Journal of Machine Learning Research 24 (2023) 1-16 Submitted 3/20; Revised 12/22; Published 1/23

On Distance and Kernel Measures of Conditional Dependence

Tianhong Sheng TXS514@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Bharath K. Sriperumbudur BKS18@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Editor: John Shawe-Taylor

Abstract

Measuring conditional dependence is one of the important tasks in statistical inference and is fundamental in causal discovery, feature selection, dimensionality reduction, Bayesian network learning, and others. In this work, we explore the connection between conditional dependence measures induced by distances on a metric space and reproducing kernels associated with a reproducing kernel Hilbert space (RKHS). For certain *distance and kernel pairs*, we show the distance-based conditional dependence measures to be equivalent to that of kernel-based measures. On the other hand, we also show that some popular kernel conditional dependence measures based on the Hilbert-Schmidt norm of a certain cross-conditional covariance operator, do not have a simple distance representation, except in some limiting cases.

Keywords: Conditional independence test, distance covariance, energy distance, Hilbert-Schmidt independence criterion, reproducing kernel Hilbert space

1. Introduction

Measuring conditional dependence between random variables plays a fundamental role in many statistical inference tasks such as causal discovery (Pearl, 2000; Spirtes et al., 2000), supervised dimensionality reduction (Cook and Li, 2002; Fukumizu et al., 2004), conditional independence testing (Su and White, 2007; Gretton et al., 2012), and others. Formally, for random variables (X, Y, Z), X is said to be *conditionally independent* of Y given Z, denoted as $X \perp\!\!\!\perp Y|Z$, if $P_{XY|Z} = P_{X|Z}P_{Y|Z}$ a.s.-P_Z, where the notation $P_{X|Z}$ denotes a *regular* conditional probability defined as $P_{X|Z}(\cdot) = \mathbb{E}[\mathbb{1}(X \in \cdot)|Z]$ a.s.-P_Z, with P_Z being the marginal distribution of Z. Given a distance measure D on the space of probability measures, $D(P_{XY|Z}, P_{X|Z}P_{Y|Z})$ measures the degree of conditional dependence between X and Y given Z, with $X \perp\!\!\!\perp Y|Z$ if and only if $D(P_{XY|Z}, P_{X|Z}P_{Y|Z}) = 0$ a.s.-P_Z. Some popular choices for D include the Kullback-Leibler divergence (more generally f-divergence), total variation distance, Hellinger distance, Wasserstein distance, among others.

Recently, a class of distances on probability measures induced by a Euclidean metric on $\mathbb{R}^d$—more generally by metrics of strongly negative type—, called the energy distance (Székely and Rizzo, 2004) and distance covariance (Székely et al., 2007; Székely and Rizzo, 2009; Lyons, 2013) has gained popularity in nonparametric hypothesis testing (e.g., two-sample and independence testing), because of their computational simplicity and elegant interpretation. Wang et al. (2015) extended distance covariance to conditional distributions on $\mathbb{R}^d$ to obtain a measure of conditional dependence, called *conditional distance covariance* (CdCov) and has been applied in conditional independence testing. We refer to these class of probability metrics as *distance-based measures* and point the reader to Section 3 for preliminaries on distance-based measures.

On the other hand, in the machine learning literature, measures of dependence have been formulated based on embedding of probability distributions into a reproducing kernel Hilbert space (RKHS; Aronszajn, 1950). This embedding into RKHS allows to capture the properties of distributions and has been used in many applications including homogeneity, independence, and conditional independence testing (for example, see Muandet et al., 2017 and references therein). Formally, given a probability measure ν defined on a measurable space $\mathscr{X}$, and a RKHS $\mathscr{H}_k$ with the reproducing kernel k, ν can be embedded into $\mathscr{H}_k$ as

$$\nu \mapsto \int_{\mathscr{X}} k(\cdot, x)\, d\nu(x) := \mu_k(\nu) \in \mathscr{H}_k,$$

where $\mu_k(\nu)$ is called the mean element or kernel mean embedding of ν. Using this notion, the *kernel distance*, also called as the maximum mean discrepancy (MMD) between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as the distance between their mean elements (Gretton et al., 2007), i.e., $D(\mathbb{P}, \mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathscr{H}_k}$. The kernel embedding and the kernel distance are well-studied in the literature and their mathematical theory is well-developed (Sriperumbudur et al., 2010, 2011; Sriperumbudur, 2016; Szabó and Sriperumbudur, 2018; Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2020). Generalizing this notion of kernel embedding to distributions defined on product spaces yields a kernel measure of dependence, called the Hilbert-Schmidt independence criterion (HSIC; Gretton et al., 2005, Gretton et al., 2008, Smola et al., 2007), which can then be used as a measure of conditional dependence by employing it to conditional probability distributions (Fukumizu et al., 2004, 2008). Fukumizu et al. (2004); Gretton et al. (2005) provided an alternate interpretation for HSIC in terms of the Hilbert-Schmidt norm of a certain cross-covariance operator, based on which the Hilbert-Schmidt norm of a conditional cross-covariance operator (we refer to it as HSC̈IC) is then proposed as a measure of conditional dependence. We point the reader to Sections 4 and 5 for details and refer to these class of probability metrics as *kernel-based measures*.

Sejdinovic et al. (2013) established an equivalence between distance-based and kernel-based dependence measures (i.e., distance covariance and HSIC) by showing that a reproducing kernel that defines HSIC induces a semi-metric of negative type which in turn defines the distance covariance (Székely et al., 2007, 2009), and vice-versa. However, despite the striking similarity, the relationship between conditional distance covariance and related kernel measures is not known. The goal of this work is to investigate the relationship between distance and kernel-based measures of conditional independence, and in particular,

understand whether these measures are equivalent (i.e., the distance measure can be obtained from the kernel measure and vice-versa).

As our contributions, first, in Theorem 1 (Section 4.2), we generalize the conditional distance covariance of Wang et al. (2015) to arbitrary metric spaces of negative type—we call this as generalized CdCov (gCdCov)—and develop a kernel measure of conditional dependence (we refer to it as HSCIC) that is *equivalent* to gCdCov. Therefore, it follows from Theorem 1 that CdCov introduced by Wang et al. (2015) is a special case of the HSCIC. In fact, the HSCIC we obtain is exactly the conditional dependence measure recently proposed by Park and Muandet (2020). Second, in Theorem 2 (Section 5), we consider the kernel measure of conditional dependence based on the Hilbert-Schmidt norm of the conditional cross-covariance operator (i.e., HS$\ddot{\text{C}}$IC) and obtain its distance-based interpretation. We show that this distance-based version of HS$\ddot{\text{C}}$IC does not have an elegant interpretation, except in limiting cases where it is related to CdCov and gCdCov (see Corollaries 3 and 4).

The paper is organized as follows. Definitions and notation that are widely used throughout the paper are collected in Section 2. The preliminaries on distance-based and kernel-based measures are presented in Sections 3 and 4.1, respectively, while main results are presented in Sections 4.2 and 5.

2. Definitions & Notation

For a non-empty set $\mathscr{X}$, a function $\rho : \mathscr{X} \times \mathscr{X} \to [0, \infty)$ is called a semi-metric on $\mathscr{X}$ if it satisfies (i) $\rho(x, x') = 0 \Leftrightarrow x = x'$ and (ii) $\rho(x, x') = \rho(x', x)$. Then $(\mathscr{X}, \rho)$ is said to be a semi-metric space. The semi-metric space, $(\mathscr{X}, \rho)$ is said to be of negative type if $\forall n \geq 2$, $\{x_i\}_{i=1}^n \in \mathscr{X}$, and $\{\alpha_i\}_{i=1}^n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$. $(\mathscr{X}, \rho)$ is said to be of strongly negative type if for all finite signed measures μ such that $\mu(\mathscr{X}) = 0$, $\int \int \rho(x, y)\, d\mu(x)\, d\mu(y) < 0$ for all $\mu \neq 0$. A real-valued symmetric function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called a positive definite (pd) kernel if, for all $n \in \mathbb{N}$, $\{\alpha_i\}_{i=1}^n \in \mathbb{R}$ and $\{x_i\}_{i=1}^n \in \mathscr{X}$, we have $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$. A function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, $(x, y) \mapsto k(x, y)$ is a *reproducing kernel* of the Hilbert space $(\mathscr{H}_k, \langle \cdot, \cdot \rangle_{\mathscr{H}_k})$ of functions if and only if (i) $\forall x \in \mathscr{X}$, $k(\cdot, x) \in \mathscr{H}$ and (ii) $\forall x \in \mathscr{X}$, $\forall f \in \mathscr{H}_k$, $\langle k(\cdot, x), f \rangle_{\mathscr{H}_k} = f(x)$ hold. If such a k exists, then $\mathscr{H}_k$ is called a *reproducing kernel Hilbert space*.

