Journal of the American Statistical Association

December 2019 ▪ Volume 114 ▪ Number 528

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Distance Metrics for Measuring Joint Dependence with Application to Causal Inference

Shubhadeep Chakraborty & Xianyang Zhang

To cite this article: Shubhadeep Chakraborty & Xianyang Zhang (2019) Distance Metrics for Measuring Joint Dependence with Application to Causal Inference, Journal of the American Statistical Association, 114:528, 1638-1650, DOI: 10.1080/01621459.2018.1513364

To link to this article: https://doi.org/10.1080/01621459.2018.1513364

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 528, 1638–1650: Theory and Methods
https://doi.org/10.1080/01621459.2018.1513364

Taylor & Francis
Taylor & Francis Group

Distance Metrics for Measuring Joint Dependence with Application to Causal Inference

Shubhadeep Chakraborty and Xianyang Zhang

Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT
Many statistical applications require the quantification of joint dependence among more than two random vectors. In this work, we generalize the notion of distance covariance to quantify joint dependence among $d \geq 2$ random vectors. We introduce the high-order distance covariance to measure the so-called Lancaster interaction dependence. The joint distance covariance is then defined as a linear combination of pairwise distance covariances and their higher-order counterparts which together completely characterize mutual independence. We further introduce some related concepts including the distance cumulant, distance characteristic function, and rank-based distance covariance. Empirical estimators are constructed based on certain Euclidean distances between sample elements. We study the large-sample properties of the estimators and propose a bootstrap procedure to approximate their sampling distributions. The asymptotic validity of the bootstrap procedure is justified under both the null and alternative hypotheses. The new metrics are employed to perform model selection in causal inference, which is based on the joint independence testing of the residuals from the fitted structural equation models. The effectiveness of the method is illustrated via both simulated and real datasets. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2017
Revised July 2018

KEYWORDS
Bootstrap; Directed acyclic graph; Distance covariance; Interaction dependence; *U*-statistic; *V*-statistic

1. Introduction

Measuring and testing dependence is of central importance in statistics, which has found applications in a wide variety of areas including independent component analysis, gene selection, graphical modeling, and causal inference. Statistical tests of independence can be associated with widely many dependence measures. Two of the most classical measures of association between two ordinal random variables are Spearman's rho and Kendall's tau. However, tests for (pairwise) independence using these two classical measures of association are not consistent, and only have power for alternatives with monotonic association. Contingency table-based methods, and in particular the power-divergence family of test statistics (Read and Cressie 1988), are the best known general purpose tests of independence, but are limited to relatively low dimensions, since they require a partitioning of the space in which each random variable resides. Another classical measure of dependence between two random vectors is the mutual information (Cover and Thomas 1991), which can be interpreted as the Kullback–Leibler divergence between the joint density and the product of the marginal densities. The idea originally dates back to the 1950s, in groundbreaking works by Shannon and Weaver (1949), Mcgill (1954), and Fano (1961). Mutual information completely characterizes independence and generalizes to more than two random vectors. However, test based on mutual information involves distributional assumptions for the random vectors and hence is not robust to model misspecification.

In the past 15 years, kernel-based methods have received considerable attention in both the statistics and machine learning literature. For instance, Bach and Jordan (2002) derived a regularized correlation operator from the covariance and cross-covariance operators and used its largest singular value to conduct independence test. Gretton et al. (2005, 2007) introduced a kernel-based independence measure, namely, the Hilbert-Schmidt Independence Criterion (HSIC), to test for independence of two random vectors. This idea was recently extended by Sejdinovic, Gretton, and Bergsma (2013) and Pfister et al. (2018) to quantify the joint independence among more than two random vectors.

Along with a different direction, Székely, Rizzo, and Bakirov (2007), in their seminal article, introduced the notion of distance covariance (dCov) and distance correlation as a measure of dependence between two random vectors of arbitrary dimensions. Given the theoretical appeal of the population quantity and the striking simplicity of the sample version, the idea has been widely extended and analyzed in various ways in Székely and Rizzo (2012, 2014), Lyons (2013), Sejdinovic, Gretton, and Bergsma (2013), Dueck et al. (2014), Bergsma and Dassios (2014), Wang et al. (2015), and Huo and Székely (2016), to mention only a few. The dCov between two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with finite first moments is defined as the positive square root of

$$\mathrm{dCov}^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2}{|t|_p^{1+p} |s|_q^{1+q}} dtds,$$

CONTACT Shubhadeep Chakraborty shubhadeep@stat.tamu.edu Department of Statistics, Texas A&M University, College Station, TX 77843.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

where f_X, f_Y, and $f_{X,Y}$ are the individual and joint characteristic functions of X and Y, respectively, $|\cdot|_p$ is the Euclidean norm of $\mathbb{R}^p$, $c_p = \pi^{(1+p)/2}/\Gamma((1+p)/2)$ is a constant with $\Gamma(\cdot)$ being the complete gamma function. An important feature of dCov is that it fully characterizes independence because $\mathrm{dCov}(X, Y) = 0$ if and only if X and Y are independent.

Many statistical applications require the quantification of joint dependence among $d \geq 2$ random variables (or vectors). Examples include model diagnostic checking for directed acyclic graph (DAG) where inferring pairwise independence is not enough in this case (see more details in Section 6), and independent component analysis which is a means for finding a suitable representation of multivariate data such that the components of the transformed data are mutually independent. In this article, we shall introduce new metrics which generalize the notion of dCov to quantify joint dependence of $d \geq 2$ random vectors. We first introduce the notion of high-order dCov to measure the so-called Lancaster interaction dependence (Lancaster 1969). We generalize the notion of Brownian covariance (Székely and Rizzo 2009) and show that it coincides with the high-order distance covariance. We then define the joint dCov (Jdcov) as a linear combination of pairwise dCov and their high-order counterparts. The proposed metric provides a natural decomposition of joint dependence into the sum of lower-order and high-order effects, where the relative importance of the lower-order effect terms and the high-order effect terms is determined by a user-chosen number. In the population case, Jdcov is equal to zero if and only if the d random vectors are mutually independent, and thus completely characterizes joint independence. It is also worth mentioning that the proposed metrics are invariant to permutation of the variables and they inherit some nice properties of dCov, see Section 2.2.

Following the idea of Streitberg (1990), we introduce the concept of distance cumulant and distance characteristic function, which leads us to an equivalent characterization of independence of the d random vectors. Furthermore, we establish a scale invariant version of Jdcov and discuss the concept of rank-based distance measures, which can be viewed as the counterparts of Spearman's rho to dCov and JdCov.

JdCov and its scale-invariant versions can be conveniently estimated in finite sample using V-statistics or their bias-corrected versions. We study the asymptotic properties of the estimators, and introduce a bootstrap procedure to approximate their sampling distributions. The asymptotic validity of the bootstrap procedure is justified under both the null and alternative hypotheses. The new metrics are employed to perform model selection in a causal inference problem, which is based on the joint independence testing of the residuals from the fitted structural equation models. We compare our tests with the bootstrap version of the d-variate HSIC (dHSIC) test recently introduced in Pfister et al. (2018) and the mutual independence test proposed by Matteson and Tsay (2017). Finally, we remark that although we focus on Euclidean space-valued random variables, our results can be readily extended to general metric spaces in view of the results in Lyons (2013).

The rest of the article is organized as follows. Section 2.1 introduces the high-order distance covariance and studies its basic properties. Section 2.2 describes the JdCov to quantity joint dependence. Sections 2.3–2.4 further introduce some related concepts including the distance cumulant, distance characteristic function, and rank-based distance covariance. We study the estimation of the distance metrics in Section 3 and present a joint independence test based on the proposed metrics in Section 4. Section 5 is devoted to numerical studies. The new metrics are employed to perform model selection in causal inference in Section 6. Section 7 discusses the efficient computation of distance metrics and future research directions. The technical details are gathered in the supplementary material.

Notations. Consider $d \geq 2$ random vectors $\mathcal{X} = \{X_1, \dots, X_d\}$, where $X_i \in \mathbb{R}^{p_i}$. Set $p_0 = \sum_{i=1}^{d} p_i$. Let $\{X_1', \dots, X_d'\}$ be an independent copy of $\mathcal{X}$. Denote by $\imath = \sqrt{-1}$ the imaginary unit. Let $|\cdot|_p$ be the Euclidean norm of $\mathbb{R}^p$ with the subscript omitted later without ambiguity. For $a, b \in \mathbb{R}^p$, let $\langle a, b\rangle = a^\top b$. For a complex number a, denote by $\bar{a}$ its conjugate. Let f_i be the characteristic function of X_i, that is, $f_i(t) = \mathbb{E}[e^{\imath\langle t, X_i\rangle}]$ with $t \in \mathbb{R}^{p_i}$. Define $w_p(t) = (c_p|t|_p^{1+p})^{-1}$ with $c_p = \pi^{(1+p)/2}/\Gamma((1+p)/2)$. Write $dw = (c_{p_1}c_{p_2}\dots c_{p_d}|t_1|_{p_1}^{1+p_1}\dots|t_d|_{p_d}^{1+p_d})^{-1}dt_1\dots dt_d$. Let I_k^d be the collection of k-tuples of indices from $\{1, 2, \dots, d\}$ such that each index occurs exactly once. Denote by $\lfloor a\rfloor$ the integer part of $a \in \mathbb{R}$. Write $X \perp\!\!\!\perp Y$ if X is independent of Y.

2. Measuring Joint Dependence

2.1. High-Order Distance Covariance

We briefly review the concept of Lancaster interactions first introduced by Lancaster (1969). The Lancaster interaction measure associated with a multidimensional probability distribution of d random variables $\{X_1, \dots, X_d\}$ with the joint distribution $F = F_{1,2,\dots,d}$, is a signed measure ΔF given by

$$\Delta F = (F_1^* - F_1)(F_2^* - F_2)\dots(F_d^* - F_d), \tag{1}$$

where after expansion, a product of the form $F_i^*F_j^*\dots F_k^*$ denotes the corresponding joint distribution function $F_{i,j,\dots,k}$ of $\{X_i, X_j, \dots, X_k\}$. For example, for $d = 4$, the term $F_1^*F_2^*F_3F_4$ stands for $F_{12}F_3F_4$, $F_1^*F_2F_3F_4$ stands for $F_1F_2F_3F_4$, etc. In particular for $d = 3$, (1) simplifies to

$$\Delta F = F_{123} - F_1F_{23} - F_2F_{13} - F_3F_{12} + 2F_1F_2F_3\,. \tag{2}$$

In light of the Lancaster interaction measure, we introduce the concept of dth-order dCov as follows.

