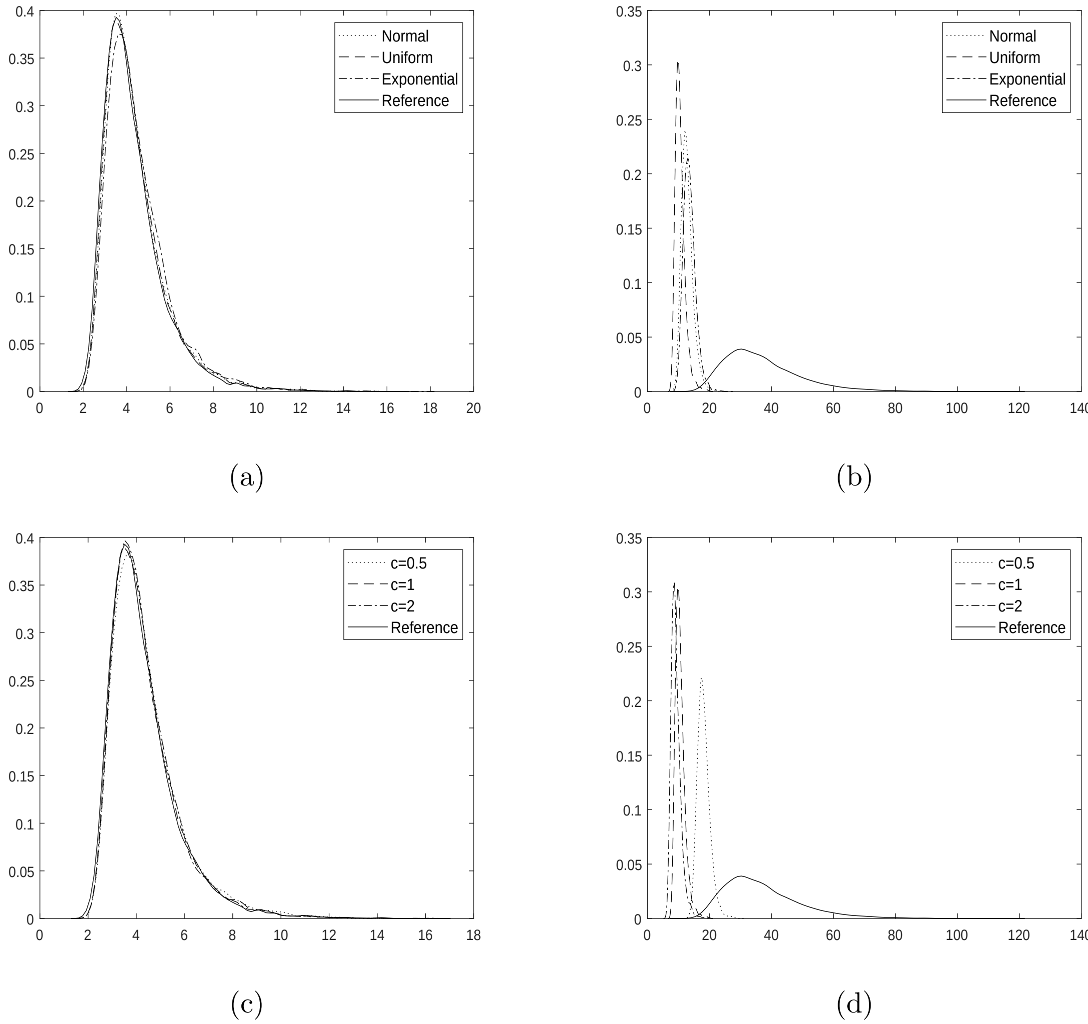

Figure 1: (a) and (b) are the simulated null distributions for $n\widehat{\rho}(X, Y \mid Z)$ and $n\widehat{\rho}_0(X, Y \mid Z)$ when data were generated from different distributions, while (c) and (d) are the simulated null distributions for $n\widehat{\rho}(X, Y \mid Z)$ and $n\widehat{\rho}_0(X, Y \mid Z)$ using different bandwidths, respectively.

M4: $X = X_1 + Z$, $Y = X_1 + X_2 + Z$.

M5: $X = \sqrt{|X_1 Z|} + Z$, $Y = 0.25X_1^2X_2^2 + X_2 + Z$.

M6: $X = \log(|X_1 Z| + 1) + Z$, $Y = 0.5(X_1^2 Z) + X_2 + Z$.

The empirical sizes for M1 and powers for the other five models at the significance levels $\alpha = 0.05$ and 0.1 are depicted in Table 2. In our simulation, we consider two sample sizes $n = 50$ and 100. Table 2 indicates that the empirical sizes of all the tests are all very close to the level α, which means that the Type I error can be controlled very well. As for the empirical power performance of models M2$-$M6, the proposed test outperforms other tests for both normal data and heavy tailed data, especially when $n = 50$.

Table 2: Empirical size and power of tests of conditional independence when Z is random variable at significance levels with $\alpha = 0.05$ and 0.1 and $n = 50$ and 100.

n	α	Test	M1	M2	M3	M4	M5	M6
50	0.05	CIT	0.056	1.000	0.572	1.000	0.954	0.888
		CDC	0.050	0.886	0.338	0.837	0.881	0.473
		CMI	0.048	0.380	0.070	0.829	0.898	0.448
		KCI	0.038	0.884	0.250	0.191	0.048	0.010
	0.1	CIT	0.098	1.000	0.712	1.000	0.974	0.938
		CDC	0.134	0.970	0.562	0.934	0.930	0.642
		CMI	0.088	0.484	0.132	0.854	0.912	0.485
		KCI	0.088	0.968	0.344	0.323	0.145	0.042
100	0.05	CIT	0.048	1.000	0.960	1.000	1.000	0.997
		CDC	0.066	0.998	0.694	0.918	0.971	0.624
		CMI	0.054	0.402	0.070	0.877	0.904	0.424
		KCI	0.040	1.000	0.444	0.371	0.095	0.020
	0.1	CIT	0.112	1.000	0.998	1.000	1.000	0.999
		CDC	0.158	1.000	0.834	0.974	0.985	0.745
		CMI	0.098	0.496	0.124	0.902	0.926	0.455
		KCI	0.088	1.000	0.598	0.513	0.199	0.057

We next examine the finite sample performance of the tests when Z is two-dimensional random vector, i.e., $\mathbf{z} = (Z_1, Z_2)$. M7 is designed for examining the size since $X \perp\!\!\!\perp Y \mid \mathbf{z}$. Five conditional dependent model M8–M12 are designed to examine the power of the tests. Similar as M1-M6, we generate $\widetilde{X}_1$, $\widetilde{X}_2$, Z_1 and Z_2 independently from $N(0, 1)$ in each of the following model.

M7: $X = \widetilde{X}_1 + Z_1 + Z_2$, $Y = \widetilde{X}_2 + Z_1 + Z_2$.

M8: $X = \widetilde{X}_1^2 + Z_1 + Z_2$, $Y = \log(\widetilde{X}_1 + 10) + Z_1 + Z_2$.

M9: $X = \tanh(\widetilde{X}_1) + Z_1 + Z_2$, $Y = \log(\widetilde{X}_1^2 + 10) + Z_1 + Z_2$.

M10: $X = \widetilde{X}_1^2 + Z_1 + Z_2$, $Y = \log(\widetilde{X}_1 Z_1 + 10) + Z_1 + Z_2$.

M11: $X = \widetilde{X}_1 + Z_1 + Z_2$, $Y = \sin(\widetilde{X}_1 Z_1) + Z_1 + Z_2$.

M12: $X = \log(\widetilde{X}_1 Z_1 + 10) + Z_1 + Z_2$, $Y = \exp(\widetilde{X}_1 Z_2) + Z_1 + Z_2$.

Lastly, we study the finite sample performance of the tests when X, Y and Z are all multivariate. Specifically, $\mathbf{x} = (X_1, X_2)$, $\mathbf{y} = (Y_1, Y_2)$, $\mathbf{z} = (Z_1, Z_2)$. M13 is designed to examine the size of the tests, while M14 -M18 are designed to study the powers. We generate $\widetilde{X}_1$, X_2, Y_2, Z_1 and Z_2 independently from $N(0, 1)$ for each model in M13-M18.

M13: $X_1 = \widetilde{X}_1 + Z_1$, $Y_1 = Z_1 + Z_2$.

M14: $X_1 = \log(\widetilde{X}_1 Z_1 + 100) + Z_1 + Z_2$, $Y_1 = \exp(\widetilde{X}_1 Z_1) + Z_1 + Z_2$.

M15: $X = \log(\widetilde{X}_1^2 + 100) + Z_1 + Z_2$, $Y_1 = 0.1\widetilde{X}_1^3 + Z_1 + Z_2$.

M16: $X = \log(\widetilde{X}_1 * Z_1 + 100) + Z_1 + Z_2$, $Y_1 = 0.5\widetilde{X}_1^3 Z_1^3 + Z_1 + Z_2$.

M17: $X = 0.1\exp(\widetilde{X}_1) + Z_1 + Z_2$, $Y_1 = \sin(\widetilde{X}_1) + |\widetilde{X}_1| + Z_1 + Z_2$.

M18: $X = \tanh(\widetilde{X}_1) + Z_1 + Z_2$, $Y_1 = 0.5\log(\widetilde{X}_1^2 + 100) + 0.5X_2 + Z_1 + Z_2$.

Table 3: Empirical size and power of conditional independence tests with $\mathbf{z}$ being random vector, level $\alpha = 0.05$ and 0.1, and sample size $n = 50$ and 100.

n	α	Test	M7	M8	M9	M10	M11	M12
50	0.05	CIT	0.046	0.672	0.906	0.686	0.440	0.788
		CDC	0.052	0.408	0.850	0.054	0.128	0.134
		CMI	0.072	0.426	0.192	0.392	0.118	0.300
		KCI	0.046	0.030	0.088	0.026	0.662	0.248
	0.1	CIT	0.092	0.792	0.948	0.798	0.582	0.874
		CDC	0.126	0.670	0.976	0.194	0.298	0.264
		CMI	0.120	0.506	0.292	0.500	0.210	0.386
		KCI	0.098	0.092	0.190	0.074	0.820	0.492
100	0.05	CIT	0.048	0.936	0.998	0.936	0.664	0.988
		CDC	0.084	0.890	1.000	0.920	0.972	0.306
		CMI	0.044	0.412	0.164	0.392	0.126	0.300
		KCI	0.044	0.038	0.172	0.028	0.990	0.358
	0.1	CIT	0.104	0.958	1.000	0.966	0.766	0.996
		CDC	0.168	0.978	1.000	0.986	0.992	0.456
		CMI	0.128	0.486	0.240	0.460	0.194	0.390
		KCI	0.090	0.092	0.358	0.108	0.998	0.608

Simulation results of models M7−M12 and models M13−M18 are summarized in Tables 3 and 4, respectively, from which it can be seen that the proposed method outperforms all other tests in terms of type I error and power. Furthermore, the numerical results seem to indicates that when the conditional set is large, the conditional mutual information and the kernel based conditional test tend to have relatively low power. The conditional distance correlation has high power but suffers huge computational burden.

4.2 Application to causal discovery

In this section, we consider a real application of conditional independence test in causal discovery of directed acyclic graphs. For a directed acyclic graph $G = (V, E)$, the nodes $V = \{1, 2, \ldots, p\}$ corresponds to a random vector $\mathbf{x} = (X_1, \ldots, X_p) \in \mathbb{R}^p$, and the set of edges $E \subset V \times V$ do not form any directed cycles. Two vertices X_1 and X_2 are d-separated by a subset of vertices S if every path between them is blocked by S. One may refer to Wasserman (2013) for a formal definition. Denote the joint distribution of $\mathbf{x}$ by $P(\mathbf{x})$. The joint distribution is said to be faithful with respect to a graph G if and only if for any $i, j \in V$, and any subset $S \subset V$,

$$X_i \perp\!\!\!\perp X_j \mid \{X_r : r \in S\} \Leftrightarrow \text{node } i \text{ and node } j \text{ are } d\text{-separated by the set } S.$$

One of the most famous algorithms for recovering the graphs satisfying the faithfulness assumption is the PC–algorithm (Spirtes et al., 2000; Kalisch and Bühlmann, 2007). The algorithm could recover the graph up to its Markov equivalence class, which are sets of graphs that entail the same set of (conditional) independencies. The performance of the PC–

Table 4: Empirical size and power of conditional independence tests when $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are random vectors. Level $\alpha = 0.05$ and 0.1, and sample size $n = 50$ and 100.

n	α	Test	M13	M14	M15	M16	M17	M18
50	0.05	CIT	0.05	1.000	1.000	1.000	0.363	0.986
		CDC	0.019	0.022	0.984	0.304	0.120	0.833
		CMI	0.01	0.582	0.812	0.205	0.082	0.020
		KCI	0.036	0.036	0.047	0.042	0.052	0.688
	0.1	CIT	0.100	1.000	1.000	1.000	0.564	0.997
		CDC	0.092	0.064	1.000	0.733	0.262	0.939
		CMI	0.028	0.6	0.915	0.294	0.170	0.054
		KCI	0.086	0.081	0.099	0.083	0.118	0.886
100	0.05	CIT	0.026	1.000	1.000	1.000	0.873	1
		CDC	0.048	0.032	1.000	0.965	0.38	0.999
		CMI	0.004	0.498	1.000	0.211	0.218	0.036
		KCI	0.044	0.042	0.052	0.05	0.068	0.999
	0.1	CIT	0.077	1.000	1.000	1.000	0.965	1.000
		CDC	0.127	0.13	1.000	1.000	0.567	1.000
		CMI	0.013	0.523	1.000	0.338	0.378	0.077
		KCI	0.087	0.077	0.102	0.091	0.119	1.000

algorithm relies heavily on the (conditional) independence tests because small mistakes at the beginning of the algorithm may lead to a totally different directed acyclic graph (Zhang et al., 2011). One of the most popular approach for testing conditional independence is the partial correlation, under the assumption that the joint distribution $P(\mathbf{x})$ follows Gaussian distribution and the nodes relationship is linear (Kalisch and Bühlmann, 2007). Conditional mutual information (Scutari, 2010) is another possible option. Zhang et al. (2011) proposed a kernel-based conditional independence test for causal discovery in directed acyclic graphs. In this section, we demonstrate how the proposed conditional independence index can be applied for causal discovery in real data. Additional simulation results are relegated into the appendix.