$\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ denote Polish spaces endowed with Borel σ-algebras. X, Y and Z denote random elements in $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$, respectively. $\ddot{X}$ is defined as (X, Z), which is a random element in $\mathscr{X} \times \mathscr{Z}$. The probability law of a random variable X is denoted by P_X, the joint probability law of random variables X and Z is denoted by P_{XZ} and the *regular* conditional probability of X given Z is defined as $P_{X|Z}(\cdot) = \mathbb{E}[\mathbb{1}(X \in \cdot)|Z]$ a.s.-P_Z such that $P_{X|Z=z}$ is a probability measure on $\mathscr{X}$ for all $z \in \mathscr{Z}$. The symbol $X \perp\!\!\!\perp Y|Z$ indicates the conditional independence of X and Y given Z. ϕ_X and ϕ_Y denote the characteristic functions of X and Y respectively and their joint characteristic function is denoted as ϕ_{XY}. The conditional characteristic functions of X, Y and (X, Y) given Z are denoted as $\phi_{X|Z}$, $\phi_{Y|Z}$ and $\phi_{XY|Z}$ respectively. A measurable, positive definite kernel on $\mathscr{X}$ is denoted as $k_{\mathscr{X}}$ and its corresponding RKHS as $\mathscr{H}_{\mathscr{X}}$. Similarly we define $k_{\mathscr{Y}}$, $\mathscr{H}_{\mathscr{Y}}$, $k_{\mathscr{Z}}$, $\mathscr{H}_{\mathscr{Z}}$, $k_{\ddot{X}}$ and $\mathscr{H}_{\ddot{X}}$. In this paper we assume that all involved RKHS's are separable.

The space of r-integrable functions w.r.t. a σ-finite measure, μ on $\mathbb{R}^d$ is denoted as $L^r(\mathbb{R}^d, \mu)$ and if μ is a Lebesgue measure on $\mathbb{R}^d$, we denote it as $L^r(\mathbb{R}^d)$.

3. Conditional Distance Covariance

Distance covariance was proposed by Székely et al. (2007) as a new measure of dependence between Euclidean random vectors in arbitrary dimension. An interesting feature of distance covariance is that unlike the classical covariance, it is zero only if the random vectors are independent. Formally, the distance covariance (dCov) between two random vectors is defined as the weighted L^2 norm between the joint characteristic function and the product of marginal characteristic functions, i.e.,

$$\mathcal{V}^2(X,Y) = \|\phi_{XY} - \phi_X\phi_Y\|^2_{L^2(w)} = \frac{1}{c_p c_q}\int\int \frac{|\phi_{XY}(t,s) - \phi_X(t)\phi_Y(s)|^2}{\|t\|^{p+1}\|s\|^{q+1}}\,dt\,ds,$$

where ϕ_{XY} denotes the joint characteristic function of random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with ϕ_X and ϕ_Y denoting their respective marginal characteristic functions. Here $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$, $c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}$ and $w(t,s) = \|t\|^{-p-1}\|s\|^{-q-1}$ with $\|t\|^2 = \sum_{i=1}^p t_i^2$ for $t = (t_1,\ldots,t_p)$. A particular advantage of distance covariance is its compact representation in terms of certain expectation of pairwise Euclidean distances (Székely et al., 2007):

$$\begin{aligned}\mathcal{V}^2(X,Y) = \mathbb{E}[\mathbb{E}[\|X-X'\|\|Y-Y'\||X,Y]] + \mathbb{E}\|X-X'\|\mathbb{E}\|Y-Y'\| \\ -2\mathbb{E}[\mathbb{E}[\|X-X'\||X]\mathbb{E}[\|Y-Y'\||Y]],\end{aligned} \tag{1}$$

where $X \overset{\text{i.i.d.}}{\sim} X'$, $Y \overset{\text{i.i.d.}}{\sim} Y'$, which leads to straightforward empirical estimates by replacing the expectations with empirical estimators. Such an estimator has been used as a test statistic in independence testing and the resulting test is shown to be consistent if the marginal distributions have finite first moment (Székely et al., 2007). As a natural generalization, Lyons (2013) extended (1) to metric spaces of negative type and showed that the corresponding distance covariance—obtained by replacing the Euclidean metric by a metric of strongly negative type—is zero if and only if X and Y are independent.

Extending the idea of distance covariance, recently, Wang et al. (2015) proposed a conditional version to measure conditional independence between random vectors of arbitrary dimension. To elaborate, let $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ and $Z \in \mathbb{R}^r$ be random vectors. The conditional distance covariance (CdCov) $\mathcal{V}(X,Y|Z)$ between random vectors X and Y with finite moments given Z is defined as

$$\mathcal{V}^2(X,Y|Z) = \|\phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}\|^2_{L^2(w)} = \int\int \frac{|\phi_{XY|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)|^2}{c_p c_q\|t\|^{p+1}\|s\|^{q+1}}\,dt\,ds,$$

where

$$\phi_{XY|Z}(t,s) = \mathbb{E}\left[e^{\sqrt{-1}\langle t,X\rangle + \sqrt{-1}\langle s,Y\rangle}|Z\right],\ \phi_{X|Z}(t) = \phi_{XY|Z}(t,0) \text{ and } \phi_{Y|Z}(s) = \phi_{XY|Z}(0,s).$$

As a crucial property, CdCov is zero P_Z-almost surely if and only if $X \perp\!\!\!\perp Y|Z$. Similar to distance covariance, one advantage of this measure is that its sample version can be expressed elegantly as a V- or U-statistic, based on which Wang et al. (2015) proposed a statistically consistent conditional independence test.

The conditional distance covariance defined above can also be computed in terms of the conditional expectations of pairwise Euclidean distances:

$$\begin{aligned}\mathcal{V}^2(X,Y|Z) = \mathbb{E}[\mathbb{E}[\|X-X'\|\|Y-Y'\||X,Y,Z]|Z] + \mathbb{E}[\|X-X'\||Z]\mathbb{E}[\|Y-Y'\||Z] \\ -2\mathbb{E}[[\mathbb{E}[\|X-X'\||X,Z]\mathbb{E}[\|Y-Y'\||Y,Z]|Z],\end{aligned} \tag{2}$$

where (X, Y) and (X', Y') are independent copies given Z. In the similar spirit of Lyons (2013), CdCov can be extended to metric spaces of negative type through conditional expectations so that (2) can be written as

$$\begin{aligned}\mathcal{V}^2_{\rho_{\mathscr{X}},\rho_{\mathscr{Y}}}(X,Y|Z) &= \mathbb{E}[\mathbb{E}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')|X,Y,Z]|Z] \\ &\quad +\mathbb{E}[\rho_{\mathscr{X}}(X,X')|Z]\mathbb{E}[\rho_{\mathscr{Y}}(Y,Y')|Z] \\ &\quad -2\mathbb{E}[\mathbb{E}[\rho_{\mathscr{X}}(X,X')|X,Z]\mathbb{E}[\rho_{\mathscr{Y}}(Y,Y')|Y,Z]|Z], \qquad (3) \\ &=: \mathbb{G}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')] =: \mathbb{G}\circ[\rho_{\mathscr{X}}\rho_{\mathscr{Y}}], \qquad (4)\end{aligned}$$

where $\rho_{\mathscr{X}}$ and $\rho_{\mathscr{Y}}$ are metrics of strongly negative type defined on spaces $\mathscr{X}$ and $\mathscr{Y}$ respectively with $\mathbb{E}[\rho^2_{\mathscr{X}}(X, x_0)|Z] < \infty$ a.s.-P_Z and $\mathbb{E}[\rho^2_{\mathscr{Y}}(Y, y_0)|Z] < \infty$ a.s.-P_Z for some $x_0 \in \mathscr{X}$ and $y_0 \in \mathscr{Y}$. The moment conditions ensure that the expectations are finite. When $\rho_{\mathscr{X}}$ and $\rho_{\mathscr{Y}}$ are strongly negative, then clearly (3) is zero if and only if $X \perp\!\!\!\perp Y|Z$.