Definition 1. The dth-order dCov is defined as the positive square root of

$$\mathrm{dCov}^2(X_1, \dots, X_d) = \int_{\mathbb{R}^{p_0}} \left|\mathbb{E}\left[\prod_{i=1}^{d}(f_i(t_i) - e^{\imath\langle t_i, X_i\rangle})\right]\right|^2 dw. \tag{3}$$

When $d = 2$, it reduces to the dCov in Székely, Rizzo, and Bakirov (2007).

The term $\mathbb{E}[\prod_{i=1}^{d}(f_i(t_i) - e^{\imath\langle t_i, X_i\rangle})]$ in the definition of dCov is a counterpart of the Lancaster interaction measure in (1) with the joint distribution functions replaced by the joint characteristic functions. When $d = 3$, $\mathrm{dCov}^2(X_1, X_2, X_3) > 0$ rules out the possibility of any factorization of the joint distribution. To see this, we note that $X_1 \perp\!\!\!\perp (X_2, X_3)$, $X_2 \perp\!\!\!\perp (X_1, X_3)$, or $X_3 \perp\!\!\!\perp (X_1, X_2)$ all lead to $\mathrm{dCov}^2(X_1, X_2, X_3) = 0$. On the other hand, $\mathrm{dCov}^2(X_1, X_2, X_3) = 0$ implies that

$$\begin{aligned} & f_{123}(t_1, t_2, t_3) - f_1(t_1)f_2(t_2)f_3(t_3) \\ & \quad = f_1(t_1)f_{23}(t_2, t_3) + f_2(t_2)f_{13}(t_1, t_3) + f_3(t_3)f_{12}(t_1, t_2) \\ & \qquad - 3f_1(t_1)f_2(t_2)f_3(t_3) \end{aligned}$$

for $t_i \in \mathbb{R}^{p_i}$ almost everywhere. In this case, the "higher-order effect" that is, $f_{123}(t_1, t_2, t_3) - f_1(t_1)f_2(t_2)f_3(t_3)$ can be represented by the "lower-order/pairwise effects" $f_{ij}(t_i, t_j) - f_i(t_i)f_j(t_j)$ for $1 \leq i \neq j \leq 3$. However, this does not necessarily imply that X_1, X_2, and X_3 are jointly independent. In other words when $d = 3$ (or more generally when $d \geq 3$), joint independence of X_1, X_2, and X_3 is not a necessary condition for dCov to be zero. To address this issue, we shall introduce a new distance metric to quantify any forms of dependence among $\mathcal{X}$ in Section 2.2.

In the following, we present some basic properties of high-order dCov. Define the bivariate function $U_i(x, x') = \mathbb{E}|x - X_i'| + \mathbb{E}|X_i - x'| - |x - x'| - \mathbb{E}|X_i - X_i'|$ for $x, x' \in \mathbb{R}^{p_i}$ with $1 \leq i \leq d$. Our definition of dCov is partly motivated by the following lemma.

Lemma 1. For $1 \leq i \leq d$,

$$U_i(x, x') = \int_{\mathbb{R}^{p_i}} \left\{ (f_i(t) - e^{\imath \langle t, x \rangle})(f_i(-t) - e^{-\imath \langle t, x' \rangle}) \right\} w_{p_i}(t) dt.$$

By Lemma 1 and Fubini's theorem, the dth-order (squared) dCov admits the following equivalent representation,

$$\begin{aligned} \mathrm{dCov}^2(X_1, \ldots, X_d) &= \int_{\mathbb{R}^{p_0}} \left| \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath \langle t_i, X_i \rangle}) \right] \right|^2 dw \\ &= \int_{\mathbb{R}^{p_0}} \mathbb{E}\left[\prod_{i=1}^{d} (f_i(t_i) - e^{\imath \langle t_i, X_i \rangle}) \right] \\ & \quad \times \mathbb{E}\left[\prod_{i=1}^{d} \overline{(f_i(t_i) - e^{\imath \langle t_i, X_i' \rangle})} \right] dw \\ &= \mathbb{E}\left[\prod_{i=1}^{d} U_i(X_i, X_i') \right]. \end{aligned} \tag{4}$$

This suggests that similar to dCov, its high-order counterpart has an expression based on the moments of U_i's, which results in very simple and applicable empirical formulas, see more details in Section 3.

From the definition of dCov in Székely, Rizzo, and Bakirov (2007), it might appear that its most natural generalization to the case of $d = 3$ would be to define a measure in the following way:

$$\frac{1}{c_p c_q c_r} \int_{\mathbb{R}^{p+q+r}} \frac{|f_{X,Y,Z}(t, s, u) - f_X(t)f_Y(s)f_Z(u)|^2}{|t|_p^{1+p} |s|_q^{1+q} |u|_r^{1+r}} dtdsdu,$$

where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, and $Z \in \mathbb{R}^r$. Assuming that the integral above exists, one can easily verify that such a measure completely characterizes joint independence among X, Y, and Z. However, it does not admit a nice equivalent representation as in (4) (unless one considers a different weighting function). We exploit this equivalent representation of the dth-order dCov to propose a V-statistic-type estimator of the population quantity (see Section 3) which is much simpler to compute rather than evaluating an integral as in the original definition in (3).

Székely and Rizzo (2009) introduced the notion of covariance with respect to a stochastic process. Theorem 8 in Székely and Rizzo (2009) shows that the population distance covariance coincides with the covariance with respect to Brownian motion (or the so-called *Brownian covariance*). Remark 1.1 in the supplementary materials generalizes the notion of Brownian covariance for $d \geq 2$ random vectors and establishes a connection with the high-order distance covariances.

The following proposition shows that the high-order distance covariances are invariant to translation, orthogonal transformation, and permutation on X_i's.

Proposition 1. For any $a_i \in \mathbb{R}^{p_i}$, $c_i \in \mathbb{R}$, and orthogonal transformations $A_i \in \mathbb{R}^{p_i \times p_i}$, $\mathrm{dCov}^2(a_1 + c_1 A_1 X_1, \ldots, a_d + c_d A_d X_d) = \prod_{i=1}^{d} |c_i| \, \mathrm{dCov}^2(X_1, \ldots, X_d)$. Moreover, dCov is invariant to any permutation of $\{X_1, X_2, \ldots, X_d\}$.

Theorem 7 in Székely, Rizzo, and Bakirov (2007) shows the relationship between distance correlation and the correlation coefficient for bivariate normal distributions. We extend that result in case of multivariate normal random variables with zero mean, unit variance, and pairwise correlation ρ. Proposition 2 establishes a relationship between the correlation coefficient and higher-order distance covariances for multivariate normal random variables.

Proposition 2. Suppose $(X_1, X_2, \ldots, X_d) \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{i,j})_{i,j=1}^{d}$ with $\sigma_{ii} = 1$ for $1 \leq i \leq d$ and $\sigma_{ij} = \rho$ for $1 \leq i \neq j \leq d$. When $d = 2k - 1$ or $d = 2k$, $\mathrm{dCov}^2(X_1, \ldots, X_d) = O(|\rho|^{2k})$ for $k \geq 2$.

Proposition 1.1 in the supplementary materials shows some additional properties of the dth-order dCov. Property (1) in Proposition 1.1 gives an upper bound for $\mathrm{dCov}^2(X_1, X_2, \ldots, X_d)$, which is motivated by Lemma 2.1 of Lyons (2013), whereas an alternative upper bound is given in Property (2) which follows directly from the Hölder's inequality. Property (3) allows us to represent dCov of random vectors of any dimensions as an integral of dCov of univariate random variables, which are the projections of the aforementioned random vectors.

2.2. Joint Distance Covariance

In this subsection, we introduce a new joint dependence measure called the joint dCov (Jdcov), which is designed to capture all types of interaction dependence among the d random vectors. To achieve this goal, we define JdCov as the linear combination of all kth-order dCov for $1 \leq k \leq d$.

Definition 2. The JdCov among $\{X_1,\dots,X_d\}$ is given by

$$\begin{aligned}&\mathrm{JdCov}^2(X_1,\dots,X_d;C_2,\dots,C_d)\\ =&C_2\sum_{(i_1,i_2)\in I_2^d}\mathrm{dCov}^2(X_{i_1},X_{i_2})+C_3\sum_{(i_1,i_2,i_3)\in I_3^d}\mathrm{dCov}^2(X_{i_1},X_{i_2},X_{i_3})\\ &+\cdots+C_d\,\mathrm{dCov}^2(X_1,\dots,X_d),\end{aligned}\tag{5}$$

for some nonnegative constants $C_i \geq 0$ with $2 \leq i \leq d$.

Proposition 3 states that JdCov completely characterizes joint independence among $\{X_1,\dots,X_d\}$.

Proposition 3. Suppose $C_i > 0$ for $2 \leq i \leq d$. Then $\mathrm{JdCov}^2(X_1,\dots,X_d;C_2,\dots,C_d) = 0$ if and only if $\{X_1,\dots,X_d\}$ are mutually independent.

Next we show that by properly choosing C_i's, $\mathrm{JdCov}^2(X_1,\dots,X_d;C_2,\dots,C_d)$ has a relatively simple expression, which does not require the evaluation of $2^d - d - 1$ dCov terms in its original definition (5). Specifically, let $C_i = c^{d-i}$ for $c \geq 0$ in the definition of JdCov and denote $\mathrm{JdCov}^2(X_1,\dots,X_d;c) = \mathrm{JdCov}^2(X_1,\dots,X_d;c^{d-2},c^{d-1},\dots,1)$. Then, we have the following result.