We analyze a real data set originally from the National Institute of Diabetes and Digestive and Kidney Diseases (Smith et al., 1988). The dataset consists of serval medical predictor variables for the outcome of diabetes. We are interested in the causal structural of five variables: age, body mass index, 2-hour serum insulin, plasma glucose concentration and diastolic blood pressure. After removing the missing data, we obtain $n = 392$ samples. The PC–algorithm is applied to examine the causal structure of the five variables based on the four different conditional independence measures. We implement the causal algorithms by the R package *pcalg* (Kalisch et al., 2012). The estimated causal structure are shown in Figure 2. The proposed test gives the same estimated graph as the partial correlation, since the data is approximately normally distributed. To interpret the graph, note that age is likely to affect the diastolic blood pressure. The plasma glucose concentration level is also likely to be related to age. This is confirmed by the causal findings of (a), (b) and (c) in Figure 2. Besides, serum insulin has plausible causal effects on body mass index, and is also related to plasma glucose concentration. The causal relationship between age and blood pressure is not confirmed in part (c), the test of conditional mutual information. This is not a surprise given the high false positive rate reported in Table 5 in the appendix. The kernel based conditional independence is a little conservative and is not able to detect some of the possible edges. To further illustrate the robustness of the proposed test, we make a logarithm transformation on the data, and apply the same procedure again. The estimated causal structures are reported in Figure 3. We observe that the proposed test results in the same estimated structure as the original data, which echos property (4) in Theorem 2, i.e., the proposed test is invariant with respect to monotone transformations. However, the partial correlation test yields more false positives, since the normality assumption is violated.

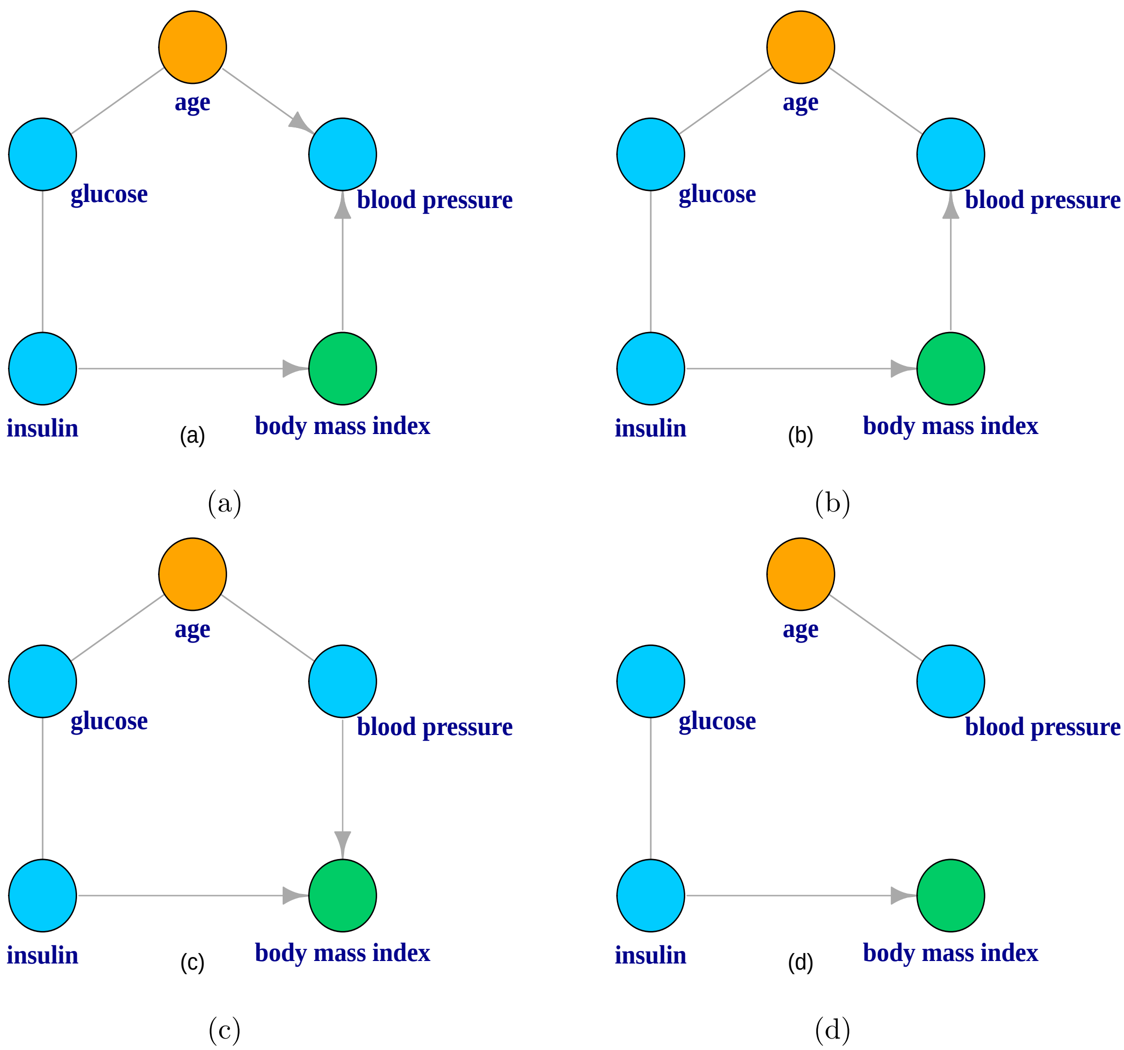

Figure 2: The estimated causal structure of the five variables by using the proposed test in (a), partial correlation in (b), conditional mutual information in (c) and kernel based conditional independence test in (d).

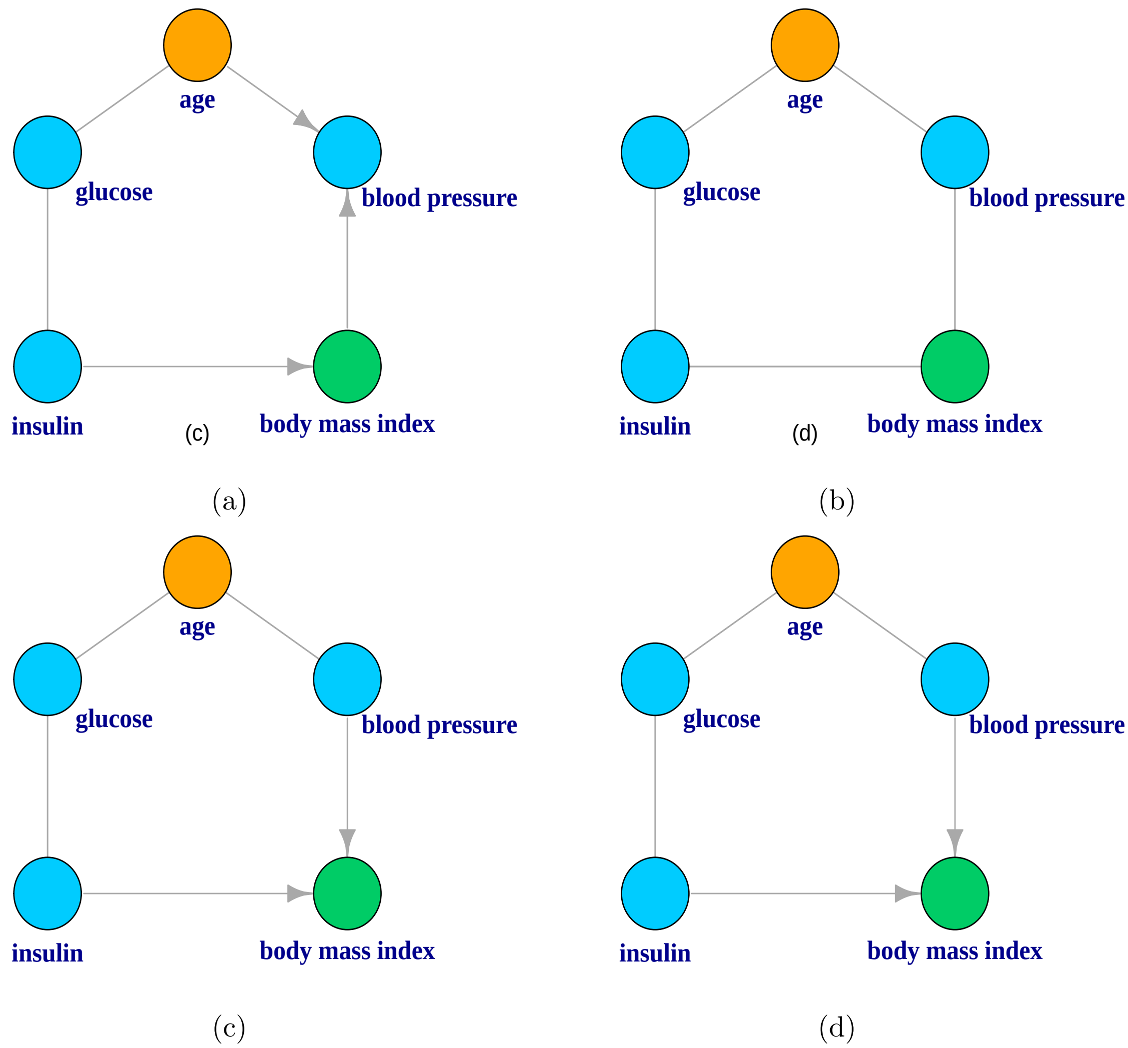

Figure 3: The estimated causal structure of the log-transformed five variables by using the four tests. Refer to the caption of Figure 2 for the four tests.

5. Discussions

In this paper we developed a new index to measure conditional dependence of random variables and vectors. The calculation of the estimated index requires low computational cost. The test of conditional independence based on the newly proposed index has nontrivial power against all fixed and local alternatives. The proposed test is distribution free under the null hypothesis, and is robust to outliers and heavy-tailed data. Numerical simulations indicate that the proposed test is more powerful than some existing ones. The proposed test is further applied to directed acyclic graphs for causal discovery and shows superior performance.

Acknowledgments

The authors would like to thank the action editor and reviewers for their constructive suggestions, which lead to a significant improvement of this work. All authors contributed equally to this work, the authors are listed in the alphabetic order, and Yaowu Zhang is the corresponding author. Cai and Li's research was supported by National Science Foundation (NSF), DMS 1820702, 1953196 and 2015539. Zhang's research was supported by National NSF of China (NNSFC) (11801349, 72192832) and the Program for Innovative Research Team of Shanghai University of Finance and Economics.