4. Kernel Measures of Conditional Dependence

First, in Section 4.1, we present preliminaries on RKHS embedding of probability measures and introduce kernel measures of dependence. Based on this discussion, in Section 4.2, we develop a kernel measure of conditional dependence (we call it as Hilbert-Schmidt conditional independence criterion—HSCIC) that is related to gCdCov (and therefore CdCov) discussed in Section 3. We also present an interpretation for gCdCov through conditional cross-covariance operator formulation for HSCIC.

4.1 RKHS embedding of probabilities

In the machine learning literature, the notion of embedding probability measures in an RKHS has gained lot of attention and has been applied in goodness-of-fit (Balasubramanian et al., 2021), two-sample (Gretton et al., 2007, 2012), independence (Gretton et al., 2008) and conditional independence (Fukumizu et al., 2008; Zhang et al., 2011) testing. To elaborate, given a probability measure P such that $\int_{\mathscr{X}} \sqrt{k(x,x)}\, dP(x) < \infty$, its RKHS embedding (Smola et al., 2007) is defined as

$$P \mapsto \mu_P := \int_{\mathscr{X}} k(\cdot, x)\, dP(x) \in \mathscr{H}_k,$$

where $\mathscr{H}_k$ is an RKHS with k as the reproducing kernel. Based on this embedding, a distance on the space of probabilities can be defined through the distance between the embeddings, i.e., $\mathcal{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathscr{H}_k}$, called the *kernel distance* or *maximum mean discrepancy* (Gretton et al., 2007). If the map $P \mapsto \mu_P$ is injective, then the kernel k that induces μ_P is said to be *characteristic* (Fukumizu et al., 2009; Sriperumbudur et al., 2010) and therefore $\mathcal{D}_k(P, Q)$ induces a metric on $\mathcal{M}^{1/2}_k(\mathscr{X}) := \{P \in \mathcal{M}^1_+(\mathscr{X}) : \int_{\mathscr{X}} \sqrt{k(x,x)}\, dP(x) < \infty\}$, where $\mathcal{M}^1_+(\mathscr{X})$ denotes the set of all probability measures on $\mathscr{X}$. Using the reproducing property of the kernel, it can be shown that

$$\mathcal{D}^2_k(P, Q) = \mathbb{E}_{XX'}k(X, X') + \mathbb{E}_{YY'}k(Y, Y') - 2\mathbb{E}_{XY}k(X, Y),$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$. Extending this distance to probability measures on product spaces, particularly the joint measure P_{XY} and product of marginals $P_X P_Y$, yields

a measure of dependence between two random variables X and Y defined on measurable spaces $\mathscr{X}$ and $\mathscr{Y}$, called the Hilbert-Schmidt Independence Criterion (HSIC), which is defined (Gretton et al., 2005) as

$$\begin{aligned}\mathcal{D}^2_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY}, P_X P_Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y') \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}k_{\mathscr{X}}(X,X')\mathbb{E}_Y\mathbb{E}_{Y'}k_{\mathscr{Y}}(Y,Y') \\ &\quad - 2\mathbb{E}_{XY}[\mathbb{E}_{X'}k_{\mathscr{X}}(X,X')\mathbb{E}_{Y'}k_{\mathscr{Y}}(Y,Y')] \\ &= \int (k_{\mathscr{X}}k_{\mathscr{Y}})(x,y,x',y')\, d[P_{XY} - P_X P_Y]^2(x,y,x',y'). \end{aligned} \tag{5}$$

If the kernels $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are characteristic, then HSIC characterizes independence (Szabó and Sriperumbudur, 2018), i.e., $\mathcal{D}_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY}, P_X P_Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. An empirical version of (5) has been used as a test statistic in independence testing and the resultant test is shown to be consistent against all alternatives as long as $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are characteristic (Gretton et al., 2008). An interesting connection between kernel-based HSIC and distance-based dCov is shown by Sejdinovic et al. (2013) that dCov in (1) is in fact a special case of HSIC and HSIC is equivalent to the generalized dCov introduced by Lyons (2013). This result provides a unifying framework for the distance and kernel-based dependence measures. With this background, in the rest of the paper, we explore the relation between distance and kernel-based measures of conditional dependence.

4.2 Hilbert-Schmidt conditional independence criterion

For appropriate choice of kernels and distances, the following result provides a kernel-equivalent of gCdCov, which we refer to as the Hilbert-Schmidt conditional independence criterion (HSCIC).

Theorem 1 *Let $(\mathscr{X}, \rho_{\mathscr{X}})$ and $(\mathscr{Y}, \rho_{\mathscr{Y}})$ be semi-metric spaces of negative type. Suppose $\mathbb{E}[\rho^2_{\mathscr{X}}(X, x_0)|Z] < \infty$ and $\mathbb{E}[\rho^2_{\mathscr{Y}}(Y, y_0)|Z] < \infty$ a.s.-P_Z for some $x_0 \in \mathscr{X}$, $y_0 \in \mathscr{Y}$. If $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are pd kernels on $\mathscr{X}$ and $\mathscr{Y}$ that are distance-induced, i.e.,*

$$k_{\mathscr{X}}(x,x') = \rho_{\mathscr{X}}(x,\theta) + \rho_{\mathscr{X}}(x',\theta) - \rho_{\mathscr{X}}(x,x')$$

and

$$k_{\mathscr{Y}}(y,y') = \rho_{\mathscr{Y}}(y,\theta') + \rho_{\mathscr{Y}}(y',\theta') - \rho_{\mathscr{Y}}(y,y')$$

for some $\theta \in \mathscr{X}$ and $\theta' \in \mathscr{Y}$. Then

$$\mathcal{V}^2_{\rho_{\mathscr{X}},\rho_{\mathscr{Y}}}(X,Y|Z) = \mathbb{G} \circ [\rho_{\mathscr{X}}\rho_{\mathscr{Y}}] = \mathcal{D}^2_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY|Z}, P_{X|Z}P_{Y|Z}), \quad \text{a.s.-}P_Z \tag{6}$$

with

$$\mathcal{D}^2_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY|Z}, P_{X|Z}P_{Y|Z}) = \mathbb{G} \circ [k_{\mathscr{X}}k_{\mathscr{Y}}],$$

where $\mathbb{G}$ is defined in (4).