Proposition 4. For any $c \geq 0$,

$$\mathrm{JdCov}^2(X_1,\dots,X_d;c) = \mathbb{E}\left[\prod_{i=1}^{d}\left(U_i(X_i,X_i')+c\right)\right] - c^d.$$

In particular, $\mathrm{JdCov}^2(X_1,X_2;c) = E[U_1(X_1,X_1')U_2(X_2,X_2')] = \mathrm{dCov}^2(X_1,X_2)$.

By (5), the dependence measured by JdCov can be decomposed into the main effect term $\sum_{(i_1,i_2)\in I_2^d}\mathrm{dCov}^2(X_{i_1},X_{i_2})$ quantifying the pairwise dependence as well as the higher-order effect terms $\sum_{(i_1,i_2,\dots,i_k)\in I_k^d}\mathrm{dCov}^2(X_{i_1},X_{i_2},\dots,X_{i_k})$ quantifying the multi-way interaction dependence among any k-tuples. The choice of c reflects the relative importance of the main effect and the higher-order effects. For $c \geq 1$, $C_i = c^{d-i}$ is nonincreasing in i. Thus, the larger c we select, the smaller weights we put on the higher-order terms. In particular, we have

$$\lim_{c\to+\infty} c^{2-d}\mathrm{JdCov}^2(X_1,\dots,X_d;c) = \sum_{(i_1,i_2)\in I_2^d}\mathrm{dCov}^2(X_{i_1},X_{i_2}),$$

that is, JdCov reduces to the main effect term as $c \to +\infty$. We remark that the main effect term fully characterizes joint dependence in the case of elliptical distribution and it has been recently used in Yao, Zhang, and Shao (2018) to test mutual independence for high-dimensional data. On the other hand, JdCov becomes the dth-order dCov as $c \to 0$, that is,

$$\lim_{c\to 0}\mathrm{JdCov}^2(X_1,\dots,X_d;c) = \mathrm{dCov}^2(X_1,\dots,X_d).$$

The choice of c depends on the types of interaction dependence of interest as well as the specific scientific problem, and thus is left for the user to decide.

It is worth noting that $\mathrm{JdCov}^2(X_1,\dots,X_d;c)$ depends on the scale of X_i. To obtain a scale-invariant metric, one can normalize U_i by the corresponding distance variance. Specifically, when $\mathrm{dCov}(X_i) := \mathrm{dCov}(X_i,X_i) > 0$, the resulting quantity is given by

$$\mathrm{JdCov}_S^2(X_1,\dots,X_d;c) = \mathbb{E}\left[\prod_{i=1}^{d}\left(\frac{U_i(X_i,X_i')}{\mathrm{dCov}(X_i)}+c\right)\right] - c^d,$$

which is scale-invariant. Another way to obtain a scale-invariant metric is presented in Section 2.4 based on the idea of rank transformation.

Below we present some basic properties of JdCov, which follow directly from Proposition 1.

Proposition 5. We have the following properties regarding JdCov:

1. For any $a_i \in \mathbb{R}^{p_i}$, $c_0 \in \mathbb{R}$, and orthogonal transformations $A_i \in \mathbb{R}^{p_i\times p_i}$, $\mathrm{JdCov}^2(a_1+c_0A_1X_1,\dots,a_d+c_0A_dX_d;|c_0|c) = |c_0|^d\mathrm{JdCov}^2(X_1,\dots,X_d;c)$. Moreover, JdCov is invariant to any permutation of $\{X_1,X_2,\dots,X_d\}$.
2. For any $a_i \in \mathbb{R}^{p_i}$, $c_i \neq 0$, and orthogonal transformations $A_i \in \mathbb{R}^{p_i\times p_i}$, $\mathrm{JdCov}_S^2(a_1+c_1A_1X_1,\dots,a_d+c_dA_dX_d;c) = \mathrm{JdCov}_S^2(X_1,\dots,X_d;c)$.

A natural question to ask is what should be a data-driven way to choose the tuning parameter c. Although we leave it for future research, here we present a heuristic idea of choosing c. In the discussion below Proposition 4 in Section 2.2, we pointed out that choosing $c > 1$ (or < 1) puts lesser (or higher) weightage on the higher-order effects. Note that if the data are Gaussian, testing for the mutual independence of $\{X_1,\dots,X_d\}$ is equivalent to testing for their pairwise independences. In that case, intuitively one should choose a larger (> 1) value of c. If, however, the data are non-Gaussian, it might be of interest to look into higher-order dependencies and thus a smaller (< 1) choice of c makes sense.

To summarize, a heuristic way to choose the tuning parameter c could be

$$\text{Choose c}\begin{cases}>1, & \text{if } \{X_1,\dots,X_d\} \text{ are jointly Gaussian}\\ <1, & \text{if } \{X_1,\dots,X_d\} \text{ are not jointly Gaussian.}\end{cases}\tag{6}$$

There is a huge literature on testing for joint normality of random vectors (see, e.g., Mardia 1970; Malkovich and Afifi 1973; Baringhaus and Henze 1988; Bowman and Foster 1993; Henze and Wagner 1997). It has been shown that the test based on energy distance is consistent against fixed alternatives (Székely and Rizzo 2004) and shows higher empirical power compared to several competing tests (see Székely and Rizzo 2005, 2013). Suppose p is the p-value of the energy distance-based test for joint normality of $\{X_1,\dots,X_d\}$ at level α. We expect c to increase (or decrease) from 1 as $p >$ (or $<$) α, so one heuristic choice of c can be

$$c = 1 + \mathrm{sign}(p-\alpha) \times |p-\alpha|^{1/4}\,,\tag{7}$$

where $\mathrm{sign}(x) = 1, 0 \; or -1$ depending on whether $x > 0$, $x = 0 \; or \; x < 0$. For example, $p = (0.001, 0.03, 0.0499, 0.0501, 0.1, 0.3)$ and $\alpha = 0.05$ yields $c = (0.53, 0.62, 0.9, 1.1, 1.47, 1.71)$.

Table 1. Comparison of various distance metrics for measuring joint dependence of $d \geq 2$ random vectors of arbitrary dimensions.

Distance metrics	Complete characterization of joint independence	Permutation invariance	Scale invariance
dHSIC	✓	✓	× (for fixed bandwidth)
T_{MT}	✓	×	×
High-order dCov	× (Captures Lancaster interactions)	✓	×
JdCov	✓	✓	×
JdCov$_S$	✓	✓	✓
JdCov$_R$	✓	✓	✓

2.3. *Distance Cumulant and Distance Characteristic Function*

As noted in Streitberg (1990), for $d \geq 4$, the Lancaster interaction measure fails to capture all possible factorizations of the joint distribution. For example, it may not vanish if $(X_1, X_2) \perp\!\!\!\perp (X_3, X_4)$. Streitberg (1990) corrected the definition of Lancaster interaction measure using a more complicated construction, which essentially corresponds to the cumulant version of dCov in our context. Specifically, Streitberg (1990) proposed a corrected version of Lancaster interaction as follows:

$$\widetilde{\Delta} F = \sum_{\pi} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{D\in\pi} F_D,$$

where π is a partition of the set {1,2,…,d}, $|\pi|$ denotes the number of blocks of the partition π, and F_D denotes the joint distribution of $\{X_i : i \in D\}$. It has been shown in Streitberg (1990) that $\widetilde{\Delta} F = 0$ whenever F is decomposable. Our definition of joint distance cumulant of $\{X_1, \dots, X_d\}$ below can be viewed as the dCov version of Streitberg's correction.

Definition 3. The joint distance cumulant among $\{X_1, \dots, X_d\}$ is defined as

$$\operatorname{cum}(X_1, \dots, X_d) = \sum_{\pi} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{D\in\pi} \times \mathbb{E}\left(\prod_{i\in D} U_i(X_i, X_i')\right), \quad (8)$$

where π runs through all partitions of $\{1, 2, \dots, d\}$.

It is not hard to verify that $\operatorname{cum}(X_1, \dots, X_d) = 0$ if $\{X_1, \dots, X_d\}$ can be decomposed into two mutually independent groups say $(X_i)_{i\in\pi_1}$ and $(X_j)_{j\in\pi_2}$ with π_1 and π_2 being a partition of $\{1, 2, \dots, d\}$. We further define the distance characteristic function.

Definition 4. The joint distance characteristic function among $\{X_1, \dots, X_d\}$ is defined as

$$\operatorname{dcf}(t_1, \dots, t_d) = \mathbb{E}\left[\exp\left(\imath \sum_{i=1}^{d} t_i U_i(X_i, X_i')\right)\right], \quad (9)$$

for $t_1, \dots, t_d \in \mathbb{R}$.

The following result shows that distance cumulant can be interpreted as the coefficient of the Taylor expansion of the log distance characteristic function.

Proposition 6. The joint distance cumulant $cum(X_{i_1}, \dots, X_{i_s})$ is given by the coefficient of $\imath^s \prod_{k=1}^{s} t_{i_k}$ in the Taylor expansion of $\log\left\{dcf(t_1, \dots, t_d)\right\}$, where $\{i_1, \dots, i_s\}$ is any subset of $\{1, 2, \dots, d\}$ with $s \leq d$.

Our next result indicates that the mutual independence among $\{X_1, \dots, X_d\}$ is equivalent to the mutual independence among $\{U_1(X_1, X_1'), \dots, U_d(X_d, X_d')\}$.

Proposition 7. The random variables $\{X_1, \dots, X_d\}$ are mutually independent if and only if $\operatorname{dcf}(t_1, \dots, t_d) = \prod_{i=1}^{d} \operatorname{dcf}(t_i)$ for t_i almost everywhere, where $\operatorname{dcf}(t_i) = \mathbb{E}[\exp\{\imath t_i U_i(X_i, X_i')\}]$.