Appendix A. Technical Proofs

A.1 Proof of Proposition 1

For $0 < u < 1$, define quantile function for $X \mid Z$ as

$$F_{X|Z}^{-1}(u \mid Z = z) = \inf\{x : F_{X|Z}(x \mid Z = z) \geq u\}.$$

Similarly, we can define $F_{Y|Z}^{-1}(v \mid Z = z)$, the quantile function for $Y \mid Z$, for $0 < v < 1$. Since X and Y have continuous conditional distribution functions for every given value of Z, it follows that when $0 < u < 1$ and $0 < v < 1$,

$$\begin{aligned} & \Pr\{F_{X|Z}(X \mid Z) \leq u, F_{Y|Z}(Y \mid Z) \leq v \mid Z = z\} \\ = \; & \Pr\{X \leq F_{X|Z}^{-1}(u \mid Z), Y \leq F_{Y|Z}^{-1}(v \mid Z) \mid Z = z\}. \end{aligned}$$

This implies that $X \perp\!\!\!\perp Y \mid Z$ is equivalent to $U \perp\!\!\!\perp V \mid Z$. In addition, conditional on $Z = z$, $F_{X|Z}(X \mid Z = z)$ is uniformly distributed on $(0, 1)$, which does not depend on the particular value of z, indicating $F_{X|Z}(X \mid Z) \perp\!\!\!\perp Z$. That is, $U \perp\!\!\!\perp Z$. Similarly, $V \perp\!\!\!\perp Z$. Thus, the conditional independence $f_{U,V|Z}(u, v \mid z) = f_{U|Z}(u \mid z) f_{V|Z}(v \mid z)$ together with $f_{U|Z}(u \mid z) = f_U(u)$ and $f_{V|Z}(v \mid z) = f_V(v)$ implies that

$$f_{U,V|Z}(u, v \mid z) = f_U(u) f_V(v).$$

Thus, U, V and Z are mutually independent.

On the other hand, the mutual independence immediately leads to the conditional independence $U \perp\!\!\!\perp V \mid Z$. Therefore, the conditional independence $X \perp\!\!\!\perp Y \mid Z$ is equivalent to the mutual independence of U, V and Z. We next show that the mutual independence of U, V and Z is equivalent to mutual independence of U, V and W.

Define $F_Z^{-1}(w) = \inf\{z : F(z) \geq w\}$ for $0 < w < 1$. If U, V and Z are mutually independent, then

$$\begin{aligned} & \Pr(U \leq u, V \leq v, W \leq w) = \Pr\{U \leq u, V \leq v, Z \leq F_Z^{-1}(w)\} \\ = \; & \Pr(U \leq u)\Pr(V \leq v)\Pr\{Z \leq F_Z^{-1}(w)\} = \Pr(U \leq u)\Pr(V \leq v)\Pr(W \leq w) \end{aligned}$$

holds for all u, v and w. On the other hand, if U, V and W are mutually independent, it follows that

$$\begin{aligned} & \Pr(U \leq u, V \leq v, Z \leq z) = \Pr\{U \leq u, V \leq v, W \leq F_Z(z)\} \\ = \; & \Pr(U \leq u)\Pr(V \leq v)\Pr\{W \leq F_Z(z)\} = \Pr(U \leq u)\Pr(V \leq v)\Pr(Z \leq z), \end{aligned}$$

holds for all u, v and z. Thus, the mutual independence of U, V and Z is equivalent to the mutual independence of U, V and W. This completes the proof. □

A.2 Proof of Theorem 2

We start with the derivation of the index ρ. U, V and W are mutually independent if and only if

$$\iiint \left\| \varphi_{U,V,W}(t_1, t_2, t_3) - \varphi_U(t_1)\varphi_V(t_2)\varphi_W(t_3) \right\|^2 \omega(t_1, t_2, t_3) dt_1 dt_2 dt_3 = 0, \quad (5)$$

for arbitrary positive weight function $\omega(\cdot)$. We now show that the proposed index ρ is proportional to the integration in (5) by choosing $\omega(t_1,t_2,t_3)$ to be the joint probability density function of three independent and identically distributed standard Cauchy random variables.

With some calculation and Fubini's theorem, we have

$$\begin{aligned}
&\iiint \|\varphi_{U,V,W}(t_1,t_2,t_3)-\varphi_U(t_1)\varphi_V(t_2)\varphi_W(t_3)\|^2\,\omega(t_1,t_2,t_3)dt_1dt_2dt_3\\
=\;& E\iiint e^{it_1(U_1-U_2)+it_2(V_1-V_2)+it_3(W_1-W_2)}\omega(t_1,t_2,t_3)dt_1dt_2dt_3\\
&-E\iiint e^{it_1(U_1-U_3)+it_2(V_1-V_4)+it_3(W_1-W_2)}\omega(t_1,t_2,t_3)dt_1dt_2dt_3\\
&-E\iiint e^{it_1(U_3-U_1)+it_2(V_4-V_1)+it_3(W_2-W_1)}\omega(t_1,t_2,t_3)dt_1dt_2dt_3\\
&+E\iiint e^{it_1(U_1-U_2)+it_2(V_3-V_4)+it_3(W_5-W_6)}\omega(t_1,t_2,t_3)dt_1dt_2dt_3.
\end{aligned}$$

According to the property of characteristic function for standard Cauchy distribution, we have

$$\int e^{it(U_1-U_2)}\pi^{-1}(1+t^2)^{-1}dt=e^{-|U_1-U_2|}.$$

Then by choosing $\omega(t_1,t_2,t_3)=\pi^{-3}(1+t_1^2)^{-1}(1+t_2^2)^{-1}(1+t_3^2)^{-1}$, i.e., the joint density function of three i.i.d. standard Cauchy distributions, we have

$$\begin{aligned}
&\iiint \|\varphi_{U,V,W}(t_1,t_2,t_3)-\varphi_U(t_1)\varphi_V(t_2)\varphi_W(t_3)\|^2\,\omega(t_1,t_2,t_3)dt_1dt_2dt_3\\
&=Ee^{-|U_1-U_2|-|V_1-V_2|-|W_1-W_2|}-2Ee^{-|U_1-U_3|-|V_1-V_4|-|W_1-W_2|}\\
&\qquad+Ee^{-|U_1-U_2|}Ee^{-|V_1-V_2|}Ee^{-|W_1-W_2|}.
\end{aligned}\tag{6}$$

Furthermore, with the fact that $U\perp\!\!\!\perp W$ and $V\perp\!\!\!\perp W$, (6) is equal to

$$E\left\{S_U(U_1,U_2)S_V(V_1,V_2)e^{-|W_1-W_2|}\right\},$$

where $S_U(U_1,U_2)$ and $S_V(V_1,V_2)$ are defined as

$$\begin{aligned}
S_U(U_1,U_2)&=E\left\{e^{-|U_1-U_2|}+e^{-|U_3-U_4|}-e^{-|U_1-U_3|}-e^{-|U_2-U_3|}\mid(U_1,U_2)\right\},\\
S_V(V_1,V_2)&=E\left\{e^{-|V_1-V_2|}+e^{-|V_3-V_4|}-e^{-|V_1-V_3|}-e^{-|V_2-V_3|}\mid(V_1,V_2)\right\}.
\end{aligned}$$

Now we calculate the normalization constant c_0. It follows by the Cauchy-Schwarz inequality that

$$\begin{aligned}
& E\left\{S_U(U_1,U_2)S_V(V_1,V_2)e^{-|W_1-W_2|}\right\} \\
= & E\left[e^{-|W_1-W_2|}E\left\{S_U(U_1,U_2)S_V(V_1,V_2)\mid(W_1,W_2)\right\}\right] \\
\leq & E\left[e^{-|W_1-W_2|}E^{1/2}\left\{S_U^2(U_1,U_2)\mid(W_1,W_2)\right\}E^{1/2}\left\{S_V^2(V_1,V_2)\mid(W_1,W_2)\right\}\right] \\
= & E\left[e^{-|W_1-W_2|}E^{1/2}\left\{S_U^2(U_1,U_2)\right\}E^{1/2}\left\{S_V^2(V_1,V_2)\right\}\right] \\
= & 2e^{-1}(6.5e^{-2}-20e^{-1}+6.5) \\
\stackrel{\text{def}}{=} & c_0^{-1},
\end{aligned}$$

where the equality holds if and only if $S_U(U_1,U_2)=\lambda\{S_V(V_1,V_2)\}$ holds with probability 1, where $\lambda\geq 0$ (because $E\{S_U(U_1,U_2)S_V(V_1,V_2)\}$ is nonnegative). Recall that U, V and W are all uniformly distributed on $(0,1)$, further calculations give us

$$\begin{aligned}
S_U(U_1,U_2) &= e^{-|U_1-U_2|}+e^{-U_1}+e^{U_1-1}+e^{-U_2}+e^{U_2-1}+2e^{-1}-4, \\
S_V(V_1,V_2) &= e^{-|V_1-V_2|}+e^{-V_1}+e^{V_1-1}+e^{-V_2}+e^{V_2-1}+2e^{-1}-4.
\end{aligned}$$

This, together with the normalization constant c_0, yield the expression of the index $\rho(X,Y\mid Z)$. Subsequently, the properties of the index $\rho(X,Y\mid Z)$ can be established.

(1) $\rho(X,Y\mid Z)\geq 0$ holds obviously. It equals 0 only when U, V, W are mutual independent, which is equivalent to the conditional independence $X\perp\!\!\!\perp Y\mid Z$. $\rho(X,Y\mid Z)\leq 1$ holds obviously according to the derivation of the index ρ. The equality holds if and only if $S_U(U_1,U_2)=\lambda\{S_V(V_1,V_2)\}$. Because $ES_U^2(U_1,U_2)=ES_V^2(V_1,V_2)$, we have $\lambda=1$. If $U=V$ or $U+V=1$, it is easy to check $S_U(U_1,U_2)=S_V(V_1,V_2)$. That is, if $F_{X|Z}(X\mid Z)=F_{Y|Z}(Y\mid Z)$ or $F_{X|Z}(X\mid Z)+F_{Y|Z}(Y\mid Z)=1$, then $\rho(X,Y\mid Z)=1$. To understand this condition better, we suppose $Y=m(X,Z)$. It is clear that $F_{X|Z}(X\mid Z)=F_{Y|Z}(Y\mid Z)$ when $m(\cdot,\cdot)$ is monotonically increasing on the first argument, whereas $F_{X|Z}(X\mid Z)+F_{Y|Z}(Y\mid Z)=1$ when $m(\cdot,\cdot)$ is monotonically decreasing on the first argument. Therefore, the index ρ is equal to one if Y is a strictly monotone transformation of X conditional on Z.

(2) This property is trivial according to the definition of ρ.

(3) For strictly monotone transformations $m_3(\cdot)$, we have when $m_1(\cdot)$ is strictly increasing, $U_m=F_{m_1(X)|m_3(Z)}\{m_1(X)\mid m_3(Z)\}$ equals $U=F_{X|Z}(X\mid Z)$, while when $m_1(\cdot)$ is strictly decreasing, it equals $1-U$. It can be easily verified that $S_U(U_1,U_2)=S_U(1-U_1,1-U_2)$, then we have $S_{U_m}(U_{m1},U_{m2})=S_U(U_1,U_2)$ no matter whether $m_1(\cdot)$ is strictly increasing or decreasing. Similarly, let $V_m=F_{m_2(Y)|m_3(Z)}\{m_2(Y)\mid m_3(Z)\}$, we obtain that $S_{V_m}(V_{m1},V_{m2})=S_V(V_1,V_2)$. It is clear that $W_m=F_{m_3(Z)}\{m_3(Z)\}$ equals either W or $1-W$, implying $e^{-|W_{m1}-W_{m2}|}=e^{-|W_1-W_2|}$. Therefore, we have

$$E\left\{S_{U_m}(U_{m1},U_{m2})S_{V_m}(V_{m1},V_{m2})e^{-|W_{m1}-W_{m2}|}\right\}=E\left\{S_U(U_1,U_2)S_V(V_1,V_2)e^{-|W_1-W_2|}\right\},$$

and it is true that $\rho\{m_1(X),m_2(Y)\mid m_3(Z)\}=\rho(X,Y\mid Z)$. □

A.3 Proof of Theorem 3

For simplicity, we denote by $g(x) = e^{-|x|}$ and $S_0(x, y) = g(x - y) + e^{-x} + e^{x-1} + e^{-y} + e^{y-1} + 2e^{-1} - 4$. We write $c_0^{-1}\widehat{\rho}(X, Y \mid Z)$ as

$$n^{-2}\sum_{i,j}\left\{S_0(\widehat{U}_i, \widehat{U}_j)S_0(\widehat{V}_i, \widehat{V}_j)g(\widehat{W}_i - \widehat{W}_j)\right\}.$$