On the other hand, let $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ be pd kernels on $\mathscr{X}$ and $\mathscr{Y}$ respectively. Suppose $\mathbb{E}[k^2_{\mathscr{X}}(X,X)|Z] < \infty$ and $\mathbb{E}[k^2_{\mathscr{Y}}(Y,Y)|Z] < \infty$ a.s.-P_Z. If $\rho_{\mathscr{X}}$ and $\rho_{\mathscr{Y}}$ are semi-metrics on $\mathscr{X}$ and $\mathscr{Y}$ that are kernel-induced, i.e.,

$$\rho_{\mathscr{X}}(x,x') = \frac{k_{\mathscr{X}}(x,x) + k_{\mathscr{X}}(x',x')}{2} - k_{\mathscr{X}}(x,x')$$

and

$$\rho_{\mathscr{Y}}(y,y') = \frac{k_{\mathscr{Y}}(y,y)+k_{\mathscr{Y}}(y',y')}{2} - k_{\mathscr{Y}}(y,y'),$$

then (6) *holds.*

Proof Suppose $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are distance-induced. Then

$$\begin{aligned}
&\mathcal{D}^2_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY|Z}, P_{X|Z}P_{Y|Z}) \\
&= \mathbb{G} \circ [k_{\mathscr{X}}k_{\mathscr{Y}}] = \mathbb{G}[k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y')] \\
&= \mathbb{G}\left[\left(\rho_{\mathscr{X}}(X,\theta)+\rho_{\mathscr{X}}(X',\theta)-\rho_{\mathscr{X}}(X,X')\right)\left(\rho_{\mathscr{Y}}(Y,\theta')+\rho_{\mathscr{Y}}(Y',\theta')-\rho_{\mathscr{Y}}(Y,Y')\right)\right] \\
&= \mathbb{G}\left[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')\right] = \mathcal{V}^2_{\rho_{\mathscr{X}},\rho_{\mathscr{Y}}}(X,Y|Z)
\end{aligned}$$

a.s.-P_Z, where we used the fact that $\mathbb{G}[g(X,Y,X',Y')]=0$ a.s.-P_Z when g does not depend on one or more of its arguments (for example, a constant function). On the other hand, suppose $\rho_{\mathscr{X}}$ and $\rho_{\mathscr{Y}}$ are kernel-induced. Clearly they are of negative type. Then

$$\begin{aligned}
&\mathcal{V}^2_{\rho_{\mathscr{X}},\rho_{\mathscr{Y}}}(X,Y|Z) \\
&= \mathbb{G}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')] \\
&= \mathbb{G}\left[\left(\frac{k_{\mathscr{X}}(X,X)+k_{\mathscr{X}}(X',X')}{2} - k_{\mathscr{X}}(X,X')\right)\left(\frac{k_{\mathscr{Y}}(Y,Y)+k_{\mathscr{Y}}(Y',Y')}{2} - k_{\mathscr{Y}}(Y,Y')\right)\right] \\
&= \mathbb{G}[k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y')] = \mathcal{D}^2_{k_{\mathscr{X}}k_{\mathscr{Y}}}(P_{XY|Z}, P_{X|Z}P_{Y|Z}),
\end{aligned}$$

a.s.-P_Z, where we again used the above mentioned facts about $\mathbb{G}$. ■

Note that, while the quantities θ and θ' induce a family of kernels as θ and θ' range over $\mathscr{X}$ and $\mathscr{Y}$ respectively, all these kernels are equivalent in the sense that they induce the same HSCIC as shown by the equivalence in (6). This means, CdCov is induced by kernels of the form $k_{\mathscr{X}}(x,x') = \|x-\theta\| + \|x'-\theta\| - \|x-x'\|$, $x,x' \in \mathbb{R}^p$ and $k_{\mathscr{Y}}(y,y') = \|y-\theta'\| + \|y'-\theta'\| - \|y-y'\|$, $y,y' \in \mathbb{R}^q$ with $\theta=\theta'=0$ being a popular choice—this choice leads to covariance function of a fractional Brownian motion.

We would like to mention that a concurrent and independent work by Park and Muandet (2020) proposed a criterion with the same name HSCIC, which is defined as the distance between the conditional mean embedding of $P_{XY|Z}$ and the product of marginal conditional mean embeddings of $P_{X|Z}$ and $P_{Y|Z}$, where the conditional mean embedding of $P_{X|Z}$ is denoted by $\mu_{P_{X|Z}}$ and $\mu_{P_{X|Z}} = \mathbb{E}[k_{\mathscr{X}}(X,\cdot)|Z]$ (the conditional mean embedding of $P_{Y|Z}$ and $P_{XY|Z}$ can be similarly defined). It is easy to verify that

$$\begin{aligned}
&\mathcal{V}^2_{\rho_{\mathscr{X}},\rho_{\mathscr{Y}}}(X,Y|Z) = \mathbb{G}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')] = \mathbb{G}[k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y')] \\
&= \|\mathbb{E}[k_{\mathscr{X}}(X,\cdot)\otimes k_{\mathscr{Y}}(Y,\cdot)|Z] - \mathbb{E}[k_{\mathscr{X}}(X,\cdot)|Z]\otimes\mathbb{E}[k_{\mathscr{Y}}(Y,\cdot)|Z]\|^2_{\mathscr{H}_{\mathscr{X}}\otimes\mathscr{H}_{\mathscr{Y}}} \\
&= \left\|\mu_{P_{XY|Z}} - \mu_{P_{X|Z}}\otimes\mu_{P_{Y|Z}}\right\|^2_{\mathscr{H}_{\mathscr{X}}\otimes\mathscr{H}_{\mathscr{Y}}}. \qquad (7)
\end{aligned}$$

While HSCIC is a natural measure of conditional dependence, in the kernel literature, however, a different measure has been widely used (Fukumizu et al., 2004, 2008; Zhang et al., 2011), which is based on the Hilbert-Schmidt norm of a certain *conditional cross-covariance operator*. Before we introduce the conditional cross-covariance operator and these other

measures of conditional dependence (which we do in Section 5), first we will briefly discuss how HSIC is related to the Hilbert-Schmidt norm of a cross-covariance operator so that its extension to the conditional version is natural.

For random variables $X \sim P_X$ and $Y \sim P_Y$ with joint distribution P_{XY} such that $\mathbb{E}[k_{\mathscr{X}}(X,X)] < \infty$ and $\mathbb{E}[k_{\mathscr{Y}}(Y,Y)] < \infty$, there exists a unique bounded linear operator, called the cross-covariance operator (Baker, 1973; Fukumizu et al., 2004), $\Sigma_{YX} : \mathscr{H}_{k_{\mathscr{X}}} \to \mathscr{H}_{k_{\mathscr{Y}}}$ such that $\forall\, f \in \mathscr{H}_{k_{\mathscr{X}}}$, $g \in \mathscr{H}_{k_{\mathscr{Y}}}$,

$$\langle g, \Sigma_{YX} f\rangle_{\mathscr{H}_{k_{\mathscr{Y}}}} = \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

In fact, using the reproducing property that $f(x) = \langle f, k_{\mathscr{X}}(\cdot, x)\rangle_{\mathscr{H}_{k_{\mathscr{X}}}}$, $\forall\, x \in \mathscr{X}$ and $g(y) = \langle g, k_{\mathscr{Y}}(\cdot, y)\rangle_{\mathscr{H}_{k_{\mathscr{Y}}}}$, $\forall\, y \in \mathscr{Y}$, it follows that

$$\Sigma_{YX} = \int\int k_{\mathscr{Y}}(\cdot, y) \otimes k_{\mathscr{X}}(\cdot, x)\, dP_{XY}(x,y) - \int k_{\mathscr{Y}}(\cdot, y)\, dP_Y(y) \otimes \int k_{\mathscr{X}}(\cdot, x)\, dP_X(x), \quad (8)$$

where $\otimes$ denotes the tensor product. Clearly, Σ_{YX} is a natural generalization of the finite-dimensional covariance matrix between two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. Based on (8) and the reproducing property, it can be verified that

$$\begin{aligned}
\|\Sigma_{YX}\|_{HS}^2 &= \left\| \int\int k_{\mathscr{X}}(\cdot, x) \otimes k_{\mathscr{Y}}(\cdot, y)\, d(P_{XY} - P_X P_Y)(x,y) \right\|_{HS}^2 \\
&= \int\int\int\int \langle k_{\mathscr{X}}(\cdot, x) \otimes k_{\mathscr{Y}}(\cdot, y), k_{\mathscr{X}}(\cdot, x') \otimes k_{\mathscr{Y}}(\cdot, y')\rangle_{HS}\, d(P_{XY} - P_X P_Y)(x,y) \\
&\qquad\qquad \times d(P_{XY} - P_X P_Y)(x', y') \\
&= \mathcal{D}^2_{k_{\mathscr{X}} k_{\mathscr{Y}}}(P_{XY}, P_X P_Y), \qquad (9)
\end{aligned}$$