2.4. *Rank-Based Metrics*

In this subsection, we briefly discuss the concept of rank-based distance measures (Table 1). For simplicity, we assume that X_i's are all univariate and remark that our definition can be generalized to the case where X_i's are random vectors without essential difficulty. The basic idea here is to apply the monotonic transformation based on the marginal distribution functions to each X_j, and then use the dCov or JdCov to quantify the interaction and joint dependence of the coordinates after transformation. Therefore, it can be viewed as the counterpart of Spearman's rho to dCov or JdCov.

Let F_j be the marginal distribution function for X_j. The squared rank dCov and JdCov among $\{X_1, \dots, X_d\}$ are defined, respectively, as

$$\operatorname{dCov}_R^2(X_1, \dots, X_d) = \operatorname{dCov}^2(F_1(X_1), \dots, F_d(X_d)),$$
$$\operatorname{JdCov}_R^2(X_1, \dots, X_d; c) = \operatorname{JdCov}^2(F_1(X_1), \dots, F_d(X_d); c).$$

The rank-based dependence metrics enjoy a few appealing features: (1) they are invariant to monotonic component wise transformations; (2) they are more robust to outliers and heavy tail of the distribution; (3) their existence require very weak moment assumption on the components of $\mathcal{X}$. In Section 5, we shall compare the finite-sample performance of JdCov_R^2 with that of JdCov and JdCov$_S$.

3. Estimation

We now turn to the estimation of the joint dependence metrics. Given n samples $\{\mathbf{X}_j\}_{j=1}^n$ with $\mathbf{X}_j = (X_{j1}, \dots, X_{jd})$, we consider the plug-in estimators based on the V-statistics as well as their bias-corrected versions to be described below. Denote by $\hat{f}_i(t_i) = n^{-1}\sum_{j=1}^{n} e^{\imath\langle t_i, X_{ji}\rangle}$ the empirical characteristic function for X_i.

3.1. Plug-In Estimators

For $1 \leq k, l \leq n$, let $\widehat{U}_i(k,l) = n^{-1}\sum_{v=1}^n |X_{ki} - X_{vi}| + n^{-1}\sum_{u=1}^n |X_{ui} - X_{li}| - |X_{ki} - X_{li}| - n^{-2}\sum_{u,v=1}^n |X_{ui} - X_{vi}|$ be the sample estimate of $U_i(X_{ki}, X_{li})$. The V-statistic-type estimators for dCov, JdCov, and its scale-invariant version are defined, respectively, as

$$\widehat{\mathrm{dCov}^2}(X_1,\dots,X_d) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \widehat{U}_i(k,l)^2, \tag{10}$$

$$\widehat{\mathrm{JdCov}^2}(X_1,\dots,X_d;c)) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \left(\widehat{U}_i(k,l) + c\right) - c^d, \tag{11}$$

$$\widehat{\mathrm{JdCov}_S^2}(X_1,\dots,X_d;c) = \frac{1}{n^2}\sum_{k,l=1}^n \prod_{i=1}^d \left(\frac{\widehat{U}_i(k,l)}{\widehat{\mathrm{dCov}}(X_i)} + c\right) - c^d, \tag{12}$$

where $\widehat{\mathrm{dCov}^2}(X_i) = n^{-2}\sum_{k,l=1}^n \widehat{U}_i(k,l)^2$ is the sample (squared) dCov. The following lemma shows that the V-statistic-type estimators are equivalent to the plug-in estimators by replacing the characteristic functions and the expectation in the definitions of dCov and JdCov with their sample counterparts.

Lemma 2. The sample (squared) dCov can be rewritten as

$$\widehat{\mathrm{dCov}^2}(X_1,\dots,X_d) = \int_{\mathbb{R}^{p_0}} \left| \frac{1}{n}\sum_{k=1}^n \left[\prod_{i=1}^d \left(\hat{f}_i(t_i) - e^{\imath\langle t_i, X_{ki}\rangle}\right)\right]\right|^2 dw. \tag{13}$$

Moreover, we have

$$\begin{aligned}
&\widehat{\mathrm{JdCov}^2}(X_1,\dots,X_d;c) \\
&= c^{d-2}\sum_{(i_1,i_2)\in I_2^d} \widehat{\mathrm{dCov}^2}(X_{i_1},X_{i_2}) + c^{d-3} \\
&\quad \times \sum_{(i_1,i_2,i_3)\in I_3^d} \widehat{\mathrm{dCov}^2}(X_{i_1},X_{i_2},X_{i_3}) \\
&\quad + \cdots + \widehat{\mathrm{dCov}^2}(X_1,\dots,X_d).
\end{aligned} \tag{14}$$

Remark 1. Consider the univariate case where $p_i = 1$ for all $1 \leq i \leq d$. Let $\widehat{F}_i$ be the empirical distribution based on $\{X_{ji}\}_{j=1}^n$ and define $Z_{ji} = \widehat{F}_i(X_{ji})$. Then, the rank-based metrics defined in Section 2.4 can be estimated in a similar way by replacing X_{ji} with Z_{ji} in the definitions of the above estimators.

Remark 2. The distance cumulant can be estimated by

$$\begin{aligned}
\widehat{\mathrm{cum}}(X_1,\dots,X_d) &= \sum_{\pi}(-1)^{|\pi|-1}(|\pi|-1)! \prod_{D\in\pi} \\
&\quad \times \left\{\frac{1}{n^2}\sum_{k,l=1}^n \left(\prod_{i\in D}\widehat{U}_i(k,l)\right)\right\}.
\end{aligned}$$

However, the combinatorial nature of distance cumulant implies that detecting interactions of higher order requires significantly more costly computation.

We study the asymptotic properties of the V-statistic-type estimators under suitable moment assumptions.

Assumption 1. Suppose for any subset S of $\{1, 2, \dots, d\}$ with $|S| \geq 2$, there exists a partition $S = S_1 \cup S_2$ such that $\mathbb{E}\prod_{i\in S_1}|X_i| < \infty$ and $\mathbb{E}\prod_{i\in S_2}|X_i| < \infty$.

Proposition 8. Under Assumption 1 , we have as $n \to \infty$,

$$\widehat{\mathrm{dCov}^2}(X_1,\dots,X_d) \xrightarrow{a.s} \mathrm{dCov}^2(X_1,\dots,X_d),$$
$$\widehat{\mathrm{JdCov}^2}(X_1,\dots,X_d;c) \xrightarrow{a.s} \mathrm{JdCov}^2(X_1,\dots,X_d;c),$$
$$\widehat{\mathrm{JdCov}_S^2}(X_1,\dots,X_d;c) \xrightarrow{a.s} \mathrm{JdCov}_S^2(X_1,\dots,X_d;c),$$

where " $\xrightarrow{\text{a.s}}$ " denotes the almost sure convergence.

When $d = 2$, Assumption 8 reduces to the condition that $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}|X_2| < \infty$ in Theorem 2 of Székely, Rizzo, and Bakirov (2007). Suppose X_i's are mutually independent. Then Assumption 8 is fulfilled provided that $\mathbb{E}|X_i| < \infty$ for all i. More generally, if $E|X_i|^{\lfloor(d+1)/2\rfloor} < \infty$ for $1 \leq i \leq d$, then Assumption 8 is satisfied.

Let $\Gamma(\cdot)$ denote a complex-valued zero mean Gaussian random process with the covariance function $R(t,t') = \prod_{i=1}^d \left(f_i(t_i - t_i') - f_i(t_i)f_i(-t_i')\right)$, where $t = (t_1, t_2, \dots, t_d)$, $t' = (t_1', t_2', \dots, t_d') \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times \cdots \times \mathbb{R}^{p_d}$.

Proposition 9. Suppose $X_1, X_2, \dots, X_d$ are mutually independent, and $\mathbb{E}|X_i| < \infty$ for $1 \leq i \leq d$. Then we have

$$n\widehat{\mathrm{dcov}^2}(X_1,X_2,\dots,X_d) \xrightarrow{d} \|\Gamma\|^2 = \sum_{j=1}^{+\infty}\lambda_j Z_j^2,$$

where $||\Gamma||^2 = \int \Gamma(t_1,t_2,\dots,t_d)^2 dw$, $Z_j \overset{\text{iid}}{\sim} N(0,1)$ and $\lambda_j > 0$ depends on the distribution of $\mathcal{X}$. As a consequence, we have

$$n\widehat{\mathrm{Jdcov}^2}(X_1,X_2,\dots,X_d;c) \xrightarrow{d} \sum_{j=1}^{+\infty}\lambda_j' Z_j^2,$$

with $\lambda_j' > 0$ and $Z_j \overset{\text{iid}}{\sim} N(0,1)$.

Proposition 9 shows that both $\widehat{\mathrm{dcov}^2}$ and $\widehat{\mathrm{Jdcov}^2}$ converge to weighted sum of chi-squared random variables, where the weights depend on the marginal characteristic functions in a complicated way. Since the limiting distribution is nonpivotal, we will introduce a bootstrap procedure to approximate their sampling distributions in the next section.

It has been pointed out in the literature that the computational complexity of dCov is $O(n^2)$ if it is implemented directly according to its definition. The computational cost of the V-statistic-type estimators and the bias-corrected estimators for JdCov are both of the order $O(n^2 p_0)$.