With Taylor's expansion, when $nh^{4m} \to 0$ and $nh^2/\log^2(n) \to \infty$, we have

$$\begin{aligned}
& S_0(\widehat{U}_i, \widehat{U}_j) \\
= & \left\{g'(U_i - U_j) + e^{U_i-1} - e^{-U_i}\right\}\Delta U_i - \left\{g'(U_i - U_j) - e^{U_j-1} + e^{-U_j}\right\}\Delta U_j \\
+ & 2^{-1}\left\{g''(U_i - U_j)(\Delta U_i - \Delta U_j)^2 + \left(e^{-U_i} + e^{U_i-1}\right)(\Delta U_i)^2 + \left(e^{-U_j} + e^{U_j-1}\right)(\Delta U_j)^2\right\} \\
+ & 6^{-1}\left\{g'''(U_i - U_j)(\Delta U_i - \Delta U_j)^3 + \left(e^{U_i-1} - e^{-U_i}\right)(\Delta U_i)^3 + \left(e^{U_j-1} - e^{-U_j}\right)(\Delta U_j)^3\right\} \\
+ & S_0(U_i, U_j) + o_p(n^{-1}) \\
\stackrel{\text{def}}{=} & S_1(U_i, U_j) + S_2(U_i, U_j) + S_3(U_i, U_j) + S_0(U_i, U_j) + o_p(n^{-1}),
\end{aligned}$$

where $\Delta U_i = \widehat{U}_i - U_i$, and $S_k(U_i, U_j), k = 1, 2, 3$ are defined to be each row in an obvious way. Similarly, we expand $S_0(\widehat{V}_i, \widehat{V}_j)$ as

$$S_0(\widehat{V}_i, \widehat{V}_j) = S_0(V_i, V_j) + S_1(V_i, V_j) + S_2(V_i, V_j) + S_3(V_i, V_j) + o_p(n^{-1}).$$

As for $g(\widehat{W}_i - \widehat{W}_j)$, we have

$$\begin{aligned}
g(\widehat{W}_i - \widehat{W}_j) &= g(W_i - W_j) + g'(W_i - W_j)(\Delta W_i - \Delta W_j) \\
&\quad +2^{-1}g''(W_i - W_j)(\Delta W_i - \Delta W_j)^2 + o_p(n^{-1}).
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
c_0^{-1}\widehat{\rho}(X, Y \mid Z) &= 2^{-1}n^{-2}\sum_{i,j} S_0(U_i, U_j)\, S_0(V_i, V_j)\, g''(W_i - W_j)(\Delta W_i - \Delta W_j)^2 \\
&\quad +n^{-2}\sum_{i,j}\sum_{0\le k+l\le 1} S_k(U_i, U_j)\, S_l(V_i, V_j)\, g'(W_i - W_j)(\Delta W_i - \Delta W_j) \\
&\quad +n^{-2}\sum_{i,j}\sum_{0\le k+l\le 3} S_k(U_i, U_j)\, S_l(V_i, V_j)\, g(W_i - W_j) + o_p(n^{-1}) \\
&\stackrel{\text{def}}{=} 2^{-1}Q_1 + Q_2 + Q_3 + o_p(n^{-1}).
\end{aligned}$$

We first show that Q_1 is of order $o_p(n^{-1})$. In fact,

$$\begin{aligned}
Q_1 &= n^{-2}\sum_{i,j} S_0(U_i, U_j)\, S_0(V_i, V_j)\, g''(W_i - W_j)(\Delta W_i - \Delta W_j)^2 \\
&= n^{-2}\sum_{i,j} S_0(U_i, U_j)\, S_0(V_i, V_j)\, g''(W_i - W_j)\left\{(\Delta W_i)^2 + (\Delta W_j)^2\right\} \\
&\quad +2n^{-2}\sum_{i,j} S_0(U_i, U_j)\, S_0(V_i, V_j)\, g''(W_i - W_j)(W_i\Delta W_j + W_j\Delta W_i) \\
&\quad -2n^{-2}\sum_{i,j} S_0(U_i, U_j)\, S_0(V_i, V_j)\, g''(W_i - W_j)(\widehat{W}_i\widehat{W}_j - W_iW_j) \\
&\stackrel{\text{def}}{=} Q_{1,1} + Q_{1,2} + Q_{1,3}.
\end{aligned}$$

Under the null hypothesis, U, V and W are mutually independent, it is easy to verify that $E\left\{S_0\left(U_i,U_j\right) \mid (U_i,V_i,W_i,V_j,W_j)\right\}=0$, and hence

$$E\left\{S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)\mid(U_i,V_i,W_i)\right\}=0.$$

Then for each fixed i, we have

$$n^{-1}\sum_j S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)=O_p(n^{-1/2}).$$

Thus, $Q_{1,1}$ is clearly of order $o_p(n^{-1})$ because $(\Delta W_i)^2=o_p(n^{-1/2})$. Now we deal with $Q_{1,2}$.

$$\begin{aligned} & n^{-2}\textstyle\sum_{i,j}S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)W_i\Delta W_j \\ = {} & n^{-3}\textstyle\sum_{i,j,k}S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)W_i\left\{\mathbb{1}(W_k\le W_j)-W_j\right\}. \end{aligned}$$

Under the null hypothesis, because $E\left\{S_0\left(U_i,U_j\right)\mid(U_i,V_i,W_i,V_j,W_j)\right\}=0$, W is uniformly distributed, we have $E\{\mathbb{1}(W_k\le W_j)\mid W_j\}=W_j$. Thus, the corresponding U-statistic of the equation above is second order degenerate. In addition, when any two of i,j,k are identical, we have

$$E\left[S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)W_i\left\{\mathbb{1}(W_k\le W_j)-W_j\right\}\right]=0.$$

Then the summations associated with any two of the i,j,k are identical is of order $o_p(1)$. Therefore, $Q_{1,2}=o_p(n^{-1})$. It remains to deal with $Q_{1,3}$. Similarly, the corresponding U-statistic of

$$n^{-4}\sum_{i,j,k,l}S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g''(W_i-W_j)\left\{\mathbb{1}(W_k\le W_i)\mathbb{1}(W_l\le W_j)-W_iW_j\right\}$$

is second order degenerate and hence we obtain that $Q_{1,3}=o_p(n^{-1})$.

Next, we show $Q_2=o_p(n^{-1})$. Recall that

$$\begin{aligned} Q_2 &= n^{-2}\sum_{i,j}S_0\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g'(W_i-W_j)(\Delta W_i-\Delta W_j) \\ &\quad +2n^{-2}\sum_{i,j}S_1\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g'(W_i-W_j)\Delta W_i \\ &\quad +2n^{-2}\sum_{i,j}S_0\left(U_i,U_j\right)S_1\left(V_i,V_j\right)g'(W_i-W_j)\Delta W_i \\ &\stackrel{\text{def}}{=} Q_{2,1}+2Q_{2,2}+2Q_{2,3}. \end{aligned}$$

Similar to dealing with $Q_{1,2}$, we have $Q_{2,1}=o_p(n^{-1})$. We now evaluate $Q_{2,2}$.

$$\begin{aligned} Q_{2,2} &= n^{-2}\sum_{i,j}\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}\Delta U_iS_0\left(V_i,V_j\right)g'(W_i-W_j)\Delta W_i \\ &\quad -n^{-2}\sum_{i,j}\left\{g'\left(U_i-U_j\right)-e^{U_j-1}+e^{-U_j}\right\}\Delta U_jS_0\left(V_i,V_j\right)g'(W_i-W_j)\Delta W_i. \end{aligned}$$

Because under the null hypothesis, $E\left\{S_0\left(V_i, V_j\right) \mid \left(U_i, V_i, W_i, U_j, W_j\right)\right\} = 0$, it follows that for each i,

$$n^{-1} \sum_j \left\{g'\left(U_i - U_j\right) + e^{U_i - 1} - e^{-U_i}\right\} S_0\left(V_i, V_j\right) g'(W_i - W_j) = O_p(n^{-1/2}),$$

and the first term of $Q_{2,2}$ is of order $o_p(n^{-1})$ because $\Delta W_i = O_p(n^{-1/2})$ and $\Delta U_i = o_p(1)$. In addition, for each j,

$$n^{-2} \sum_{i,k} \left\{g'\left(U_i - U_j\right) - e^{U_j - 1} + e^{-U_j}\right\} S_0\left(V_i, V_j\right) g'(W_i - W_j) \left\{\mathbb{1}(W_k \leq W_i) - W_i\right\}$$

is degenerate and hence the second term of $Q_{2,2}$ is also of order $o_p(n^{-1})$ because $\Delta U_j = o_p(1)$, indicating $Q_{2,2} = o_p(n^{-1})$. Similarly, we have $Q_{2,3} = o_p(n^{-1})$. Thus it follows that $Q_2 = o_p(n^{-1})$.

Finally, we show that $Q_3 = n^{-2} \sum_{i,j} S_0\left(U_i, U_j\right) S_0\left(V_i, V_j\right) g\left(W_i - W_j\right) + o_p(n^{-1})$. Or equivalently, we show that $Q_{3,1}, Q_{3,2}$ and $Q_{3,3}$ are all of order $o_p(n^{-1})$, where

$$\begin{aligned}
Q_{3,1} &\stackrel{\text{def}}{=} n^{-2} \sum_{i,j} \sum_{k+l=1} S_k\left(U_i, U_j\right) S_l\left(V_i, V_j\right) g\left(W_i - W_j\right), \\
Q_{3,2} &\stackrel{\text{def}}{=} n^{-2} \sum_{i,j} \sum_{k+l=2} S_k\left(U_i, U_j\right) S_l\left(V_i, V_j\right) g\left(W_i - W_j\right), \\
Q_{3,3} &\stackrel{\text{def}}{=} n^{-2} \sum_{i,j} \sum_{k+l=3} S_k\left(U_i, U_j\right) S_l\left(V_i, V_j\right) g\left(W_i - W_j\right).
\end{aligned}$$

We first show that $Q_{3,1} \stackrel{\text{def}}{=} Q_{3,1,1} + Q_{3,1,2} = o_p(n^{-1})$, where

$$\begin{aligned}
Q_{3,1,1} &= n^{-2} \sum_{i,j} S_1\left(U_i, U_j\right) S_0\left(V_i, V_j\right) g\left(W_i - W_j\right), \\
Q_{3,1,2} &= n^{-2} \sum_{i,j} S_0\left(U_i, U_j\right) S_1\left(V_i, V_j\right) g\left(W_i - W_j\right).
\end{aligned}$$

Without loss of generality, we only show that $Q_{3,1,1} = o_p(n^{-1})$. Calculate

$$\begin{aligned}
S_1\left(U_i, U_j\right) &= \left\{g'\left(U_i - U_j\right) + e^{U_i - 1} - e^{-U_i}\right\} \Delta U_i - \left\{g'\left(U_i - U_j\right) - e^{U_j - 1} + e^{-U_j}\right\} \Delta U_j, \\
\Delta U_i &= n^{-1} \sum_{k=1}^{n} \left[\frac{K_h(Z_k - Z_i)\mathbb{1}(X_k \leq X_i)}{f(Z_i)} - U_i - \frac{U_i\left\{K_h(Z_k - Z_i) - f(Z_i)\right\}}{f(Z_i)}\right] \\
&\quad + O_p(h^{2m} + n^{-1}h^{-1}\log^2 n).
\end{aligned}$$

Thus, when $nh^{4m} \to 0$ and $nh^2/\log^2(n) \to \infty$,

$$
\begin{aligned}
Q_{3,1,1} &= 2n^{-2}\sum_{i,j}\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\Delta U_i \\
&= 2n^{-3}\sum_{i,j,k}\Bigg(\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&\quad \cdot\left[\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)}{f(Z_i)}-U_i-\frac{U_i\left\{K_h(Z_k-Z_i)-f(Z_i)\right\}}{f(Z_i)}\right]\Bigg)+o_p(n^{-1}) \\
&= \frac{2}{n(n-1)}\sum_{i\neq j}\Bigg(\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&\quad \cdot\left[\frac{E\left\{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)-U_iK_h(Z_k-Z_i)\mid(X_i,Z_i)\right\}}{f(Z_i)}\right]\Bigg)+o_p(n^{-1}),
\end{aligned}
$$

where the last equality holds due to equations (2)-(3) of section 5.3.4 in Serfling (2009) and the fact that $\mathrm{var}\{K_h(Z_i-Z_j)\}=O(h^{-1})$. Therefore, since

$$
\sup_{X_i,Z_i}\left|E\left\{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)-U_iK_h(Z_k-Z_i)\mid(X_i,Z_i)\right\}\right|=O(h^m),
$$

when the $(m-1)$th derivatives of $F_{X|Z}(x\mid z)f_Z(z)$ and $f_Z(z)$ with respect to z are locally Lipschitz-continuous, $Q_{3,1,1}$ is clearly of order $o_p(n^{-1})$ by noting that the summation in the last display is degenerate.