where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm. Since HSCIC is a conditional version of HSIC and since the latter is the Hilbert-Schmidt norm of the cross-variance operator, it is natural to extend Σ_{YX} to its conditional version as a P_Z-measurable bounded linear operator $\dot{\Sigma}_{YX|Z} : \mathscr{H}_{k_{\mathscr{X}}} \to \mathscr{H}_{k_{\mathscr{Y}}}$ such that $\forall\, f \in \mathscr{H}_{k_{\mathscr{X}}}$, $g \in \mathscr{H}_{k_{\mathscr{Y}}}$,

$$\langle g, \dot{\Sigma}_{YX|Z} f\rangle_{\mathscr{H}_{k_{\mathscr{Y}}}} = \mathbb{E}[f(X)g(Y)|Z] - \mathbb{E}[f(X)|Z]\mathbb{E}[g(Y)|Z], \ \text{ a.s.-}P_Z,$$

thereby yielding

$$\dot{\Sigma}_{YX|Z} = \mathbb{E}[k_{\mathscr{X}}(\cdot, X) \otimes k_{\mathscr{Y}}(\cdot, Y)|Z] - \mathbb{E}[k_{\mathscr{X}}(\cdot, X)|Z] \otimes \mathbb{E}[k_{\mathscr{Y}}(\cdot, Y)|Z].$$

Similar to (9), it is easy to verify that

$$\|\dot{\Sigma}_{YX|Z}\|_{HS}^2 = \mathcal{D}^2_{k_{\mathscr{X}} k_{\mathscr{Y}}}(P_{XY|Z}, P_{X|Z} P_{Y|Z})$$

a.s.-P_Z. Therefore if $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are characteristic, then $X \perp\!\!\!\perp Y|Z \iff \dot{\Sigma}_{YX|Z} = 0$, P_Z-a.s.

However, in the kernel literature, to the best of our knowledge, besides the concurrent and independent work by Park and Muandet (2020) in which a quantity similar to HSCIC is proposed, HSCIC has not been used as a measure of conditional independence probably because it is a random operator. We can obtain a single measure of conditional dependence by considering the expectation of HSCIC over $Z \sim P_Z$, i.e.,

$$\mathcal{D}_{P_Z}(P_{XY|Z}, P_{X|Z} P_{Y|Z}) := \mathbb{E}_Z[\|\dot{\Sigma}_{YX|Z}\|_{HS}^2]. \quad (10)$$

This single measure of conditional dependence by taking HSCIC over Z is not discussed in Park and Muandet (2020).

5. Relation between RKHS and Distance-based Conditional Dependence Measures

Instead of $\dot{\Sigma}_{YX|Z}$, Fukumizu et al. (2004) considered an alternate operator, called the *conditional cross-covariance operator*, which is defined as follows. Suppose $\mathbb{E}_X[k_{\mathscr{X}}(X,X)] < \infty$, $\mathbb{E}_Y[k_{\mathscr{Y}}(Y,Y)] < \infty$ and $\mathbb{E}_Z[k_{\mathscr{Z}}(Z,Z)] < \infty$. Then there exists a unique bounded linear operator $\Sigma_{YX|Z}$ such that

$$\begin{aligned}\langle g, \Sigma_{YX|Z} f\rangle_{\mathscr{H}_{k_{\mathscr{Y}}}} &= \mathbb{E}[f(X)g(Y)] - \mathbb{E}[\mathbb{E}[f(X)|Z]\mathbb{E}[g(Y)|Z]] \\ &= \mathbb{E}[\mathrm{Cov}(f(X), g(Y)|Z)]\end{aligned}$$

for all $f \in \mathscr{H}_{k_{\mathscr{X}}}$ and $g \in \mathscr{H}_{k_{\mathscr{Y}}}$. As above, using the reproducing property, it can be shown that

$$\Sigma_{YX|Z} = \mathbb{E}[\mathbb{E}[k_{\mathscr{Y}}(\cdot,Y) \otimes k_{\mathscr{X}}(\cdot,X)|Z] - \mathbb{E}[k_{\mathscr{X}}(\cdot,X)|Z] \otimes \mathbb{E}[k_{\mathscr{Y}}(\cdot,Y)|Z]] = \mathbb{E}_Z[\dot{\Sigma}_{YX|Z}].$$

However, unlike $\dot{\Sigma}_{YX|Z}$, the conditional cross-covariance operator $\Sigma_{YX|Z}$ does not characterize conditional independence since $\Sigma_{YX|Z} = 0$—assuming $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ to be characteristic—only implies $P_{XY} = \mathbb{E}_Z[P_{X|Z}P_{Y|Z}]$ and not $\dot{\Sigma}_{YX|Z} = 0$, a.s.-P_Z (Fukumizu et al., 2004, Theorem 8). Therefore, Fukumizu et al. (2004, Corollary 9) considered Z as a part of X by defining $\ddot{X} := (X,Z)$ and showed that $\Sigma_{Y\ddot{X}|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$, assuming $k_{\mathscr{X}}$, $k_{\mathscr{Y}}$ and $k_{\mathscr{Z}}$ to be characteristic. This is indeed the case since if $k_{\mathscr{X}}$, $k_{\mathscr{Y}}$ and $k_{\mathscr{Z}}$ are characteristic, then $\Sigma_{Y\ddot{X}|Z} = 0$ implies $\mathbb{E}_Z[\dot{\Sigma}_{Y\ddot{X}|Z}] = 0$ and therefore

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\mathbb{1}\{X \in A, Y \in B, Z \in C\}|Z]] \\ &\qquad -\mathbb{E}[\mathbb{E}[\mathbb{1}\{X \in A, Z \in C\}|Z]\mathbb{E}_{Y|Z}[\mathbb{1}\{Y \in B\}|Z]] \\ &= \mathbb{E}[\mathbb{1}\{X \in A, Y \in B, Z \in C\}] - \mathbb{E}[\mathbb{E}[\mathbb{1}\{X \in A, Z \in C\}|Z]\mathbb{E}[\mathbb{1}\{Y \in B\}|Z]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}\{X \in A, Y \in B\}|Z]\mathbb{1}\{Z \in C\}] \\ &\qquad -\mathbb{E}[\mathbb{E}[\mathbb{1}\{X \in A\}|Z]\mathbb{1}\{Z \in C\}]\mathbb{E}[\mathbb{1}\{Y \in B\}|Z]] \\ &= \mathbb{E}[[P_{XY|Z}(A \times B|Z) - P_{X|Z}(A|Z)P_{Y|Z}(B|Z)]\mathbb{1}\{Z \in C\}] = 0,\end{aligned}$$

for all $A \in \mathcal{B}_{\mathscr{X}}$, $B \in \mathcal{B}_{\mathscr{Y}}$ and $C \in \mathcal{B}_{\mathscr{Z}}$, where $\mathcal{B}_{\mathscr{X}}$, $\mathcal{B}_{\mathscr{Y}}$ and $\mathcal{B}_{\mathscr{Z}}$ are the Borel σ-algebras associated with $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ respectively. This implies,

$$P_{XY|Z}(A \times B|Z) - P_{X|Z}(A|Z)P_{Y|Z}(B|Z) = 0, \quad \text{a.s.-}P_Z,$$

implying $X \perp\!\!\!\perp Y|Z$, a.s.-P_Z. Hence $\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS}$ can be used as a measure of conditional independence, which we refer to it as HS$\ddot{\text{C}}$IC.

The goal of this section is to explore the distance counterpart of HS$\ddot{\text{C}}$IC and understand how it is related to CdCov, gCdCov, and $\mathcal{D}_{P_Z}$ defined in (10). To this end, we first provide an expression for $\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS}$ in terms of kernels, using which we obtain an expression in terms of distances.