3.2. Bias-Corrected Estimators

It is well known that V-statistic leads to biased estimation. To remove the bias, one can construct an estimator for the dth-order dCov based on a dth-order U-statistic. However, the computational complexity for the dth-order U-statistic is of the order $O(dn^d)$, which is computationally prohibitive when n and d are both large. Adopting the $\mathcal{U}$-centering idea in Székely and Rizzo (2014), we propose bias-corrected estimators which do not bring extra computational cost as compared to the plug-in estimators. Specifically, for $1 \le i \le d$, we define the $\mathcal{U}$-centered version of $|X_{ki} - X_{li}|$ as

$$\begin{aligned}\widetilde{U}_i(k,l) = &\frac{1}{n-2}\sum_{u=1}^{n}|X_{ui} - X_{li}| \\ &+ \frac{1}{n-2}\sum_{v=1}^{n}|X_{ki} - X_{vi}| - |X_{ki} - X_{li}| \\ &- \frac{1}{(n-1)(n-2)}\sum_{u,v=1}^{n}|X_{ui} - X_{vi}|\end{aligned}$$

when $k \neq l$, and $\widetilde{U}_i(k,l) = 0$ when $k = l$. One can verify that $\sum_{v \neq k}\widetilde{U}_i(k,v) = \sum_{u \neq l}\widetilde{U}_i(u,l) = 0$, which mimics the double-centered property $\mathbb{E}[U_i(X_i, X_i')|X_i] = \mathbb{E}[U_i(X_i, X_i')|X_i'] = 0$ for its population counterpart. Let $\widetilde{\mathrm{dCov}}^2(X_i, X_j) = \sum_{k \neq l}\widetilde{U}_i(k,l)\widetilde{U}_j(k,l)/\{n(n-3)\}$ and write $\widetilde{\mathrm{dCov}}(X_i) = \widetilde{\mathrm{dCov}}(X_i, X_i)$. We define the bias-corrected estimators as

$$\begin{aligned}&\widetilde{\mathrm{JdCov}^2}(X_1,\dots,X_d;c) \\ &= \frac{1}{n(n-3)}\sum_{k,l=1}^{n}\prod_{i=1}^{d}\left(\widetilde{U}_i(k,l) + c\right) - \frac{n}{n-3}c^d, \\ &\widetilde{\mathrm{JdCov}_S^2}(X_1,\dots,X_d;c) \\ &= \frac{1}{n(n-3)}\sum_{k,l=1}^{n}\prod_{i=1}^{d}\left(\frac{\widetilde{U}_i(k,l)}{\widetilde{\mathrm{dCov}}(X_i)} + c\right) - \frac{n}{n-3}c^d.\end{aligned}$$

Direct calculation yields that

$$\begin{aligned}\widetilde{\mathrm{JdCov}^2}(X_1,\dots,X_n;c) = c^{d-2}\sum_{(i,j)\in I_2^d}\widetilde{\mathrm{dCov}}^2(X_i,X_j) \\ + \text{higher-order terms}. \qquad (15)\end{aligned}$$

It has been shown in Proposition 1 of Székely and Rizzo (2014) that $\widetilde{\mathrm{dCov}}^2(X_i, X_j)$ is an unbiased estimator for $\mathrm{dCov}^2(X_i, X_j)$. In the supplementary material, we provide an alternative proof which simplifies the arguments in Székely and Rizzo (2014). Our argument relies on a new decomposition of $\widetilde{U}_i(k,l)$, which provides some insights on the $\mathcal{U}$-centering idea. See Lemma 1.1 and Proposition 1.2 in the supplementary material. In view of (15) and Proposition 1.2, the main effect in $\mathrm{JdCov}^2(X_1,\dots,X_n;c)$ can be unbiasedly estimated by the main effect of $\widetilde{\mathrm{JdCov}^2}(X_1,\dots,X_n;c)$. However, it seems very challenging to study the impact of $\mathcal{U}$-centering on the bias of the high-order effect terms. We shall leave this problem to our future research.

4. Testing for Joint Independence

In this section, we consider the problem of testing the null hypothesis

$$H_0 : X_1,\dots,X_d \text{ are mutually independent} \qquad (16)$$

against the alternative H_A : negation of H_0. For the purpose of illustration, we use $n\widehat{\mathrm{JdCov}^2}$ as our test statistic and set

$$\phi_n(\mathbf{X}_1,\dots,\mathbf{X}_n) := \begin{cases} 1 & \text{if} \quad n\widehat{\mathrm{JdCov}^2}(X_1,\dots,X_d) > c_n\,, \\ 0 & \text{if} \quad n\widehat{\mathrm{JdCov}^2}(X_1,\dots,X_d) \le c_n\,, \end{cases} \qquad (17)$$

where the threshold c_n remains to be chosen. Consequently, we define a decision rule as follows: reject H_0 if $\phi_n = 1$ and fail to reject H_0 if $\phi_n = 0$.

Below we introduce a bootstrap procedure to approximate the sampling distribution of $n\widehat{\mathrm{JdCov}}$ under H_0. Let $\widehat{F}_i$ be the empirical distribution function based on the data points $\{X_{ji}\}_{j=1}^n$. Conditional on the original sample, we define $\mathbf{X}_j^* = (X_{j1}^*,\dots,X_{jd}^*)$, where X_{ji}^* are generated independently from $\widehat{F}_i$ for $1 \le i \le d$. Let $\{\mathbf{X}_j^*\}_{j=1}^n$ be n bootstrap samples. Then we can compute the bootstrap statistics $\widehat{\mathrm{dCov}^2}^*$ and $\widehat{\mathrm{JdCov}^2}^*$ in the same way as $\widehat{\mathrm{dCov}^2}$ and $\widehat{\mathrm{JdCov}^2}$ based on $\{\mathbf{X}_j^*\}_{j=1}^n$. In particular, we note that the bootstrap version of the dth-order dCov is given by

$$n\widehat{\mathrm{dCov}^2}^*(X_1,\dots,X_d) = \|\Gamma_n^*\|^2 = \int \Gamma_n^*(t_1,\dots,t_d)^2\,dw,$$

where

$$\Gamma_n^*(t) = n^{-1/2}\sum_{j=1}^{n}\prod_{i=1}^{d}(\hat{f}_i^*(t_i) - e^{\imath\langle t_i, X_{ji}^*\rangle}).$$

Denote by "$\xrightarrow{d^*}$" the weak convergence in the bootstrap world conditional on the original sample $\{\mathbf{X}_j\}_{j=1}^n$.

Proposition 10. Suppose $\mathbb{E}|X_i| < \infty$ for $1 \le i \le d$. Then

$$n\widehat{\mathrm{dCov}^2}^*(X_1,\dots,X_d) \xrightarrow{d^*} \sum_{j=1}^{+\infty}\lambda_j Z_j^2,$$

$$n\widehat{\mathrm{JdCov}^2}^*(X_1,\dots,X_d) \xrightarrow{d^*} \sum_{j=1}^{+\infty}\lambda_j' Z_j^2,$$

almost surely as $n \to \infty$.

Proposition 10 shows that the bootstrap statistic is able to imitate the limiting distribution of the test statistic. Thus, we shall choose c_n to be the $1-\alpha$ quantile of the distribution of $n\widehat{\mathrm{JdCov}^2}^*$ conditional on the sample $\{\mathbf{X}_j\}_{j=1}^n$. The validity of the bootstrap-assisted test can be justified as follows.

Proposition 11. For all $\alpha \in (0,1)$, the α-level bootstrap-assisted test has asymptotic level α when testing H_0 against H_A. In other words, under H_0, $\limsup_{n\to\infty} P\,(\,\phi_n(\mathbf{X}_1,\dots,\mathbf{X}_n) = 1\,) = \alpha$.

Proposition 12. For all $\alpha \in (0,1)$, the α-level bootstrap-assisted test is consistent when testing H_0 against H_A. In other words, under H_A, $\lim_{n\to\infty} P\,(\,\phi_n(\mathbf{X}_1,\dots,\mathbf{X}_n) = 1\,) = 1.$

5. Numerical Studies

We investigate the finite-sample performance of the proposed methods. Our first goal is to test the joint independence among the variables $\{X_1, \dots, X_d\}$ using the new dependence metrics, and compare the performance with some existing alternatives in the literature in terms of size and power. Throughout the simulation, we set $c = 0.5, 1, 2$ in JdCov and implement the bootstrap-assisted test based on the bias-corrected estimators. We compare our tests with the dHSIC-based test in Pfister et al. (2018) and the mutual independence test proposed in Matteson and Tsay (2017), which is defined as

$$T_{MT} := \sum_{i=1}^{d-1} \mathrm{dCov}^2(X_i, X_{(i+1):d}), \tag{18}$$

where $X_{(i+1):d} = \{X_{i+1}, X_{i+2}, \dots, X_d\}$. We consider both Gaussian and non-Gaussian distributions and study the following models, motivated from Sejdinovic, Gretton, and Bergsma (2013) and Yao, Zhang, and Shao (2018).

Example 1 (Gaussian copula model). The data $\mathbf{X} = (X_1, \dots, X_d)$ are generated as follows:

1. $\mathbf{X} \sim N(0, I_d)$;
2. $\mathbf{X} = Z^{1/3}$ and $Z \sim N(0, I_d)$;
3. $\mathbf{X} = Z^3$ and $Z \sim N(0, I_d)$.

Example 2 (Multivariate Gaussian model). The data $\mathbf{X} = (X_1, \dots, X_d)$ are generated from the multivariate normal distribution with the following three covariance matrices $\Sigma = (\sigma_{ij}(\rho))_{i,j=1}^d$ with $\rho = 0.25$:

1. AR(1): $\sigma_{ij} = \rho^{|i-j|}$ for all $i, j \in \{1, \dots, d\}$;
2. Banded: $\sigma_{ii} = 1$ for $i = 1, \dots, d$; $\sigma_{ij} = \rho$ if $1 \le |i-j| \le 2$ and $\sigma_{ij} = 0$ otherwise;
3. Block: Define $\Sigma_{\text{block}} = (\sigma_{ij})_{i,j=1}^5$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ if $i \neq j$. Let $\Sigma = I_{\lfloor d/5 \rfloor} \otimes \Sigma_{\text{block}}$, where $\otimes$ denotes the Kronecker product.

Example 3. The data $\mathbf{X} = (X, Y, Z)$ are generated as follows:

1. $X, Y \overset{\text{iid}}{\sim} N(0,1)$, $Z = \text{sign}(XY)\, W$, where W follows an exponential distribution with mean $\sqrt{2}$;
2. X, Y are independent Bernoulli random variables with the success probability 0.5, and $Z = \mathbf{1}\{X = Y\}$.