Next, we consider $Q_{3,2}$, where

$$
\begin{aligned}
Q_{3,2} &= n^{-2}\sum_{i,j}S_1\left(U_i,U_j\right)S_1\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&+ n^{-2}\sum_{i,j}S_2\left(U_i,U_j\right)S_0\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&+ n^{-2}\sum_{i,j}S_0\left(U_i,U_j\right)S_2\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&\overset{\text{def}}{=} Q_{3,2,1}+Q_{3,2,2}+Q_{3,2,3}.
\end{aligned}
$$

We first show that $Q_{3,2,1}=o_p(n^{-1})$. It follows that

$$
\begin{aligned}
Q_{3,2,1} &= 2n^{-2}\sum_{i,j}\Bigg[\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}\left\{g'\left(V_i-V_j\right)+e^{V_i-1}-e^{-V_i}\right\} \\
&\quad \cdot g\left(W_i-W_j\right)\Delta U_i\Delta V_i\Bigg]+2n^{-2}\sum_{i,j}\Bigg[\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\} \\
&\quad \cdot\left\{g'\left(V_i-V_j\right)-e^{V_j-1}+e^{-V_j}\right\}g\left(W_i-W_j\right)\Delta U_i\Delta V_j\Bigg] \\
&\overset{\text{def}}{=} Q_{3,2,1,1}+Q_{3,2,1,2}.
\end{aligned}
$$

Because $E\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\mid(U_i,V_i,W_i,V_j,W_j)\}=0$. For each i,

$$
n^{-1}\sum_{j=1}^{n}\left[\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\}\left\{g'\left(V_i-V_j\right)+e^{V_i-1}-e^{-V_i}\right\}g\left(W_i-W_j\right)\right]
$$

is of order $O_p(n^{-1/2})$. Then $Q_{3,2,1,1} = o_p(n^{-1})$ because $\Delta U_i \Delta V_i = o_p(n^{-1/2})$. For $Q_{3,2,1,2}$, $-\Delta U_i \Delta V_j = (U_i \Delta V_j + U_j \Delta V_i) + (\widehat{U}_i \widehat{V}_j - U_i V_j)$. By expanding the $\widehat{U}_i$, $\widehat{U}_j$, $\widehat{V}_i$, $\widehat{V}_j$ in $Q_{3,2,1,2}$ as U statistics and apply the same technique as showing $Q_{3,1,1} = o_p(n^{-1})$ and $Q_{1,3} = o_p(n^{-1})$, it follows immediately that $Q_{3,2,1,2}$ is of order $o_p(n^{-1})$. Thus $Q_{3,2,1}$ is of order $o_p(n^{-1})$.

For $Q_{3,2,2}$ and $Q_{3,2,3}$, we only show that $Q_{3,2,2}$ is of order $o_p(n^{-1})$, for simplicity. Recall that $S_2(U_i, U_j)$ is defined as

$$\begin{aligned} S_2(U_i, U_j) &= 2^{-1}\Big[g''(U_i - U_j)(\Delta U_i)^2 + g''(U_i - U_j)(\Delta U_j)^2 + \left(e^{-U_i} + e^{U_i - 1}\right)(\Delta U_i)^2 \\ &\qquad + \left(e^{-U_j} + e^{U_j - 1}\right)(\Delta U_j)^2 - 2g''(U_i - U_j)\Delta U_i \Delta U_j\Big]. \end{aligned}$$

The summations associated with either $(\Delta U_i)^2$ or $(\Delta U_j)^2$ are of order $o_p(n^{-1})$ following similar reasons as showing $Q_{3,2,1,1} = o_p(n^{-1})$, and that associated with $\Delta U_i \Delta U_j$ are of order $o_p(n^{-1})$ similar to dealing with $Q_{3,2,1,2}$. As a result, $Q_{3,2}$ is of order $o_p(n^{-1})$.

For $Q_{3,3}$, we have

$$Q_{3,3} = n^{-2} \sum_{i,j} \sum_{k=1}^{4} S_{4-k}(U_i, U_j) S_{k-1}(V_i, V_j) g(W_i - W_j) \stackrel{\text{def}}{=} \sum_{k=1}^{4} Q_{3,3,k}.$$

We only show that $Q_{3,3,1} = o_p(n^{-1})$ because the other terms are similar. Calculate

$$\begin{aligned} & 6S_3(U_i, U_j) \\ = \; & g'''(U_i - U_j)\left\{(\Delta U_i)^3 - 3(\Delta U_i)^2 \Delta U_j + 3\Delta U_i (\Delta U_j)^2 - (\Delta U_j)^3\right\} \\ & \qquad + \left(e^{U_i - 1} - e^{-U_i}\right)(\Delta U_i)^3 + \left(e^{U_j - 1} - e^{-U_j}\right)(\Delta U_j)^3. \end{aligned}$$

Then the summations associated with either $(\Delta U_i)^3$ or $(\Delta U_j)^3$ are of order $o_p(n^{-1})$ similar to dealing with $Q_{3,2,1,1}$, and that associated with $\Delta U_i(\Delta U_j)^2$ or $(\Delta U_i)^2 \Delta U_j$ are of order $o_p(n^{-1})$ similar to the second term of $Q_{2,2}$.

To sum up, we have shown that

$$c_0^{-1} \widehat{\rho}(X, Y \mid Z) = n^{-2} \sum_{i,j} S_0(U_i, U_j) S_0(V_i, V_j) g(W_i - W_j) + o_p(n^{-1}),$$

where the right hand side is essentially a first order degenerate V-statistics. Thus by applying Theorem 6.4.1.B of Serfling (2009),

$$n\widehat{\rho}(X, Y \mid Z) \xrightarrow{d} c_0 \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1)$$

where $\chi_j^2(1)$, $j = 1, 2, \ldots$ are independent $\chi^2(1)$ random variables, and λ_j, $j = 1, 2, \ldots$ are the corresponding eigenvalues of $h(u, v, w; u', v', w')$. It is worth mentioning that the kernel is positive definite and hence all the λ_js are positive. Therefore, the proof is completed. □

A.4 Proof of Theorem 4

Since we generate $\{U_i^*, V_i^*, W_i^*\}$, $i = 1, \ldots, n$ independently from uniform distribution, it is quite straightforward that U^*, V^* and W^* are mutually independent. In addition, we can write $\widehat{\rho}^*$ as

$$\widehat{\rho}^* = n^{-2} \sum_{i,j} c_0 S_0\left(U_i^*, U_j^*\right) S_0\left(V_i^*, V_j^*\right) g\left(W_i^* - W_j^*\right),$$

which clearly converges in distribution to $c_0 \sum_{j=1}^{\infty} \widetilde{\lambda}_j \chi_j^2(1)$, where $\chi_j^2(1)$, $j = 1, 2, \ldots$ are independent $\chi^2(1)$ random variables, and $\widetilde{\lambda}_j$, $j = 1, 2, \ldots$ are the eigenvalues of $h(u, v, w; u', v', w')$, implying $\widetilde{\lambda}_j = \lambda_j$, for $j = 1, 2, \ldots$, and hence the proof is completed.

□

A.5 Proof of Theorem 5

We use the same notation as the proof in Theorem 3. With Taylor's expansion, when $nh^{4m} \to 0$ and $nh^2/\log^2(n) \to \infty$, we have

$$\begin{aligned}
S_0(\widehat{U}_i, \widehat{U}_j) &= S_0(U_i, U_j) + S_1(U_i, U_j) + o_p(n^{-1/2}), \\
S_0(\widehat{V}_i, \widehat{V}_j) &= S_0(V_i, V_j) + S_1(V_i, V_j) + o_p(n^{-1/2}), \\
g(\widehat{W}_i - \widehat{W}_j) &= g(W_i - W_j) + g'(W_i - W_j)(\Delta W_i - \Delta W_j) + o_p(n^{-1/2}).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
c_0^{-1}\widehat{\rho}(X, Y \mid Z) &= n^{-2} \sum_{i,j} S_0\left(U_i, U_j\right) S_0\left(V_i, V_j\right) g\left(W_i - W_j\right) \\
&\quad + n^{-2} \sum_{i,j} S_0\left(U_i, U_j\right) S_0\left(V_i, V_j\right) g'(W_i - W_j)(\Delta W_i - \Delta W_j) \\
&\quad + n^{-2} \sum_{i,j} S_1\left(U_i, U_j\right) S_0\left(V_i, V_j\right) g(W_i - W_j) \\
&\quad + n^{-2} \sum_{i,j} S_0\left(U_i, U_j\right) S_1\left(V_i, V_j\right) g(W_i - W_j) + o_p(n^{-1/2}) \\
&\overset{\text{def}}{=} P_1 + P_2 + P_3 + P_4 + o_p(n^{-1/2}).
\end{aligned}$$

We deal with the four terms, respectively. For P_1, by applying Lemma 5.7.3 and equation (2) in section 5.3.1 of Serfling (2009), we have

$$\begin{aligned}
P_1 - c_0^{-1}\rho(X, Y \mid Z) &= 2n^{-1} \sum_{i=1}^{n} E\left[\left\{S_0\left(U_i, U\right) S_0\left(V_i, V\right) g\left(W_i - W\right)\right\} \mid (U_i, V_i, W_i)\right] \\
&\qquad\qquad - 2c_0^{-1}\rho(X, Y \mid Z) + o_p(n^{-1/2}) \\
&\overset{\text{def}}{=} 2n^{-1} \sum_{i=1}^{n} \left\{P_{1,i} - c_0^{-1}\rho(X, Y \mid Z)\right\} + o_p(n^{-1/2}).
\end{aligned} \tag{7}$$

Next, we deal with P_2. Recall that

$$\begin{aligned} P_2 &= n^{-2}\sum_{i,j} S_0\left(U_i,U_j\right) S_0\left(V_i,V_j\right) g'(W_i-W_j)(\Delta W_i-\Delta W_j) \\ &= 2n^{-3}\sum_{i,j,k} S_0\left(U_i,U_j\right) S_0\left(V_i,V_j\right) g'(W_i-W_j)\left\{I(W_k\le W_i)-W_i\right\}. \end{aligned}$$

By applying Lemma 5.7.3 and equation (2) in section 5.3.1 of Serfling (2009) again, we can obtain that

$$\begin{aligned} P_2 &= 2n^{-1}\sum_{i=1}^{n} E\left[S_0(U,U')S_0(V,V')g'(W-W')\left\{I(W_i\le W)-W\right\}\mid W_i\right]+o_p(n^{-1/2}) \\ &\stackrel{\text{def}}{=} 2n^{-1}\sum_{i=1}^{n} P_{2,i}+o_p(n^{-1/2}), \end{aligned} \tag{8}$$

where (U',V',W') is an independent copy of (U,V,W).

It remains to deal with P_3 and P_4. P_3 equals

$$\begin{aligned} P_3 &= 2n^{-3}\sum_{i,j,k}\Bigg[\left\{g'\left(U_i-U_j\right)+e^{U_i-1}-e^{-U_i}\right\} S_0\left(V_i,V_j\right) g(W_i-W_j) \\ &\qquad \cdot\left\{\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)-U_iK_h(Z_k-Z_i)}{f(Z_i)}\right\}\Bigg]+o_p(n^{-1/2}). \end{aligned}$$

By definition, we have $V \perp\!\!\!\perp W$ and hence it can be verified that

$$E\left\{S_0\left(V_i,V_j\right) g(W_i-W_j)\mid V_i,W_i\right\}=0.$$

Denote $P_3^{k,i}=\{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)-U_iK_h(Z_k-Z_i)\}/f_Z(Z_i)$. Thus,

$$\begin{aligned} & E\left\{\left(e^{U_i-1}-e^{-U_i}\right) S_0\left(V_i,V_j\right) g(W_i-W_j)P_3^{k,i}\mid X_k,Z_k\right\} \\ =\ & E\left\{\left(e^{U_i-1}-e^{-U_i}\right) S_0\left(V_i,V_j\right) g(W_i-W_j)P_3^{k,i}\mid X_k,Z_k,U_i\right\}=0. \end{aligned}$$

Thus when $nh^{2m}\to 0$ and $nh\to\infty$, we have

$$\begin{aligned} P_3 &= 2n^{-1}\sum_{k=1}^{n} E\left[\left\{I(X\ge X_k)-U\right\} g'(U-U')S_0(V,V')\right. \\ &\qquad \left.\cdot g(W-W')\mid X_k,Z_k\right]+o_p(n^{-1/2}) \\ &\stackrel{\text{def}}{=} 2n^{-1}\sum_{i=1}^{n} P_{3,i}+o_p(n^{-1/2}). \end{aligned} \tag{9}$$

Following similar arguments, we can show that

$$\begin{aligned} P_4 &= 2n^{-1}\sum_{i=1}^{n} E\left[\left\{I(Y\ge Y_i)-V\right\} g'(V-V')S_0(U,U')\right. \\ &\qquad \left.\cdot g(W-W')\mid Z=Z_i\right]+o_p(n^{-1/2}) \\ &\stackrel{\text{def}}{=} 2n^{-1}\sum_{i=1}^{n} P_{4,i}+o_p(n^{-1/2}). \end{aligned} \tag{10}$$

To sum up, it is shown that $c_0^{-1}\widehat{\rho}(X, Y \mid Z)$ could be written as

$$\begin{aligned} & c_0^{-1}\widehat{\rho}(X, Y \mid Z) - c_0^{-1}\rho(X, Y \mid Z) \\ = \quad & 2n^{-1}\sum_{i=1}^{n}\left\{P_{1,i} + P_{2,i} + P_{3,i} + P_{4,i} - c_0^{-1}\rho(X, Y \mid Z)\right\} + o_p(n^{-1/2}), \end{aligned}$$

where $P_{1,i}, P_{2,i}, P_{3,i}$ and $P_{4,i}$ are defined in (7)-(10), respectively. Thus the asymptotic normality follows.