Theorem 2 *Suppose* $\mathbb{E}_X[k^2_{\mathscr{X}}(X,X)] < \infty$, $\mathbb{E}_Y[k^2_{\mathscr{Y}}(Y,Y)] < \infty$ *and* $\mathbb{E}_Z[k^2_{\mathscr{Z}}(Z,Z)] < \infty$. *Denote* $\ddot{X} = (X,Z)$ *Then*

$$\begin{aligned}\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS} &= \mathbb{E}_Z\mathbb{E}_{Z'}\left[k_{\mathscr{Z}}(Z,Z')\left\langle \dot{\Sigma}_{YX|Z}, \dot{\Sigma}_{YX|Z'}\right\rangle_{HS}\right] \\ &= \mathbb{E}_Z\mathbb{E}_{Z'}[k_{\mathscr{Z}}(Z,Z')h(Z,Z')], \qquad (11)\end{aligned}$$

where $h(Z,Z') := \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y')]$, $\mathbb{F}_{YX|Z} := \mathbb{E}_{XY|Z} - \mathbb{E}_{Y|Z}\mathbb{E}_{X|Z}$ *and* $\mathbb{E}_{XY|Z} := \mathbb{E}[\cdot|Z]$ *(*$\mathbb{E}_{Y|Z}$ *and* $\mathbb{E}_{X|Z}$ *are defined similarly).*

Suppose $k_{\mathscr{X}}$ *and* $k_{\mathscr{Y}}$ *are distance-induced, i.e.,*

$$k_{\mathscr{X}}(x,x') = \rho_{\mathscr{X}}(x,\theta)+\rho_{\mathscr{X}}(x',\theta)-\rho_{\mathscr{X}}(x,x') \text{ and } k_{\mathscr{Y}}(y,y') = \rho_{\mathscr{Y}}(y,\theta')+\rho_{\mathscr{Y}}(y',\theta')-\rho_{\mathscr{Y}}(y,y')$$

for some $\theta \in \mathscr{X}$ *and* $\theta' \in \mathscr{Y}$*. Then* $h(Z,Z') = \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')]$.

Proof Note that

$$\begin{aligned}
&\Sigma_{Y\ddot{X}|Z} = \mathbb{E}[\dot{\Sigma}_{Y\ddot{X}|Z}]\\
&= \mathbb{E}[\mathbb{E}[k_{\mathscr{Y}}(\cdot,Y)\otimes(k_{\mathscr{X}}k_{\mathscr{Z}})(\cdot,\ddot{X})|Z]] - \mathbb{E}[\mathbb{E}[k_{\mathscr{Y}}(\cdot,Y)|Z]\otimes\mathbb{E}[(k_{\mathscr{X}}k_{\mathscr{Z}})(\cdot,\ddot{X})|Z]]\\
&= \mathbb{E}[\mathbb{E}[k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X)|Z]\otimes k_{\mathscr{Z}}(\cdot,Z)]\\
&\qquad -\mathbb{E}[\mathbb{E}[k_{\mathscr{Y}}(\cdot,Y)|Z]\otimes\mathbb{E}[k_{\mathscr{X}}(\cdot,X)|Z]\otimes k_{\mathscr{Z}}(\cdot,Z)]\\
&= \mathbb{E}_Z[\dot{\Sigma}_{YX|Z}\otimes k_{\mathscr{Z}}(\cdot,Z)].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\Sigma_{Y\ddot{X}|Z}\|_{HS}^2 &= \left\|\mathbb{E}[\dot{\Sigma}_{YX|Z}\otimes k_{\mathscr{Z}}(\cdot,Z)]\right\|_{HS}^2\\
&= \left\langle \mathbb{E}_Z[\dot{\Sigma}_{YX|Z}\otimes k_{\mathscr{Z}}(\cdot,Z)], \mathbb{E}_Z[\dot{\Sigma}_{YX|Z}\otimes k_{\mathscr{Z}}(\cdot,Z)]\right\rangle_{HS}\\
&= \mathbb{E}_Z\mathbb{E}_{Z'}\left\langle \dot{\Sigma}_{YX|Z}\otimes k_{\mathscr{Z}}(\cdot,Z), \dot{\Sigma}_{YX|Z'}\otimes k_{\mathscr{Z}}(\cdot,Z')\right\rangle_{HS}\\
&= \mathbb{E}_Z\mathbb{E}_{Z'}\left\langle \dot{\Sigma}_{YX|Z}, \dot{\Sigma}_{YX|Z'}\right\rangle_{HS}\langle k_{\mathscr{Z}}(\cdot,Z), k_{\mathscr{Z}}(\cdot,Z')\rangle_{\mathscr{H}_{k_{\mathscr{Z}}}}\\
&= \mathbb{E}_Z\mathbb{E}_{Z'}\left\langle \dot{\Sigma}_{YX|Z}, \dot{\Sigma}_{YX|Z'}\right\rangle_{HS} k_{\mathscr{Z}}(Z,Z'). \qquad (12)
\end{aligned}$$

Note that $\dot{\Sigma}_{YX|Z} = \mathbb{F}_{YX|Z}[k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X)]$. Therefore,

$$\begin{aligned}
\left\langle \dot{\Sigma}_{YX|Z}, \dot{\Sigma}_{YX|Z'}\right\rangle_{HS} &= \left\langle \mathbb{F}_{YX|Z}\left[k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X)\right], \mathbb{F}_{YX|Z'}\left[k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X)\right]\right\rangle_{HS}\\
&= \left\langle \mathbb{F}_{YX|Z}\left[k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X)\right], \mathbb{F}_{Y'X'|Z'}\left[k_{\mathscr{Y}}(\cdot,Y')\otimes k_{\mathscr{X}}(\cdot,X')\right]\right\rangle_{HS}\\
&= \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}\left[\left\langle k_{\mathscr{Y}}(\cdot,Y)\otimes k_{\mathscr{X}}(\cdot,X), k_{\mathscr{Y}}(\cdot,Y')\otimes k_{\mathscr{X}}(\cdot,X')\right\rangle_{HS}\right]\\
&= \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}\left[\left\langle k_{\mathscr{Y}}(\cdot,Y), k_{\mathscr{Y}}(\cdot,Y')\right\rangle_{\mathscr{H}_{k_{\mathscr{Y}}}}\left\langle k_{\mathscr{X}}(\cdot,X), k_{\mathscr{X}}(\cdot,X')\right\rangle_{\mathscr{H}_{k_{\mathscr{X}}}}\right]\\
&= \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[k_{\mathscr{X}}(X,X')k_{\mathscr{Y}}(Y,Y')] = h(Z,Z'),
\end{aligned}$$

using which in (12) yields the result. If $k_{\mathscr{X}}$ and $k_{\mathscr{Y}}$ are distance-induced, then using the fact that $\mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[g(X,X',Y,Y')] = 0$ when g does not depend on one or more of its arguments—basically, the same argument that we carried out in the proof of Theorem 1—we have

$$h(Z,Z') = \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[\rho_{\mathscr{X}}(X,X')\rho_{\mathscr{Y}}(Y,Y')],$$

and the result follows. ■

While $h(Z,Z')$ has a distance interpretation as shown in Theorem 2, $\|\Sigma_{Y\ddot{X}|Z}\|_{HS}^2$ does not

have an elegant representation in terms of distances. Suppose $k_{\mathscr{Z}}$ is also distance-induced, i.e., $k_{\mathscr{Z}}(z,z') = \rho_{\mathscr{Z}}(z,\theta'') + \rho_{\mathscr{Z}}(x',\theta'') - \rho_{\mathscr{Z}}(z,z')$ for some $\theta'' \in \mathscr{Z}$. Then

$$\begin{aligned}\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS} &= \int\int h(z,z')k_{\mathscr{Z}}(z,z')\,dP_Z(z)\,dP_Z(z') \\ &= \int\int \left[\rho_{\mathscr{Z}}(z,\theta'') + \rho_{\mathscr{Z}}(X',\theta'') - \rho_{\mathscr{Z}}(z,z')\right] h(z,z')\,dP_Z(z)\,dP_Z(z'). \quad (13)\end{aligned}$$

Unfortunately, (13) cannot be related in a simple manner to gCdCov or HSCIC. However, some simplifications occur based on certain assumptions on $k_{\mathscr{Z}}$, as shown in the following corollaries. Under an appropriate choice of $k_{\mathscr{Z}}$, Corollary 3 shows HSC̈IC to be asymptotically equivalent to the weighted average of HSCIC (equivalently, the weighted average of gCdCov) defined in (10) while Corollary 4 shows the asymptotic equivalence between HSC̈IC and CdCov.