Example 4. In this example, we consider a triplet of random vectors (X, Y, Z) on $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$, with $X, Y \overset{\text{iid}}{\sim} N(0, I_p)$. We focus on the following cases:

1. $Z_1 = \text{sign}(X_1 Y_1)\, W$ and $Z_{2:p} \sim N(0, I_{p-1})$, where W follows an exponential distribution with mean $\sqrt{2}$;
2. $Z_{2:p} \sim N(0, I_{p-1})$ and

$$Z_1 = \begin{cases} X_1^2 + \epsilon, & \text{with probability } 1/3, \\ Y_1^2 + \epsilon, & \text{with probability } 1/3, \\ X_1 Y_1 + \epsilon, & \text{with probability } 1/3, \end{cases}$$

where $\epsilon \sim U(-1, 1)$.

We conduct tests for joint independence among the random variables described in the above examples. For each example, we draw 1000 simulated datasets and perform tests of joint independence with 500 bootstrap resamples. We try small and moderate sample sizes, that is, $n = 50, 100$, or 200. Figures 1 and 2 display the proportion of rejections (out of 1000 simulation runs) for the five different tests, based on the statistics $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}_S^2$, $\widetilde{\mathrm{JdCov}}_R^2$, dHSIC, and T_{MT}. The detailed figures are reported in Tables 1 and 2 in the supplementary materials.

In Example 1, the data-generating scheme suggests that the variables are jointly independent. The plots in Figure 1 show that all the five tests perform more or less equally well in Examples 1.1 and 1.2, and the rejection probabilities are quite close to the 10% or 5% nominal level. In Example 1.3, the tests based on our proposed statistics show greater conformation of the empirical size to the actual size of the test than T_{MT}. In Example 2, the tests based on $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}_S^2$, and $\widetilde{\mathrm{JdCov}}_R^2$ as well as T_{MT} significantly outperform the dHSIC-based test. Note that the empirical power becomes higher when c increases to 2. From Figure 2, we observe that in Example 3 all the tests perform very well in the second case. However, in the first case, our tests and the dHSIC-based test deliver higher power as compared to T_{MT}. Finally, in Example 4, we allow X, Y, Z to be random vectors with dimension $p = 5, 10$. The rejection probabilities for each of the five tests increase with n, and the proposed tests provide better performances in comparison with the other two competitors. In particular, the test based on $\widetilde{\mathrm{JdCov}}_S^2$ outperforms all the others in a majority of the cases. In Examples 3 and 4, the power becomes higher when c decreases to 0.5. These results are consistent with our statistical intuition and the discussions in Section 2.2. For the Gaussian copula model, only the main effect term matters, so a larger c is preferable. For non-Gaussian models, the high-order terms kick in and hence a smaller c may lead to higher power.

We have considered U-statistic-type estimators of JdCov^2, JdCov_S^2, and JdCov_R^2 so far in all the above computations, as they remove the bias due to the main effects (see Section 3.2). However, it might be interesting to see if the bias correction has any empirical impact. We conduct tests for joint independence of the random variables in some of the above examples, this time using the V-statistic-type estimators (described in Section 3.1). Table 4 (in the supplementary materials) shows the proportion of rejections (out of 1000 simulation runs) for the tests based on $\widehat{\mathrm{JdCov}}^2$, $\widehat{\mathrm{JdCov}}_S^2$, and $\widehat{\mathrm{JdCov}}_R^2$, setting $c = 1$. The results indicate that use of the bias-corrected estimators lead to greater conformation of the empirical size to the actual size of the test (in Example 1), and slightly better power in Example 3.

In connection to the heuristic idea discussed in Section 2.2 about choosing the tuning parameter c, we conduct tests for joint independence of the random variables in all the above examples, choosing c in that way. Table 4 (in the supplementary materials) presents the proportion of rejections for the proposed tests and the values of c for each example, averaged over the 1000 simulated datasets. The plots in Figures 1 and 2 reveal some interesting features. In Example 2, we have Gaussian data, so a larger c is preferable. Clearly the proportion of rejections is a little higher (or lower) in most of the cases when we choose

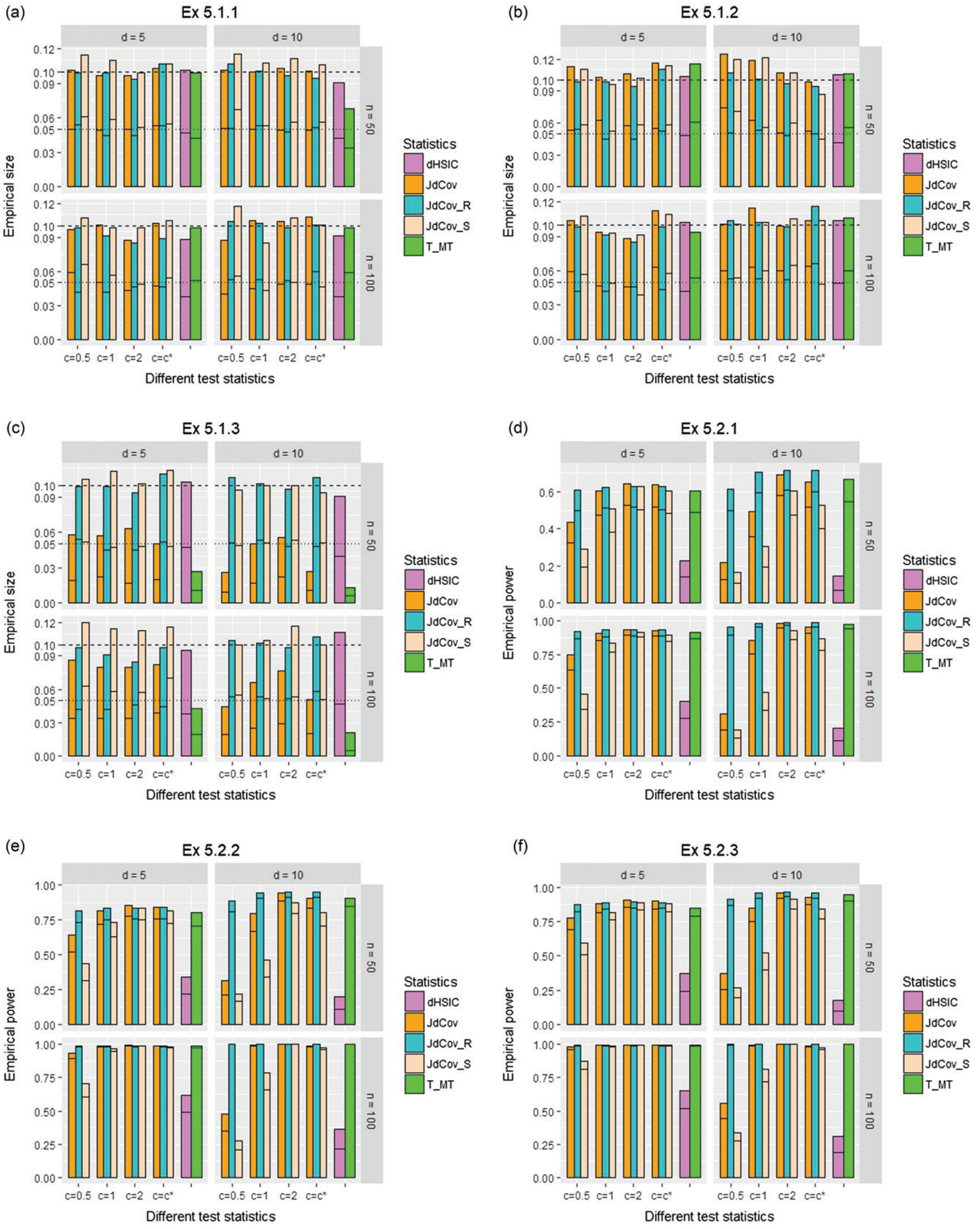


Figure 1. Figures showing the empirical size and power for the different tests statistics in Examples 1 and 2. c^* denotes the data-driven choice of c. The vertical height of a bar and a line on a bar stand for the empirical size or power at levels $\alpha = 0.1$ or $\alpha = 0.05$, respectively.