Under the local alternative, we have $U = F(X \mid Y, Z) + n^{-1/2}\ell(X, Y, Z)$, and it is easy to verify that $\breve{U} \stackrel{\text{def}}{=} F(X \mid Y, Z)$, V and W are mutually independent. With Taylor's expansion, we have

$$\begin{aligned} & S_0(\widehat{U}_i, \widehat{U}_j) \\ = \quad & \left\{g'(\breve{U}_i - \breve{U}_j) + e^{\breve{U}_i - 1} - e^{-\breve{U}_i}\right\}\Delta\breve{U}_i - \left\{g'(\breve{U}_i - \breve{U}_j) - e^{\breve{U}_j - 1} + e^{-\breve{U}_j}\right\}\Delta\breve{U}_j \\ + \quad & 2^{-1}\left\{g''(\breve{U}_i - \breve{U}_j)(\Delta\breve{U}_i - \Delta\breve{U}_j)^2 + \left(e^{-\breve{U}_i} + e^{\breve{U}_i - 1}\right)(\Delta\breve{U}_i)^2 + \left(e^{-\breve{U}_j} + e^{\breve{U}_j - 1}\right)(\Delta\breve{U}_j)^2\right\} \\ + \quad & 6^{-1}\left\{g'''(\breve{U}_i - \breve{U}_j)(\Delta\breve{U}_i - \Delta\breve{U}_j)^3 + \left(e^{\breve{U}_i - 1} - e^{-\breve{U}_i}\right)(\Delta\breve{U}_i)^3 + \left(e^{\breve{U}_j - 1} - e^{-\breve{U}_j}\right)(\Delta\breve{U}_j)^3\right\} \\ + \quad & S_0(\breve{U}_i, \breve{U}_j) + o_p(n^{-1}) \\ \stackrel{\text{def}}{=} \quad & S_1(\breve{U}_i, \breve{U}_j) + S_2(\breve{U}_i, \breve{U}_j) + S_3(\breve{U}_i, \breve{U}_j) + S_0(\breve{U}_i, \breve{U}_j) + o_p(n^{-1}), \end{aligned}$$

where $\Delta\breve{U}_i = \widehat{U}_i - \breve{U}_i$. Then we can write $c_0^{-1}\widehat{\rho}(X, Y \mid Z)$ as

$$\begin{aligned} & c_0^{-1}\widehat{\rho}(X, Y \mid Z) \\ = \quad & 2^{-1}n^{-2}\sum_{i,j} S_0(\breve{U}_i, \breve{U}_j)S_0\left(V_i, V_j\right)g''(W_i - W_j)(\Delta W_i - \Delta W_j)^2 \\ & + n^{-2}\sum_{i,j}\sum_{0 \le k+l \le 1} S_k(\breve{U}_i, \breve{U}_j)S_l\left(V_i, V_j\right)g'(W_i - W_j)(\Delta W_i - \Delta W_j) \\ & + n^{-2}\sum_{i,j}\sum_{0 \le k+l \le 3} S_k(\breve{U}_i, \breve{U}_j)S_l\left(V_i, V_j\right)g\left(W_i - W_j\right) + o_p(n^{-1}) \\ \stackrel{\text{def}}{=} \quad & 2^{-1}\widetilde{Q}_1 + \widetilde{Q}_2 + \widetilde{Q}_3 + o_p(n^{-1}). \end{aligned}$$

With the same arguments as that in deriving $Q_1 = o_p(n^{-1})$ in the proof of Theorem 3, we have $\widetilde{Q}_1 = o_p(n^{-1})$.

Now we deal with $\widetilde{Q}_2$. For ease of notation, we write $\ell(X_i, Y_i, Z_i)$ as ℓ_i in the remaining proof. By decomposing $\Delta\breve{U}_i$ as $\Delta\breve{U}_i = \Delta U_i + n^{-1/2}\ell_i$, we have

$$\begin{aligned} \widetilde{Q}_2 \quad = \quad & n^{-2}\sum_{i,j} S_1(\breve{U}_i, \breve{U}_j)S_0\left(V_i, V_j\right)g'(W_i - W_j)(\Delta W_i - \Delta W_j) + o_p(n^{-1}) \\ = \quad & 2n^{-2}\sum_{i,j}\left\{g'(\breve{U}_i - \breve{U}_j) + e^{\breve{U}_i - 1} - e^{-\breve{U}_i}\right\}\Delta\breve{U}_i S_0\left(V_i, V_j\right)g'(W_i - W_j)\Delta W_i \\ & - 2n^{-2}\sum_{i,j}\left\{g'(\breve{U}_i - \breve{U}_j) + e^{\breve{U}_i - 1} - e^{-\breve{U}_i}\right\}\Delta\breve{U}_i \\ & \qquad \cdot S_0\left(V_i, V_j\right)g'(W_i - W_j)\Delta W_j + o_p(n^{-1}) \\ \stackrel{\text{def}}{=} \quad & 2\widetilde{Q}_{2,1} - 2\widetilde{Q}_{2,2} + o_p(n^{-1}). \end{aligned}$$

$\widetilde{Q}_{2,1}$ is clearly of order $o_p(n^{-1})$ because for each fixed i,

$$E\left[\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}S_0\left(V_i,V_j\right)g'(W_i-W_j)\;\Big|\;(\breve{U}_i,V_i,W_i)\right]=0.$$

$\widetilde{Q}_{2,2}$ is also of order $o_p(n^{-1})$ because for each i,

$$n^{-2}\sum_{j,k}\left[\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}S_0\left(V_i,V_j\right)g'(W_i-W_j)\left\{\mathbb{1}(W_k\le W_j)-W_j\right\}\right]$$

is degenerate.

Then we deal with the last quantity, $\widetilde{Q}_3$, where

$$\begin{aligned}
\widetilde{Q}_3 &= n^{-2}\sum_{i,j}S_0(\breve{U}_i,\breve{U}_j)S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
&\quad +n^{-2}\sum_{i,j}\sum_{k+l=1}S_k(\breve{U}_i,\breve{U}_j)S_l\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
&\quad +n^{-2}\sum_{i,j}\sum_{k+l=2}S_k(\breve{U}_i,\breve{U}_j)S_l\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
&\quad +n^{-2}\sum_{i,j}\sum_{k+l=3}S_k(\breve{U}_i,\breve{U}_j)S_l\left(V_i,V_j\right)g\left(W_i-W_j\right)+o_p(n^{-1})\\
&\stackrel{\text{def}}{=} \widetilde{Q}_{3,0}+\widetilde{Q}_{3,1}+\widetilde{Q}_{3,2}+\widetilde{Q}_{3,3}+o_p(n^{-1}).
\end{aligned}$$

We simplify $\widetilde{Q}_{3,1}$ first. According to the proof of Theorem 3, we have

$$\begin{aligned}
\widetilde{Q}_{3,1} &= n^{-2}\sum_{i,j}S_1(\breve{U}_i,\breve{U}_j)S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)+o_p(n^{-1})\\
&= 2n^{-5/2}\sum_{i,j}\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}\ell_i S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
&\quad +2n^{-3}\sum_{i,j,k}\left[\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\right.\\
&\qquad\qquad \left.\cdot\left\{\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)}{f(Z_i)}-\frac{U_iK_h(Z_k-Z_i)}{f(Z_i)}\right\}\right]+o_p(n^{-1})\\
&\stackrel{\text{def}}{=} \widetilde{Q}_{3,1,1}+2\widetilde{Q}_{3,1,2}+o_p(n^{-1}).
\end{aligned}$$

As we can see, $\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)}{f(Z_i)}-\frac{U_iK_h(Z_k-Z_i)}{f(Z_i)}$ is of order h^m. Then we can derive that

$$\begin{aligned}
\widetilde{Q}_{3,1,2} &= n^{-1}\sum_{j=1}^{n}E\left[\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\right.\\
&\quad \left.\cdot\left\{\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)}{f(Z_i)}-\frac{U_iK_h(Z_k-Z_i)}{f(Z_i)}\right\}\;\Big|\;(X_j,Y_j,Z_j)\right]+O_p(n^{-1}h^m).
\end{aligned}$$

It can be verified that

$$\begin{aligned}
& E\left\{\frac{K_h(Z_k-Z_i)\mathbb{1}(X_k\le X_i)}{f(Z_i)}-\frac{U_iK_h(Z_k-Z_i)}{f(Z_i)}\;\Big|\;(X_i,Z_i)\right\}\\
= & E\left\{\frac{K_h(Z_k-Z_i)F(X_i\mid Z_k)}{f(Z_i)}-\frac{U_iK_h(Z_k-Z_i)}{f(Z_i)}\;\Big|\;(X_i,Z_i)\right\}\\
= & f^{-1}(Z_i)\int K(u)\left\{F(X_i\mid Y_i,Z_i+uh)+n^{-1/2}\ell(X_i,Y_i,Z_i+uh)\right\}f(uh+Z_i)du\\
& -f^{-1}(Z_i)\breve{U}_iE\left\{K_h(Z_k-Z_i)\mid Z_i\right\}-n^{-1/2}f^{-1}(Z_i)\ell_iE\left\{K_h(Z_k-Z_i)\mid Z_i\right\}.
\end{aligned}$$

And

$$f^{-1}(Z_i)\int K(u)F(X_i\mid Y_i,Z_i+uh)f(uh+Z_i)du-f^{-1}(Z_i)\breve{U}_iE\left\{K_h(Z_k-Z_i)\mid Z_i\right\}$$

is of order h^m and is only a function of $(\breve{U}_i,Z_i)$, which is independent of V_i. Substituting this into $\widetilde{Q}_{3,1,2}$, we have

$$\begin{aligned}
\widetilde{Q}_{3,1,2} = & n^{-3/2}\sum_{j=1}^{n}E\Bigg(\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
& \cdot\left[\frac{\int K(u)\left\{\ell(X_i,Y_i,Z_i+uh)-\ell_i\right\}f(Z_i+uh)du}{f(Z_i)}\right]\;\Big|\;(X_j,Y_j,Z_j)\Bigg)+O_p(n^{-1}h^m).
\end{aligned}$$

Then $\widetilde{Q}_{3,1,2}$ is clearly of order $o_p(n^{-1})$ by noting that the conditional expectation of the above display is of order h^m while the unconditional expectation is zero.

Next, we deal with $\widetilde{Q}_{3,2}$. It is straightforward that

$$\begin{aligned}
\widetilde{Q}_{3,2} = & n^{-2}\sum_{i,j}S_1(\breve{U}_i,\breve{U}_j)S_1\left(V_i,V_j\right)g\left(W_i-W_j\right)\\
& +n^{-2}\sum_{i,j}S_2(\breve{U}_i,\breve{U}_j)S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)+o_p(n^{-1})\\
\stackrel{\text{def}}{=} & \widetilde{Q}_{3,2,1}+\widetilde{Q}_{3,2,2}+o_p(n^{-1}).
\end{aligned}$$

Similar to dealing with $\widetilde{Q}_{3,1,2}$, we can show that

$$\widetilde{Q}_{3,2,1} = 2n^{-5/2}\sum_{i,j}\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}\ell_iS_1\left(V_i,V_j\right)g\left(W_i-W_j\right)+o_p(n^{-1}).$$

Then $\widetilde{Q}_{3,2,1}$ is of order $o_p(n^{-1})$ because $S_1\left(V_i,V_j\right)=o_p(1)$ and

$$\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}\ell_iS_1\left(V_i,V_j\right)g\left(W_i-W_j\right)$$

is also $o_p(1)$ with the expectation being zero. Similar as before, we can show that

$$\begin{aligned}
\widetilde{Q}_{3,2,2} &= 2^{-1}n^{-3}\sum_{i,j}\Bigg[\Bigg\{g''(\breve{U}_i-\breve{U}_j)(\ell_i-\ell_j)^2+\left(e^{-\breve{U}_i}+e^{\breve{U}_i-1}\right)\ell_i^2 \\
&\qquad +\left(e^{-\breve{U}_j}+e^{\breve{U}_j-1}\right)\ell_j^2\Bigg\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)\Bigg]+o_p(n^{-1}) \\
&= 2^{-1}n^{-1}E\Bigg[\left\{g''(\breve{U}_1-\breve{U}_2)(\ell_1-\ell_2)^2+\left(e^{-\breve{U}_1}+e^{\breve{U}_1-1}\right)\ell_1^2+\left(e^{-\breve{U}_2}+e^{\breve{U}_2-1}\right)\ell_2^2\right\} \\
&\qquad \cdot S_0\left(V_1,V_2\right)g\left(W_1-W_2\right)\Bigg]+o_p(n^{-1}) \\
&= -n^{-1}E\left\{g''(\breve{U}_1-\breve{U}_2)\ell_1\ell_2S_0\left(V_1,V_2\right)g\left(W_1-W_2\right)\right\}.
\end{aligned}$$