Corollary 3 *Suppose the assumptions of Theorem 2 hold and P_Z has a density p_Z w.r.t. the Lebesgue measure on $\mathbb{R}^d$ such that $h(z,\cdot)p_Z$ is uniformly continuous and bounded for all $z \in \mathbb{R}^d$. For $t > 0$, let*

$$k_{\mathscr{Z}}(z,z') = \frac{1}{t^d}\psi\left(\frac{z-z'}{t}\right),\ z,z' \in \mathbb{R}^d,$$

where $\psi \in L^1(\mathbb{R}^d)$ is a bounded continuous positive definite function with $\int_{\mathbb{R}^d}\psi(z)\,dz = 1$. Then

$$\lim_{t\to 0}\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS} = \mathbb{E}_Z[\|\dot{\Sigma}_{YX|Z}\|^2_{HS}\,p_Z(Z)] = \mathcal{D}^2_{P_Z}(P_{XY|Z}, P_{X|Z}P_{Y|Z}).$$

Proof Define $\psi_t(z) := t^{-d}\psi\left(\frac{z}{t}\right)$. From (11), it follows that

$$\begin{aligned}\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS} &= \mathbb{E}_Z\mathbb{E}_{Z'}[\psi_t(Z-Z')h(Z,Z')] \\ &= \int p_Z(z)\left(\int \psi_t(z-z')h(z,z')p_Z(z')\,dz'\right)dz \\ &= \int p_Z(z)(\psi_t * (h(z,\cdot)p_Z)(z)\,dz,\end{aligned}$$

where $*$ denotes convolution. Taking the limit on both sides as $t \to 0$ and applying dominated convergence theorem, we obtain

$$\lim_{t\to 0}\|\Sigma_{Y\ddot{X}|Z}\|^2_{HS} = \lim_{t\to 0}\int p_Z(z)(\psi_t * (h(z,\cdot)p_Z)(z)\,dz = \int p_Z(z)\lim_{t\to 0}(\psi_t * (h(z,\cdot)p_Z)(z)\,dz.$$

The result follows from Folland (1999, Theorem 8.14) which yields $\lim_{t\to 0}(\psi_t * (h(z,\cdot)p_Z)(z) = h(z,z)p_Z(z)$ for all $z \in \mathbb{R}^d$ and by noting that $h(Z,Z) = \|\dot{\Sigma}_{YX|Z}\|^2_{HS}$. ■

Corollary 4 *Suppose the assumptions of Theorem 2 hold with $\rho_{\mathscr{X}}(x,x') = \|x-x'\|$, $x,x' \in \mathbb{R}^p$ and $\rho_{\mathscr{Y}}(y,y') = \|y-y'\|$, $y,y' \in \mathbb{R}^q$. Let $k_{\mathscr{Z}}(z,z') = \eta(z)\eta(z')$, $z,z' \in \mathbb{R}^d$ for some real-valued function η on $\mathbb{R}^d$ and*

$$\mathbb{E}_Z\left[|\eta(Z)|\left\|\phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}\right\|_{L^2(w)}\right] < \infty. \quad (14)$$

Then

$$\|\Sigma_{Y\ddot{X}|Z}\|_{HS}^2 = \left\|\mathbb{E}_Z\left[\eta(Z)\left(\phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}\right)\right]\right\|_{L^2(w)}^2, \tag{15}$$

where $w(t,s) = \frac{1}{c_p c_q}\|t\|^{-p-1}\|s\|^{-q-1}$, $t \in \mathbb{R}^p$, $s \in \mathbb{R}^q$. *In particular, for* $t > 0$ *and some* $a \in \mathbb{R}^d$, *if* $\eta(z) = \frac{1}{t^d}\theta\left(\frac{a-z}{t}\right)$, $z \in \mathbb{R}^d$ *where* θ *is a bounded continuous function with* $\int \theta(z)\,dz = 1$ *and* P_Z *has a bounded uniformly continuous density* p_Z *on* $\mathbb{R}^d$ *such that*

$$\int \operatorname*{ess\,sup}_Z \left|\phi_{XY|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)\right|^2 dw(t,s) < \infty, \tag{16}$$

then

$$\lim_{t\to 0}\|\Sigma_{Y\ddot{X}|Z}\|_{HS}^2 = p_Z^2(a)\mathcal{V}^2(X,Y|Z=a). \tag{17}$$

Proof In the following, we show that

$$h(Z,Z') = \left\langle \phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}, \phi_{XY|Z'} - \phi_{X|Z'}\phi_{Y|Z'}\right\rangle_{L^2(w)} \tag{18}$$

and therefore (15) follows by using (18) in (11) with $k(z,z') = \eta(z)\eta(z')$ and applying dominated convergence theorem through (14). We now prove (18). Consider

$$\begin{aligned}
&\left\langle \phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}, \phi_{XY|Z'} - \phi_{X|Z'}\phi_{Y|Z'}\right\rangle_{L^2(w)} \\
&= \int\int w(t,s)\left[\phi_{XY|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)\right]\left[\overline{\phi_{XY|Z'}(t,s) - \phi_{X|Z'}(t)\phi_{Y|Z'}(s)}\right] dt\,ds \\
&= \int\int w(t,s)\Lambda(t,s)\,dt\,ds,
\end{aligned} \tag{19}$$

where

$$\begin{aligned}
\Lambda(t,s,Z,Z') &= \left[\phi_{XY|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)\right]\left[\overline{\phi_{XY|Z'}(t,s) - \phi_{X|Z'}(t)\phi_{Y|Z'}(s)}\right] \\
&= \left[\mathbb{E}\left[e^{i(\langle t,X\rangle + \langle s,Y\rangle)}|Z\right] - \mathbb{E}\left[e^{i\langle t,X\rangle}|Z\right]\mathbb{E}\left[e^{i\langle s,Y\rangle}|Z\right]\right] \\
&\qquad \cdot\left[\overline{\mathbb{E}\left[e^{i(\langle t,X\rangle + \langle s,Y\rangle)}|Z'\right] - \mathbb{E}\left[e^{i\langle t,X\rangle}|Z'\right]\mathbb{E}\left[e^{i\langle s,Y\rangle}|Z'\right]}\right] \\
&= \mathbb{E}_{XY|Z}\mathbb{E}_{X'Y'|Z'}e^{i(\langle t,X-X'\rangle + \langle s,Y-Y'\rangle)} - \mathbb{E}_{XY|Z}\mathbb{E}_{X'|Z'}\mathbb{E}_{Y'|Z'}e^{i(\langle t,X-X'\rangle + \langle s,Y-Y'\rangle)} \\
&\qquad -\mathbb{E}_{X|Z}\mathbb{E}_{Y|Z}\mathbb{E}_{X'Y'|Z'}e^{i(\langle t,X-X'\rangle + \langle s,Y-Y'\rangle)} \\
&\qquad\qquad +\mathbb{E}_{X|Z}\mathbb{E}_{Y|Z}\mathbb{E}_{X'|Z'}\mathbb{E}_{Y'|Z'}e^{i(\langle t,X-X'\rangle + \langle s,Y-Y'\rangle)} \\
&= \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}e^{i(\langle t,X-X'\rangle + \langle s,Y-Y'\rangle)},
\end{aligned} \tag{20}$$

where $\mathbb{F}_{YX|Z} := \mathbb{E}_{XY|Z} - \mathbb{E}_{Y|Z}\mathbb{E}_{X|Z}$. Using (20) in (19), we obtain