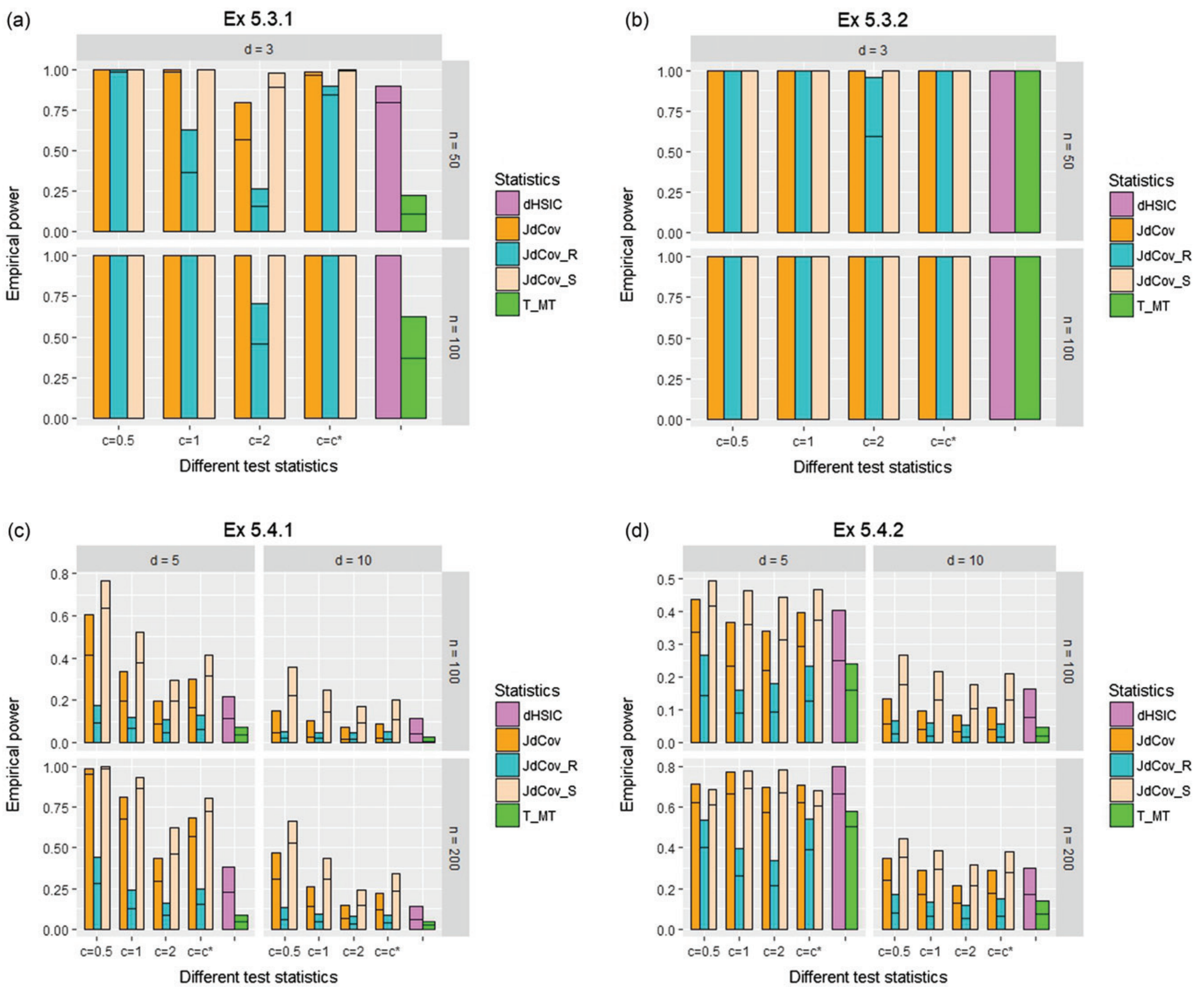


Figure 2. Figures showing the empirical power for the different tests statistics in Examples 3 and 4. c^* denotes the data-driven choice of c. The vertical height of a bar and a line on a bar stand for the empirical power at levels $\alpha = 0.1$ or $\alpha = 0.05$, respectively.

c in the data-driven way (c turns out to be around 1.6 or 1.7), than when c is subjectively chosen to be 0.5 (or 2). On the contrary, in Example 3, the data are non-Gaussian and a smaller c is preferable. Evidently choosing c in the data-driven way leads to nearly equally good power compared to when $c = 0.5$, and higher power compared to when $c = 2$.

6. Application to Causal Inference

6.1. Model Diagnostic Checking for Directed Acyclic Graph (DAG)

We employ the proposed metrics to perform model selection in causal inference which is based on the joint independence testing of the residuals from the fitted structural equation models. Specifically, given a candidate DAG $\mathcal{G}$, we let $\mathrm{Par}(j)$ denote the index associated with the parents of the jth node. Following Peters et al. (2014) and Bühlmann, Peters, and Ernest (2014), we consider the structural equation models with additive components

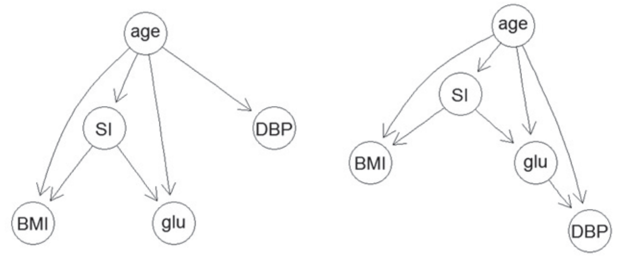
$$X_j = \sum_{k \in \mathrm{Par}(j)} f_{j,k}(X_k) + \epsilon_j\,,\ j = 1, 2, \ldots, d, \tag{19}$$

where the noise variables $\epsilon_1, \ldots, \epsilon_d$ are jointly independent variables. Given n observations $\{\mathbf{X}_i\}_{i=1}^n$ with $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$, we use generalized additive regression (Wood and Augustin 2002) to regress X_j on all its parents $\{X_k, k \in \mathrm{Par}(j)\}$ and denote the resulting residuals by

$$\hat{\epsilon}_{ij} = X_{ij} - \sum_{k \in \mathrm{Par}(j)} \hat{f}_{j,k}(X_{ik}), \quad 1 \le j \le d, \quad 1 \le i \le n,$$

where $\hat{f}_{j,k}$ is the B-spline estimator for $f_{j,k}$. To check the goodness of fit of $\mathcal{G}$, we test the joint independence of the residuals. Let T_n be the statistic (e.g., $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}_S^2$, or $\widetilde{\mathrm{JdCov}}_R^2$) to test the joint dependence of $(\epsilon_1, \ldots, \epsilon_d)$ constructed based on the fitted residuals $\hat{\epsilon}_i = (\hat{\epsilon}_{i1}, \ldots, \hat{\epsilon}_{id})$ for $1 \le i \le n$. Following the idea presented in Sen and Sen (2014), it seems that T_n might have a limiting distribution different from the one mentioned in Proposition 9. So to approximate the sampling distribution of T_n, we introduce the following residual bootstrap procedure.

1. Randomly sample $\epsilon_j^* = (\epsilon_{1j}^*, \ldots, \epsilon_{nj}^*)$ with replacement from the residuals $\{\hat{\epsilon}_{1j}, \ldots, \hat{\epsilon}_{nj}\}$, $1 \le j \le d$. Construct the bootstrap sample $X_{ij}^* = \sum_{k \in \mathrm{Par}(j)} \hat{f}_{j,k}(X_{ik}) + \epsilon_{ij}^*$.



(a) $\widetilde{JdCov}^2$, $\widetilde{JdCov}^2_S$, $\widetilde{JdCov}^2_R$ and T_{MT}

(b) dHSIC

Figure 3. The DAG models corresponding to the largest p-values from the five tests.

2. Based on the bootstrap sample $\{\mathbf{X}_i^*\}_{i=1}^n$ with $\mathbf{X}_i^* = (X_{i1}^*, \dots, X_{id}^*)$, estimate $f_{j,k}$ for $k \in \mathrm{Par}(j)$, and denote the corresponding residuals by $\hat{\epsilon}_{ij}^*$.
3. Calculate the bootstrap statistic T_n^* based on $\{\hat{\epsilon}_{ij}^*\}$.
4. Repeat the above steps B times and let $\{T_{b,n}^*\}_{b=1}^B$ be the corresponding values of the bootstrap statistics. The p-value is given by $B^{-1}\sum_{b=1}^B \{T_{b,n}^* > T_n\}$.

Pfister et al. (2018) proposed to bootstrap the residuals directly and used the bootstrapped residuals to construct the test statistic. In contrast, we suggest the use of the above residual bootstrap to capture the estimation effect caused by replacing $f_{j,k}$ with the estimate $\hat{f}_{j,k}$.

6.2. Real Data Example

We now apply the model diagnostic checking procedure for DAG to one real world dataset. A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the U.S. National Institute of Diabetes and Digestive and Kidney Diseases. We downloaded the data from *https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes*. We focus only on the following five variables: Age, Body Mass Index (BMI), 2-Hour Serum Insulin (SI), Plasma Glucose Concentration (glu), and Diastolic Blood Pressure (DBP). Further, we only selected the instances with nonzero values, as it seems that zero values encode missing data. This yields $n = 392$ samples.

Now, age is likely to affect all the other variables (but of course not the other way round). Moreover, serum insulin also has plausible causal effects on BMI and plasma glucose concentration. We try to determine the correct causal structure out of 48 candidate DAG models and perform model diagnostic checking for each of the 48 models, as illustrated in Section 6.1. We first center each of the variables and scale them so that l_2 norm of each of the variables is $\sqrt{n}$. We perform the mutual independence test of residuals based on the statistics $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}^2_S$, and $\widetilde{\mathrm{JdCov}}^2_R$ with $c = 1$, and compare with the bootstrap-assisted version of the dHSIC-based test proposed in Pfister et al. (2018) and T_{MT}. For each of the tests, we implement the residual bootstrap to obtain the p-value with $B = 1000$. Figure 3 shows the selected DAG models corresponding to the largest p-values from each of the five tests. A directed edge from one variable to another indicates a causal influence of the former on the latter.

Figure 3(a) shows the model with the maximum p-value among all the 48 candidate DAG models, when the test for joint independence of the residuals is conducted based on $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}^2_S$, and $\widetilde{\mathrm{JdCov}}^2_R$ and T_{MT}. This graphical structure goes in tune with the biological evidences of causal relationships among these five variables. Figure 3(b) stands for the model with the maximum p-value when the test is based on dHSIC. Its only difference with Figure 3(a) is that, it has an additional edge from glu to DBP, indicating a causal effect of Plasma Glucose Concentration on Diastolic Blood Pressure. We are unsure of any biological evidence that supports such a causal relationship in reality.

It might be intriguing to take into account the heuristic data-driven way of determining c (see Section 2.2) in the above example, instead of setting c at a default value of 1. Our findings indicate that choosing c in the data-driven way leads to a slightly different result. The tests based on dHSIC and $\widetilde{\mathrm{JdCov}}^2_S$ select the DAG model shown in Figure 3(b) (considering the maximum p-value among all the 48 candidate DAG models), whereas Figure 3(a) is the DAG model selected when the test is based on $\widetilde{\mathrm{JdCov}}^2$, $\widetilde{\mathrm{JdCov}}^2_R$, and T_{MT}. The proposed tests (based on $\widetilde{\mathrm{JdCov}}^2$ and $\widetilde{\mathrm{JdCov}}^2_R$) still perform well.