Now we show that $\widetilde{Q}_{3,3}=o_p(n^{-1})$. Because $(\Delta U_i)^2=o_p(n^{-1/2})$, we have

$$\begin{aligned}
\widetilde{Q}_{3,3} &= 2n^{-2}\sum_{i,j}\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}\Delta U_iS_2\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&\quad +2^{-1}n^{-2}\sum_{i,j}\Bigg\{g''(\breve{U}_i-\breve{U}_j)(\Delta U_i-\Delta U_j)^2+\left(e^{-\breve{U}_i}+e^{\breve{U}_i-1}\right)(\Delta U_i)^2 \\
&\qquad +\left(e^{-\breve{U}_j}+e^{\breve{U}_j-1}\right)(\Delta U_j)^2\Bigg\}S_1\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
&\quad +6^{-1}n^{-2}\sum_{i,j}\Bigg\{g'''(\breve{U}_i-\breve{U}_j)(\Delta U_i-\Delta U_j)^3+\left(e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right)(\Delta U_i)^3 \\
&\qquad +\left(e^{\breve{U}_j-1}-e^{-\breve{U}_j}\right)(\Delta U_j)^3\Bigg\}S_0\left(V_i,V_j\right)g\left(W_i-W_j\right)+o_p(n^{-1/2}).
\end{aligned}$$

Similar to dealing with $\widetilde{Q}_{3,1,2}$, we can obtain that $\widetilde{Q}_{3,3}=o_p(n^{-1})$.

Combining these results together, we have

$$\begin{aligned}
& c_0^{-1}\widehat{\rho}(X,Y\mid Z) \\
=\; & n^{-2}\sum_{i,j}S_0(\breve{U}_i,\breve{U}_j)S_0\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
& +2n^{-5/2}\sum_{i,j}\left\{g'(\breve{U}_i-\breve{U}_j)+e^{\breve{U}_i-1}-e^{-\breve{U}_i}\right\}\ell_iS_0\left(V_i,V_j\right)g\left(W_i-W_j\right) \\
& -n^{-1}E\left\{g''(\breve{U}_1-\breve{U}_2)\ell_1\ell_2S_0\left(V_1,V_2\right)g\left(W_1-W_2\right)\right\}+o_p(n^{-1}).
\end{aligned}$$

Then we can verify that $c_0^{-1}\widehat{\rho}(X,Y\mid Z)$ can be written as

$$\begin{aligned}
c_0^{-1}\widehat{\rho}(X,Y\mid Z) &= \iiint\Bigg\|n^{-1}\sum_{j=1}^{n}\Bigg[\left\{e^{it_1\breve{U}_j}-\varphi_{\breve{U}}(t_1)\right\}\left\{e^{it_2V_j}-\varphi_V(t_2)\right\}e^{it_3W_j} \\
&\qquad +it_1n^{-1/2}\ell_je^{it_1\breve{U}_j}\left\{e^{it_2V_j}-\varphi_V(s)\right\}e^{it_3W_j}\Bigg]\Bigg\|^2\omega(t_1,t_2,t_3)dt_1dt_2dt_3+o_p(n^{-1}).
\end{aligned}$$

It is clear that the empirical process

$$n^{-1/2}\sum_{j=1}^{n}\left[\left\{e^{it_1\breve{U}_j}-\varphi_{\breve{U}}(t_1)\right\}\left\{e^{it_2V_j}-\varphi_V(t_2)\right\}e^{it_3W_j}+it_1\ell_je^{it_1\breve{U}_j}\left\{e^{it_2V_j}-\varphi_V(t_2)\right\}e^{it_3W_j}\right]$$

converges in distribution to a complex valued gaussian process $\zeta(t_1,t_2,t_3)$ with mean function

$$E\left[it_1\ell(X,Y,Z)e^{it_1\breve{U}}\left\{e^{it_2V}-\varphi_V(t_2)\right\}e^{it_3W}\right],$$

and covariance function $\text{cov}\{\zeta(t_1,t_2,t_3),\overline{\zeta(t_{10},t_{20},t_{30})}\}$ given by

$$\{\varphi_U(t_1-t_{10})-\varphi_U(t_1)\varphi_U(-t_{10})\}\{\varphi_V(t_2-t_{20})-\varphi_V(t_2)\varphi_V(-t_{20})\}\varphi_W(t_3-t_{30}). \quad (11)$$

Therefore, by employing empirical process technology, we can derive that

$$c_0^{-1}n\widehat{\rho}(X,Y\mid Z)\xrightarrow{d}\iiint\|\zeta(t_1,t_2,t_3)\|^2\,\omega(t_1,t_2,t_3)dt_1dt_2dt_3.$$

Hence we conclude the proof for local alternatives.

□

A.6 Proof of Theorem 6

Firstly, $\mathbf{x}\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z}$ is equivalent to $(X_1,\ldots,X_p)\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z}$. According to Proposition 4.6 of Cook (2009), it is also equivalent to

$$X_1\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z},\quad X_2\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},X_1),\quad\ldots,\quad X_p\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},X_1,\ldots,X_{p-1}).$$

Following similar arguments for proving the equivalence between $X\perp\!\!\!\perp Y\mid Z$ and $U\perp\!\!\!\perp V\mid Z$ in the proof of Proposition 1, the above conditional independence series are equivalent to

$$\widetilde{U}_1\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z},\quad \widetilde{U}_2\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},X_1),\quad\ldots,\quad \widetilde{U}_p\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},X_1,\ldots,X_{p-1}).$$

According to the proof of Proposition 1, we know that $\widetilde{U}_k\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},X_1,\ldots,X_{k-1})$ is equivalent to $\widetilde{U}_k\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},\widetilde{U}_1,\ldots,\widetilde{U}_{k-1})$ for $k=1,\ldots,p-1$. Hence the conditional independence series hold if and only if

$$\widetilde{U}_1\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z},\quad \widetilde{U}_2\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},\widetilde{U}_1),\quad\ldots,\quad \widetilde{U}_p\perp\!\!\!\perp\mathbf{y}\mid(\mathbf{z},\widetilde{U}_1,\ldots,\widetilde{U}_{p-1}).$$

Then by applying Proposition 4.6 of Cook (2009) again, we know that $(X_1,\ldots,X_p)\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z}$ is equivalent to $\mathbf{u}\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z}$. Furthermore, with the same arguments for dealing with $\mathbf{y}$, we can obtain that it is additionally equivalent to $\mathbf{u}\perp\!\!\!\perp\mathbf{v}\mid\mathbf{z}$. Besides, with the fact that $\mathbf{u}\perp\!\!\!\perp\mathbf{z}$ and $\mathbf{v}\perp\!\!\!\perp\mathbf{z}$, we can get the conditional independence $\mathbf{x}\perp\!\!\!\perp\mathbf{y}\mid\mathbf{z}$ is equivalent to the mutual independence of $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$. Therefore, the proof is completed by following similar arguments with the proof of Proposition 1. □

A.7 Proof of Theorem 7

Following the proof of Theorem 3, we denote by $\widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) = e^{-\|\mathbf{u}_1 - \mathbf{u}_2\|_1}$, $\widetilde{g}(\mathbf{v}_1, \mathbf{v}_2) = e^{-\|\mathbf{v}_1 - \mathbf{v}_2\|_1}$ and $\widetilde{g}(\mathbf{w}_1, \mathbf{w}_2) = e^{-\|\mathbf{w}_1 - \mathbf{w}_2\|_1}$. Then we have

$$\begin{aligned} S_{\mathbf{u}}(\mathbf{u}_1, \mathbf{u}_2) &= E\left\{\widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) + \widetilde{g}(\mathbf{u}_3, \mathbf{u}_4) - \widetilde{g}(\mathbf{u}_1, \mathbf{u}_3) - \widetilde{g}(\mathbf{u}_2, \mathbf{u}_3) \mid (\mathbf{u}_1, \mathbf{u}_2)\right\}, \\ S_{\mathbf{v}}(\mathbf{v}_1, \mathbf{v}_2) &= E\left\{\widetilde{g}(\mathbf{v}_1, \mathbf{v}_2) + \widetilde{g}(\mathbf{v}_3, \mathbf{v}_4) - \widetilde{g}(\mathbf{v}_1, \mathbf{v}_3) - \widetilde{g}(\mathbf{v}_2, \mathbf{v}_3) \mid (\mathbf{v}_1, \mathbf{v}_2)\right\}. \end{aligned}$$

Therefore, $\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ can be written as

$$\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = n^{-2} \sum_{i,j} \left\{S_{\mathbf{u}}(\widehat{\mathbf{u}}_i, \widehat{\mathbf{u}}_j) S_{\mathbf{v}}(\widehat{\mathbf{v}}_i, \widehat{\mathbf{v}}_j) \widetilde{g}(\widehat{\mathbf{w}}_i, \widehat{\mathbf{w}}_j)\right\}.$$

With Taylor's expansion, when $nh^{4m} \to 0$, $nh^{2(r+p-1)}/\log^2(n) \to \infty$, under conditions 2′ and 3′, we have

$$\widetilde{g}(\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2) = \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) + \sum_{k=1}^{3} (\Delta\mathbf{u}_1^{\mathrm{T}}, \Delta\mathbf{u}_2^{\mathrm{T}})^{\otimes k} D^{\otimes k} \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) + o_p(n^{-1}),$$

where $\mathbf{A}^{\otimes k}$ denotes the k-th Kronecker power of the matrix $\mathbf{A}$, $\Delta\mathbf{u}_i = \widehat{\mathbf{u}}_i - \mathbf{u}_i$ and

$$D^{\otimes k} \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) = \frac{\partial^k \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2)}{\{\partial(\mathbf{u}_1^{\mathrm{T}}, \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}}\}^{\otimes k}}.$$

In addition, we can expand $\widetilde{g}(\widehat{\mathbf{u}}_1, \mathbf{u}_2)$ as

$$\widetilde{g}(\widehat{\mathbf{u}}_1, \mathbf{u}_2) = \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) + \sum_{k=1}^{3} (k!)^{-1} (\Delta\mathbf{u}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\otimes k} D^{\otimes k} \widetilde{g}(\mathbf{u}_1, \mathbf{u}_2) + o_p(n^{-1}).$$

Therefore, by the definition of $S_{\mathbf{u}}(\mathbf{u}_1, \mathbf{u}_2)$, we have

$$\begin{aligned} & S_{\mathbf{u}}(\widehat{\mathbf{u}}_i, \widehat{\mathbf{u}}_j) \\ = & E\left\{\widetilde{g}(\mathbf{u}_i, \mathbf{u}_j) + \widetilde{g}(\mathbf{u}, \mathbf{u}') - \widetilde{g}(\mathbf{u}_i, \mathbf{u}) - \widetilde{g}(\mathbf{u}, \mathbf{u}_j) \mid (\mathbf{u}_i, \mathbf{u}_j)\right\} \\ + & \sum_{k=1}^{3} (k!)^{-1} E\Big\{(\Delta\mathbf{u}_i^{\mathrm{T}}, \Delta\mathbf{u}_j^{\mathrm{T}})^{\otimes k} D^{\otimes k} \widetilde{g}(\mathbf{u}_i, \mathbf{u}_j) - (\Delta\mathbf{u}_i^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\otimes k} D^{\otimes k} \widetilde{g}(\mathbf{u}_i, \mathbf{u}) \\ & \qquad - (\mathbf{0}^{\mathrm{T}}, \Delta\mathbf{u}_j^{\mathrm{T}})^{\otimes k} D^{\otimes k} \widetilde{g}(\mathbf{u}, \mathbf{u}_j) \mid \mathbf{u}_i, \mathbf{u}_j\Big\} + o_p(n^{-1}), \\ \stackrel{\text{def}}{=} & \sum_{k=0}^{3} \widetilde{S}_k(\mathbf{u}_i, \mathbf{u}_j) + o_p(n^{-1}), \end{aligned}$$

where $\widetilde{S}_0(\mathbf{u}_i, \mathbf{u}_j) = S_{\mathbf{u}}(\mathbf{u}_i, \mathbf{u}_j)$ and $\widetilde{S}_k(\mathbf{u}_i, \mathbf{u}_j), k = 1, 2, 3$ are defined obviously. Similarly, when $nh^{2(r+q-1)}/\log^2(n) \to \infty$, and $nh^{4m} \to 0$, we can expand $S_{\mathbf{v}}(\widehat{\mathbf{v}}_i, \widehat{\mathbf{v}}_j)$ as