$$\int\int w(t,s)\Lambda(t,s,Z,Z')\,dt\,ds = \int\int \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}\cos\langle t, X-X'\rangle\cos\langle s, Y-Y'\rangle w(t,s)\,dt\,ds \tag{21}$$

by noting that $\sin\langle t, X - X'\rangle$ and $\sin\langle s, Y - Y'\rangle$ are odd functions w.r.t. t and s respectively. Since $\cos\langle t, X - X'\rangle \cos\langle s, Y - Y'\rangle = 1 - (1 - \cos\langle t, X - X'\rangle) - (1 - \cos\langle s, Y - Y'\rangle) + (1 - \cos\langle t, X - X'\rangle)(1 - \cos\langle s, Y - Y'\rangle)$ and

$$\mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'} \cdot [f(X, X', Y, Y')] = 0$$

for $f(X, X', Y, Y') = 1$, $f(X, X', Y, Y') = 1 - \cos\langle t, X - X'\rangle$ and $f(X, X', Y, Y') = 1 - \cos\langle s, Y - Y'\rangle$, (21) reduces to

$$\begin{aligned}
&\int\int w(t,s)\Lambda(t,s,Z,Z')\,dt\,ds \\
&= \int\int \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}\left[\frac{1-\cos\langle t, X-X'\rangle}{c_p\|t\|^{p+1}} \cdot \frac{1-\cos\langle s, Y-Y'\rangle}{c_q\|s\|^{q+1}}\right] dt\,ds \\
&= \mathbb{F}_{YX|Z}\mathbb{F}_{Y'X'|Z'}[\|X-X'\|\|Y-Y'\|] \\
&= h(Z, Z'),
\end{aligned}$$

where the last equality follows from Lemma 1 of Székely et al. (2007) through $\int \frac{1-\cos\langle t,x\rangle}{c_p\|t\|^{p+1}}\,dt = \|x\|$, thereby proving the result in (15). By defining $\theta_t(z) = t^{-d}\theta\left(\frac{z}{t}\right)$, we have

$$\mathbb{E}_Z[\eta(Z)\left(\phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}\right)] = \theta_t * \left(\left(\phi_{XY|Z} - \phi_{X|Z}\phi_{Y|Z}\right)p_Z\right)(a),$$

which by (Folland, 1999, Theorem 8.14) converges to $\left(\phi_{XY|Z=a} - \phi_{X|Z=a}\phi_{Y|Z=a}\right)p_Z(a)$ as $t \to 0$. Using these in (15) along with dominated convergence theorem combined with (16) yields (17). ■

Remark 5 *Informally, the result of Corollary 3 can be obtained by choosing* $k_{\mathscr{Z}}(z, z') = \delta(z - z')$, $z, z' \in \mathbb{R}^d$, *where* $\delta(\cdot)$ *is the Dirac distribution. Since such a choice does not correspond to a valid reproducing kernel—Dirac distribution is not a function but a* distribution *that does not belong to an RKHS—, the rigorous argument involves considering a family of kernels indexed by bandwidth* t *which in the limiting case of* $t \to 0$ *achieves the behavior of the Dirac distribution. Similar argument applies to Corollary 4 as well.*

6. Discussion

Conditional distance covariance is a commonly used metric for measuring conditional dependence in the statistics community. In the machine learning community, a conditional dependence measure based on reproducing kernels is popularly used in applications such as conditional independence testing. In this work, we have explored the connection between these two conditional dependence measures where we showed the distance-based measure to be a limiting version of the kernel-based measure, where we may view conditional distance covariance as a member of a much larger class of kernel-based conditional dependence measures. This may enable to design more powerful conditional independence tests by choosing a richer class of kernels.

Having understood the relation between these various measures of conditional dependence, an important question to understand is the statistical behavior of conditional independence tests based on these measures. Fukumizu et al. (2004, Proposition 5) provides an alternate

representation for the conditional covariance operator, $\Sigma_{Y\ddot{X}|Z}$ in terms of only covariance operators (this is reminiscent of the situation when (X, Y, Z) are jointly normal so that the conditional covariance matrix can be represented in terms of the joint covariance matrices) as $\Sigma_{Y\ddot{X}|Z} = \Sigma_{Y\ddot{X}} - \Sigma_{YZ}\tilde{\Sigma}_{ZZ}^{-1}\Sigma_{Z\ddot{X}}$ where $\tilde{\Sigma}_{ZZ}^{-1}$ is the right inverse of Σ_{ZZ} on $(\mathrm{Ker}(\Sigma_{ZZ}))^{\perp}$. The advantage of this alternate form is that $\Sigma_{Y\ddot{X}|Z}$ can be estimated from data $(X_i, Y_i, Z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XYZ}$ by simply estimating the (cross) covariance operators $\Sigma_{Y\ddot{X}}$, Σ_{YZ}, $\Sigma_{Z\ddot{X}}$, and replacing $\tilde{\Sigma}_{ZZ}^{-1}$ by an inverse of the regularized version of an empirical estimator of Σ_{ZZ}. Using these, a plug-in (biased) estimator $\|\hat{\Sigma}_{Y\ddot{X}|Z}\|_{HS}^2$ of HS$\ddot{\text{C}}$IC (i.e., $\|\Sigma_{Y\ddot{X}|Z}\|_{HS}^2$), can be shown to be consistent and to have a computational complexity of $O(n^3)$, where $\hat{\Sigma}_{Y\ddot{X}|Z} := \hat{\Sigma}_{Y\ddot{X}} - \hat{\Sigma}_{YZ}(\hat{\Sigma}_{ZZ} + \lambda I)^{-1}\hat{\Sigma}_{Z\ddot{X}}$ and $\lambda > 0$—these claims can be proved using the ideas in Fukumizu et al. (2008) where such claims are proved for a normalized version of $\Sigma_{Y\ddot{X}|Z}$. Similar results are shown for the kernel version of HSCIC (see (7)) by Park and Muandet (2020). To elaborate, (Park and Muandet, 2020, Section 5.2) proposed a biased estimator of HSCIC (see r.h.s. of (7)), which is based on Gram matrices on $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ and associated regularized inverse, yielding a computational complexity of $O(n^3)$. On the other hand, Wang et al. (2015) proposed a (biased) estimator of CdCov—the same idea can be used to estimate gCdCov and therefore HSCIC—based on a Nadarya-Watson type density estimator of $P_{XY|Z}$, where it can be shown that HSCIC can be consistently estimated with a computational complexity of $O(n^3)$. This means, all these different estimators of HSCIC and HS$\ddot{\text{C}}$IC are consistent and have same computational complexity. However, the statistical performance of these estimators as test statistics to test for conditional independence remains open.

Acknowledgements

BKS is partially supported by National Science Foundation (NSF) award DMS-1713011 and CAREER award DMS-1945396.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45, 2021.

R. D. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, USA, 1999.

K. Fukumizu, F. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

K. Fukumizu, A. Gretton, Xiaohai S., and B. Schölkopf. Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496. Curran Associates, Inc., 2008.

K. Fukumizu, A. Gretton, B. Schölkopf, and B. K. Sriperumbudur. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 473–480. Curran Associates, Inc., 2009.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, pages 63–77, Berlin, Heidelberg, 2005. Springer-Verlag.

A. Gretton, K. Borgwardt, R. Malte, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. Curran Associates, Inc., 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10 (1-2):1–141, 2017.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, USA, 2000.

D. Sejdinovic, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

C-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

C-J. Simon-Gabriel, A. Barp, B. Schölkopf, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. 2020. https://arxiv.org/pdf/2006.09268.pdf.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In M. Hutter, R. A. Servedio, and E. Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, Prediction, and Search*. MIT press, Cambridge, MA, USA, 2000.

B. K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2011.

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

Z. Szabó and B. K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, (5), 2004.

G. Székely and M. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 4 (3):1233–1303, 2009.

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

Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The Annals of Applied statistics*, 3(4):1236–1265, 2009.

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.