6.3. A Simulation Study

We conduct a simulation study based on our findings in the previous real data example. To save the computational cost, we focus our attention on three of the five variables, viz., age, glu, and DBP. In the correct causal structure among these three variables, there are directed edges from age to glu and age to DBP. We consider the additive structural equation models

$$X_j = \sum_{k \in \mathrm{Par}(j)} \hat{f}_{j,k}(X_k) + e_j\,,\ j = 1, 2, 3, \tag{20}$$

where X_1, X_2, X_3 correspond to age, glu, and DBP (after centering and scaling), respectively, and $\hat{f}_{j,k}$ denotes the estimated function from the real data. Note that X_1 is the only variable without any parent. In Section 6.2, we get from our numerical studies that the standard deviation of X_1 is 1.001, and the standard deviations of the residuals when X_2 and X_3 are regressed on X_1 (according to the structural equation models in (19)), are 0.918 and 0.95, respectively. In this simulation study, we simulate X_1 from a zero mean Gaussian distribution with standard deviation 1. For X_2 and X_3, we simulate the noise variables from zero mean Gaussian distributions with standard deviations 0.918 and 0.95, respectively. The same $n = 392$ is considered for the number of generated observations, and based on this simulated dataset we perform the model diagnostic checking for 27 candidate DAG models. The number of bootstrap replications is set to be $B = 100$ (to save the computational cost). This procedure is repeated 100 times to note how many times out of 100 that the five tests select the correct model, based on the largest p-value. The results in Table 2 indicate that the proposed tests with $c = 1$ and the dHSIC-based test outperform T_{MT}.

A natural question to raise is why do we bootstrap the residuals and not test for the joint independence of the estimated

Table 2. The number of times (out of 100) that the true model is being selected.

$\widetilde{\mathrm{JdCov}^2}$	$\widetilde{\mathrm{JdCov}_S^2}$	$\widetilde{\mathrm{JdCov}_R^2}$	dHSIC	T_{MT}
45	61	54	52	32

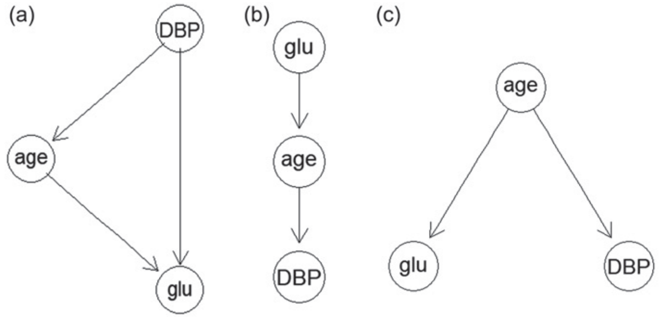


Figure 4. The DAG models selected (most frequently out of 100 times) by the five tests, without doing residual bootstrap to reestimate $f_{j,k}$.

residuals directly, to check for the goodness of fit of the DAG model. From the idea in Sen and Sen (2014), it appears that the joint distance covariance of the estimated residuals might have a limiting distribution different from the one stated in Proposition 9. We leave the formulation of a rigorous theory in support of that for future research. We present below the models selected most frequently (out of 100 times) by the different test statistics if we repeat the simulation study done above in Section 6.3 without using residual bootstrap to reestimate $f_{j,k}$. We immediately see that joint independence tests of the estimated residuals based on all of the five statistics we consider, select a DAG model that is meaningless and far away from the correct one.

It might be intriguing to take into account the heuristic data-driven way of choosing c (see Section 2.2) in the simulation study in Section 6.3, instead of setting c at a default value of 1. Our findings indicate that our proposed tests and the dHSIC-based test still outperform T_{MT}. If we repeat the simulation study done in Section 6.3 (choosing c in the heuristic way) without using residual bootstrap to reestimate $f_{j,k}$, we still end up selecting the same models presented in Figure 4.

7. Future Research

Huo and Székely (2016) proposed an $O(n \log n)$ algorithm to compute dCov of univariate random variables. In a more recent work, Huang and Huo (2017) introduced a fast method for multivariate cases which is based on random projection and has computational complexity $O(nK \log n)$, where K is the number of random projections. One of the possible directions for future research is to come up with a fast algorithm to compute JdCov. When $p_i = 1$, we can indeed use the method in Huo and Székely (2016) to compute JdCov. But their method may be inefficient when d is large and it is not applicable to the case where $p_i > 1$. Another direction is, to introduce the notion of Conditional JdCov in light of Wang et al. (2015), to test if the variables $(X_1, \dots, X_d)$ are jointly independent given another variable Z.

Supplementary Materials

The supplementary materials contain some additional proofs of the main results in the article and tabulated numerical results from Section 5.

Acknowledgments

The authors are grateful to the editor and two very careful reviewers for their immensely helpful comments and suggestions that greatly improved the article. It would also be authors' pleasure to thank Niklas Pfister for sharing with them the R codes used in Pfister et al. (2018).

Funding

The authors acknowledge partial funding from the NSF grant DMS-1607320.

References

Bach, F. R., and Jordan, M. I. (2002), "Kernel Independent Component Analysis," *Journal of Machine Learning Research*, 3, 1–48. [1638]

Baringhaus, L., and Henze, N. (1988), "A Consistent Test for Multivariate Normality Based on the Empirical Characteristic Function," *Metrika*, 35, 339–348. [1641]

Bergsma, W., and Dassios, A. (2014), "A Consistent Test of Independence Based on a Sign Covariance Related to Kendall's Tau," *Bernoulli*, 20, 1006–1028. [1638]

Bowman, A. W., and Foster, P. J. (1993), "Adaptive Smoothing and Density-Based Tests of Multivariate Normality," *Journal of the American Statistical Association*, 88, 529–537. [1641]

Bühlmann, P., Peters, J., and Ernest, J. (2014), "CAM : Causal Additive Models, High-Dimensional Order Search and Penalized Regression," *The Annals of Statistics*, 42, 2526–2556. [1647]

Cover, T. M., and Thomas, J. A. (1991), "*Elements of Information Theory*, New York: Wiley. [1638]

Dueck, J., Edelmann, D., Gneiting, T., and Richards, D. (2014), "The Affinely Invariant Distance Correlation," *Bernoulli*, 20, 2305–2330. [1638]

Fano, R. M. (1961), *Transmission of Information*, Cambridge, MA: MIT Press. [1638]

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005), "Measuring Statistical Dependence with Hilbert-Schmidt Norms," in *Algorithmic Learning Theory*, eds. S. Jain, H. U. Simon, and E. Tomita, Berlin: Springer-Verlag, pp. 63–77. [1638]

Gretton, A., Fukumizu, C. H. Teo., Song, L., Schölkopf, B., and Smola, A. (2007), "A Kernel Statistical Test of Independence," *Advances in Neural Information Processing Systems*, 20, 585–592. [1638]

Henze, N., and Wagner, T. (1997), "A New Approach to the BHEP Tests for Multivariate Normality," *Journal of Multivariate Analysis*, 62, 1–23. [1641]

Huang, C., and Huo, X. (2017), "A Statistically and Numerically Efficient Independence Test Based on Random Projections and Distance Covariance." arXiv:1701.06054. [1649]

Huo, X., and Székely, G. J. (2016), "Fast Computing for Distance Covariance," *Technometrics*, 58, 435–446. [1638,1649]

Lancaster, H. O. (1969), *The Chi-Squared Distribution*, London: Wiley. [1639]

Lyons, R. (2013), "Distance Covariance in Metric Spaces," *Annals of Probability*, 41, 3284–3305. [1638,1639,1640]

Malkovich, J. F., and Afifi, A. A. (1973), "On Tests for Multivariate Normality," *Journal of the American Statistical Association*, 68, 176–179. [1641]

Mardia, K. V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, 57, 519–530. [1641]

Matteson, D. S., and Tsay, R. S. (2017), "Independent Component Analysis Via Distance Covariance," *Journal of the American Statistical Association*, 112, 623–637. [1639,1645]

Mcgill, W. J. (1954), "Multivariate Information Transmission," *Psychometrika*, 19, 97–116. [1638]

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014), "Causal Discovery with Continuous Additive Noise Models," *Journal of Machine Learning Research*, 15, 2009–2053. [1647]

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018), "Kernel-based Tests for Joint Independence," *Journal of the Royal Statistical Society*, Series B, 80, 5–31. [1638,1639,1645,1648,1649]

Read, T., and Cressie, N. (1988), *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*, New York: Springer-Verlag. [1638]

Sejdinovic, D., Gretton, A., and Bergsma, W. (2013), "A Kernel Test for Three-variable Interactions," in *Advances in Neural Information Processing Systems* (NIPS 26), pp. 1124–1132. [1638,1645]

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing," *Annals of Statistics*, 41, 2263–2291. [1638,1645]

Sen, A., and Sen, B. (2014), "Testing Independence and Goodness-of-Fit in Linear Models," *Biometrika*, 101, 927–942. [1647,1649]

Shannon, C. E., and Weaver, W. (1949), *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press. [1638]

Streitberg, B. (1990), "Lancaster Interactions Revisited," *Annals of Statistics*, 18, 1878–1885. [1639,1642]

Székely, G. J., and Rizzo, M. L. (2004), "Testing for Equal Distributions in High Dimension," *InterStat*, 5. [1641]

——— (2005), "Hierarchical Clustering Via Joint Between-Within Distances: Extending Ward's Minimum Variance Method," *Journal of Classification*, 22, 151–183. [1641]

——— (2009), "Brownian Distance Covariance," *Annals of Applied Statistics*, 3, 1236–1265. [1639,1640]

——— (2012), "On the Uniqueness of Distance Covariance," *Statistics and Probability Letters*, 82, 2278–2282. [1638]

——— (2013), "Energy Statistics: A Class of Statistics Based on Distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272. [1641]

——— (2014), "Partial Distance Correlation with Methods for Dissimilarities," *Annals of Statistics*, 42, 2382–2412. [1638,1644]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Independence by Correlation of Distances," *Annals of Statistics*, *35*, 2769–2794. [1638,1639,1640,1643]

Wang, X., Wenliang, P., Hu, W., Tian, Y., and Zhang, H. (2015), "Conditional Distance Correlation," *Journal of the American Statistical Association*, 110, 1726–1734. [1638,1649]

Wood, S. N., and Augustin, N. H. (2002), "GAMs with Integrated Model Selection Using Penalized Regression Splines and Applications to Environmental Modelling," *Ecological Modelling*, 157, 157–177. [1647]

Yao, S., Zhang, X., and Shao, X. (2018), "Testing Mutual Independence in High Dimension Via Distance Covariance," *Journal of the Royal Statistical Society*, Series B, 80, 455–480. [1641,1645]