$$S_{\mathbf{v}}(\widehat{\mathbf{v}}_i, \widehat{\mathbf{v}}_j) \stackrel{\text{def}}{=} \sum_{k=0}^{3} \widetilde{S}_k(\mathbf{v}_i, \mathbf{v}_j) + o_p(n^{-1}),$$

and it follows that

$$\begin{aligned}
\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) &= n^{-2} \sum_{i,j} \sum_{0 \le k+l \le 3} \widetilde{S}_k(\mathbf{u}_i, \mathbf{u}_j) \widetilde{S}_l(\mathbf{v}_i, \mathbf{v}_j) \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) \\
&\quad + n^{-2} \sum_{i,j} \sum_{0 \le k+l \le 2} \widetilde{S}_k(\mathbf{u}_i, \mathbf{u}_j) \widetilde{S}_l(\mathbf{v}_i, \mathbf{v}_j) (\Delta \mathbf{w}_i^{\mathrm{T}}, \Delta \mathbf{w}_j^{\mathrm{T}}) D \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) \\
&\quad + 2^{-1} n^{-2} \sum_{i,j} \sum_{0 \le k+l \le 1} \widetilde{S}_k(\mathbf{u}_i, \mathbf{u}_j) \widetilde{S}_l(\mathbf{v}_i, \mathbf{v}_j) (\Delta \mathbf{w}_i^{\mathrm{T}}, \Delta \mathbf{w}_j^{\mathrm{T}})^{\otimes 2} D^{\otimes 2} \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) \\
&\quad + 6^{-1} n^{-2} \sum_{i,j} \widetilde{S}_0(\mathbf{u}_i, \mathbf{u}_j) \widetilde{S}_0(\mathbf{v}_i, \mathbf{v}_j) (\Delta \mathbf{w}_i^{\mathrm{T}}, \Delta \mathbf{w}_j^{\mathrm{T}})^{\otimes 3} D^{\otimes 3} \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) + o_p(n^{-1}) \\
&\stackrel{\text{def}}{=} Q_1' + Q_2' + Q_3' + Q_4' + o_p(n^{-1}).
\end{aligned}$$

Then following similar arguments in the proof of Theorem 3, we have Q_2', Q_3' and Q_4' are all of order $o_p(n^{-1})$ and Q_1' equals $n^{-2} \sum_{i,j} S_{\mathbf{u}}(\mathbf{u}_i, \mathbf{u}_j) S_{\mathbf{v}}(\mathbf{v}_i, \mathbf{v}_j) \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) + o_p(n^{-1})$. Combing these results, we have

$$\widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = n^{-2} \sum_{i,j} S_{\mathbf{u}}(\mathbf{u}_i, \mathbf{u}_j) S_{\mathbf{v}}(\mathbf{v}_i, \mathbf{v}_j) \widetilde{g}(\mathbf{w}_i, \mathbf{w}_j) + o_p(n^{-1}),$$

where the right hand side is a first order degenerate V statistics. Thus by applying Theorem 6.4.1.B of Serfling (2009),

$$n \widehat{\rho}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1)$$

where $\chi_j^2(1)$, $j = 1, 2, \ldots$ are independent $\chi^2(1)$ random variables, and λ_j, $j = 1, 2, \ldots$ are the corresponding eigenvalues of $\widetilde{h}(\mathbf{u}, \mathbf{v}, \mathbf{w}; \mathbf{u}', \mathbf{v}', \mathbf{w}')$. Therefore, the proof is completed. □

A.8 Proof of Theorem 8

It suffices to show that $X \perp\!\!\!\perp Y \mid Z$ if and only if $U \perp\!\!\!\perp V \mid Z$ because $U \perp\!\!\!\perp V \mid Z$ is equivalent to U, V and Z are mutually independent under $U \perp\!\!\!\perp Z$ and $V \perp\!\!\!\perp Z$.

We only show that $X \perp\!\!\!\perp Y \mid Z$ if and only if $U \perp\!\!\!\perp Y \mid Z$, because similar arguments will yield that it is also equivalent to $U \perp\!\!\!\perp V \mid Z$. It is quite straightforward that $X \perp\!\!\!\perp Y \mid Z$ implies $U \perp\!\!\!\perp Y \mid Z$. While when $U \perp\!\!\!\perp Y \mid Z$, we have for each $Z = z$, u and y in the corresponding support,

$$\Pr(U \le u, Y \le y \mid Z = z) = u F_{Y|Z}(y \mid z).$$

Substituting $U = (1 - U_X) F_{X|Z}(X- \mid Z) + U_X F_{X|Z}(X \mid Z)$ into the above equation, with some straight calculation, the left hand side is

$$\begin{aligned}
&\Pr\left\{ \Pr(\widetilde{X} < X \mid X, \widetilde{Z} = z) + \Pr(\widetilde{X} = X \mid X, \widetilde{Z} = z) U_X \le u, Y \le y \mid Z = z \right\} \\
= \; &\Pr\left\{ \Pr(\widetilde{X} < X \mid X, \widetilde{Z} = z) + \Pr(\widetilde{X} = X \mid X, \widetilde{Z} = z) U_X \le u, F_{Y|X,Z}(y \mid X, z) \mid Z = z \right\}.
\end{aligned}$$

Because U_X is standard uniformly distributed, we obtain

$$E\left[g\left\{\frac{u-\Pr(\widetilde{X}<X\mid X,\widetilde{Z}=Z=z)}{\Pr(\widetilde{X}=X\mid X,\widetilde{Z}=Z=z)}\right\}F_{Y|X,Z}(y\mid X,z)\right]=uF_{Y|Z}(y\mid z),$$

where $g(\cdot)$ is the cumulative distribution function of a standard uniformly distributed random variable. Now assume that conditional on $Z=z$, the support of X is $\{x_1,\ldots,x_N\}$, where $x_1<\ldots<x_N$. Therefore, when $0<u<F_{X|Z}(x_1\mid z)$, the expectation in the above equation is

$$\begin{aligned}&\sum_{i=1}^{N} g\left\{\frac{u-\Pr\left(X<x_i\mid Z=z\right)}{\Pr\left(X=x_i\mid Z=z\right)}\right\}F_{Y|X,Z}(y\mid x_i,z)\Pr\left(X=x_i\mid Z=z\right)\\ =\;&\frac{u-\Pr\left(X<x_1\mid Z=z\right)}{\Pr\left(X=x_1\mid Z=z\right)}\Pr\left(X=x_1\mid Z=z\right)F_{Y|X,Z}(y\mid x_1,z)\\ =\;&uF_{Y|X,Z}(y\mid x_1,z).\end{aligned}$$

The expectation equals $uF_{Y|Z}(y\mid z)$. That is, $F_{Y|X,Z}(y\mid x_1,z)=F_{Y|Z}(y\mid z)$.

When $F_{X|Z}(x_1\mid z)<u<F_{X|Z}(x_2\mid z)$, we can calculate the expectation as

$$\begin{aligned}&\sum_{i=1}^{N} g\left\{\frac{u-\Pr\left(X<x_i\mid Z=z\right)}{\Pr\left(X=x_i\mid Z=z\right)}\right\}F_{Y|X,Z}(y\mid x_i,z)\Pr\left(X=x_i\mid Z=z\right)\\ =\;&F_{Y|X,Z}(y\mid x_1,z)\Pr\left(X=x_1\mid Z=z\right)+\left\{u-\Pr\left(X<x_2\mid Z=z\right)\right\}F_{Y|X,Z}(y\mid x_2,z).\end{aligned}$$

Since we have shown that $F_{Y|X,Z}(y\mid x_1,z)=F_{Y|Z}(y\mid z)$, with the fact that the expectation equals $uF_{Y|Z}(y\mid z)$, we can get $F_{Y|X,Z}(y\mid x_2,z)=F_{Y|Z}(y\mid z)$.

Similarly, we can obtain that $F_{Y|X,Z}(y\mid x_k,z)=F_{Y|Z}(y\mid z)$, $k=3,\ldots,N$. Consequently, we have $F_{Y|X,Z}(y\mid x,z)=F_{Y|Z}(y\mid z)$ for all x,y and z in their support. That is, $X\perp\!\!\!\perp Y\mid Z$. Therefore, the proof is completed.

□

Appendix B. Additional Simulations Results

We consider the directed acyclic graph with 5 nodes, i.e., $X=(X_1,\ldots,X_5)$, and only allow directed edge from X_i and X_j for $i<j$. Denote the adjacency matrix A. The existence of the edge follows a Bernoulli distribution, and we set $\Pr(A_{i,j}=1)=0.4$, for $i<j$. When $A_{i,j}=1$, we replace $A_{i,j}$ with independent realizations of a uniform $U(0.1,1)$ random variable. The value of the first random variable X_1 is randomly sampled from some distribution $\widetilde{P}$. Specifically,

$$\epsilon_1\sim\widetilde{P},\text{ and } X_1=\epsilon_1.$$

The value of the next nodes is

$$\epsilon_j\sim\widetilde{P},\text{ and } X_j=\sum_{k=1}^{j-1}A_{j,k}X_k+\epsilon_j$$

for $j = 1, \cdots, p$. All random errors $\epsilon_1, \ldots, \epsilon_p$ are independently sampled from the distribution $\widetilde{P}$. We consider two scenarios, where $\widetilde{P}$ follows either normal distribution $N(0,1)$ or uniform distribution $U(0,1)$. We compare our proposed conditional independence test (denoted by "CIT") with other popular conditional dependence measure. They are, respectively, the partial correlation (, denoted by "PCR") conditional mutual information (Scutari, 2010, denoted by "CMI"), and the KCI.test (Zhang et al., 2011, denoted by "KCI"). We set the sample size n = 50, 100, 200 and 300. The true positive rate and false positive rate for the four different tests are reported in Tables 5, from which we can see that as the sample size increases, the true positive rate of the proposed method steadily grows, and the proposed method outperforms the other tests, while the false positive rate remains under control with slightly decrease.

Table 5: The true positive rate and false positive rate for the causal discovery of the directed acyclic graph with different tests

Samples	50	100	200	300	50	100	200	300
Tests	$\widetilde{P} \sim N(0,1)$							
	true positive rate				false positive rate			
CIT	0.555	0.658	0.734	0.789	0.117	0.112	0.107	0.103
PCR	0.479	0.489	0.546	0.589	0.101	0.110	0.130	0.135
CMI	0.472	0.530	0.604	0.590	0.097	0.127	0.143	0.150
KCI	0.360	0.516	0.592	0.634	0.072	0.135	0.144	0.168
Tests	$\widetilde{P} \sim U(0,1)$							
	true positive rate				false positive rate			
CIT	0.468	0.587	0.734	0.736	0.070	0.099	0.095	0.113
PCR	0.469	0.526	0.588	0.545	0.111	0.129	0.140	0.140
CMI	0.497	0.568	0.566	0.633	0.103	0.127	0.138	0.149
KCI	0.386	0.458	0.523	0.564	0.082	0.099	0.123	0.122

References

AE Brockwell. Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478, 2007.

Shubhadeep Chakraborty and Xianyang Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528):1638–1650, 2019.

R Dennis Cook. *Regression Graphics: Ideas for Studying Regressions through Graphics*, volume 482. John Wiley & Sons, 2009.

Mathias Drton, Fang Han, and Hongjian Shi. High dimensional independence testing with maxima of rank correlations. *The Annals of Statistics*, To Appear, 2020.

Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

Michael Irwin Jordan. *Learning in Graphical Models*, volume 89. Springer Science & Business Media, 1998.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.

AJ Lawrance. On conditional and partial correlation. *The American Statistician*, 30(3): 146–149, 1976.

Qi Li and Jeffrey Scott Racine. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, 2007.

Oliver Linton and Pedro Gozalo. Conditional independence restrictions: Testing and estimation. *Cowels Foundation Discussion Paper*, 1996.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

Rohit K Patra, Bodhisattva Sen, and Gábor J Székely. On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213, 2016.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceeding of International Conference on Artificial Intelligence and Statistics*, pages 938–947, 2018.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software, Articles*, 35(3):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, 2009.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

Kyungchul Song. Testing conditional independence via rosenblatt transforms. *The Annals of Statistics*, 37(6B):4011–4045, 2009.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

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

Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

Liangjun Su and Halbert White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

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

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer Science & Business Media, 2013.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813. AUAI Press, 2011.

Yeqing Zhou, Jingyuan Liu, and Liping Zhu. Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis*, 175:104557, 2020.

Liping Zhu, Yaowu Zhang, and Kai Xu. Measuring and testing for interval quantile dependence. *The Annals of Statistics*, 46(6A):2683–2710, 2018.