Self-Discrepancy Conditional Independence Test

Sanghack Lee and **Vasant Honavar**
Artificial Intelligence Research Laboratory
College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802
{sxl439, vhonavar}@ist.psu.edu

Abstract

Tests of conditional independence (CI) of random variables play an important role in machine learning and causal inference. Of particular interest are kernel-based CI tests which allow us to test for independence among random variables with complex distribution functions. The efficacy of a CI test is measured in terms of its power and its calibratedness. We show that the Kernel CI Permutation Test (KCIPT) suffers from a loss of calibratedness as its power is increased by increasing the number of bootstraps. To address this limitation, we propose a novel CI test, called Self-Discrepancy Conditional Independence Test (SDCIT). SDCIT uses a test statistic that is a modified unbiased estimate of maximum mean discrepancy (MMD), the largest difference in the means of features of the given sample and its permuted counterpart in the kernel-induced Hilbert space. We present results of experiments that demonstrate SDCIT is, relative to the other methods: (i) competitive in terms of its power and calibratedness, outperforming other methods when the number of conditioning variables is large; (ii) more robust with respect to the choice of the kernel function; and (iii) competitive in run time.

1 INTRODUCTION

Random variables X and Y are said to be conditionally independent given Z, denoted by $X \perp\!\!\!\perp Y \mid Z$, if a joint probability distribution $\mathrm{P}_{xyz} = \mathrm{P}_{xz}\mathrm{P}_{y|xz}$ can be expressed as $\mathrm{P}_{xz}\mathrm{P}_{y|z}$. Tests of conditional independence (CI) play a central role in statistics (Dawid, 1979), machine learning including dimensionality reduction (Fukumizu et al., 2004, 2009), independent component analysis (Bach and Jordan, 2002), probabilistic graphical models (Koller and Friedman, 2009) and causal inference (Pearl, 2000; Spirtes et al., 2000).

In principle, testing for conditional independence is quite straightforward: Obtain a sample S of i.i.d. observations drawn from a distribution P_{xyz}, and evenly split into two subsamples S_1 and S_2 of equal size. Leaving S_1 is intact, permute S_2 so as to simulate a sample from $\mathrm{P}_{xz}\mathrm{P}_{y|z}$ (i.e., $X \perp\!\!\!\perp Y \mid Z$); Apply a two-sample test to determine if S_1 and S_2 are different. Based on the results of the test if the null hypothesis $\mathrm{P}_{xyz} = \mathrm{P}_{xz}\mathrm{P}_{y|z}$ cannot be rejected, we conclude that X is independent of Y given Z. This procedure can be repeated multiple times (i.e., bootstrap Efron, 1979) to improve the power of the test.

In order for the preceding approach to CI testing is feasible, we need an effective two-sample test to determine if S_1 and S_2 are different. Gretton et al. (2012) introduced a framework for designing such tests using a well-behaved (e.g., smooth) function, which is large on the points drawn from one distribution and small on the points from the other distribution. The framework uses as test statistic, MMD, the largest difference between the mean function values on the two samples; when MMD is large, the samples are likely from different distributions. In order for MMD to be effective in practice, the class of functions used to define it should be (i) rich enough to ensure that the MMD vanishes if and only if the two distributions being compared are identical; and (ii) sufficiently restricted so as to ensure that the empirical estimate of MMD converges quickly to its expected value as the sample size is increased. As shown by Gretton et al. (2012), these requirements are met by unit balls in reproducing kernel Hilbert spaces (RKHS).

Doran et al. (2014) introduced the Kernel Conditional Independence Permutation Test (KCIPT), using MMD in a kernel-induced feature space as the test statistic for determining whether $X \perp\!\!\!\perp Y \mid Z$. Doran et al. (2014)

evaluated several kernel-based independence tests based on how well a test correctly rejects the null hypothesis $\mathrm{P}_{xyz} = \mathrm{P}_{xz}\mathrm{P}_{y|z}$ by estimating the *power* of the test (as measured by the Area Under the Power Curve) and the *calibratedness* of the test, i.e., the extent to which it accurately estimates the probability distribution of the test statistic under the null hypothesis (as measured by a Kolmogorov-Smirnov divergence of the observed distribution of p-values from the uniform distribution). Using the preceding procedure, with the number of bootstraps B set equal to 25, Doran et al. (2014) asserted that KCIPT "has power competitive with existing kernel-based approaches" and that it is well-calibrated compared to other kernel independence tests such as KCIT (Zhang et al., 2011) and CHSIC (Fukumizu et al., 2008).

Because increasing the number of bootstraps *always* improves its power, it is natural to ask: How does the power and calibratedness of KCIPT change as a function of B? We present results that show that as B increases, the calibratedness of KCIPT degrades. This suggests that increase in the power of KCIPT comes at the expense of its calibratedness. Based on our analysis of the limitations of KCIPT, we propose a new CI test, which we call the Self-Discrepancy CI Test (SDCIT). SDCIT is based on a modified unbiased estimate of MMD and its distribution under the null hypothesis based on half-sampling without replacement. We present results of experiments that demonstrate several advantages of SDCIT over the existing kernel-based CI tests.

2 PRELIMINARIES

We mostly follow notational conventions of (Doran et al., 2014): We use upper case letters, e.g., X, to denote random variables, and the corresponding lowercase letters $x \in \mathcal{X}$ to denote an instantiation of X in its domain $\mathcal{X}$. We use a bold lowercase letters to denote sets (or vectors) of instantiations, e.g., $\mathbf{x} = (x_i)_{i=1}^n$. We denote by $x \sim \mathbf{x}$ the fact that x is an observation of a random variable sampled from the empirical distribution constructed from the finite sample $\mathbf{x}$. We denote by $\binom{\mathbf{x}}{\ell}$ a set of all ℓ-sized subsets of $\mathbf{x}$. Let $\boldsymbol{\pi}(n)$ be a group of all possible permutations of $(1, 2, \ldots, n)$ with an additional constraint that every permutation $\pi \in \boldsymbol{\pi}(n)$ satisfies that $\forall_{i=1}^n \pi(i) \neq i$. We denote the application of a permutation π to a sequence $\mathbf{y}$ by $\pi\mathbf{y} := (y_{\pi(i)})_{i=1}^n$. We use μ, the average function, to denote the average of a given parameter, e.g., a vector, set, or matrix.

2.1 KERNEL TWO-SAMPLE TEST

Let k_x be a kernel function $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space induced by a kernel k. For instance, $k_x(x', x'') = \langle \phi_x(x'), \phi_x(x'') \rangle_{\mathcal{H}_{k_x}}$ where ϕ_x is a feature mapping from $\mathcal{X}$ to $\mathcal{H}_{k_x}$. We use K to denote a Gram matrix (i.e., kernel matrix) corresponding to a kernel function k. Thus, $\mathrm{K}_x(\mathbf{x}, \mathbf{x}')_{ij} := k_x(x_i, x'_j)$. We use k_{xy} to denote the product of kernels k_x and k_y and define it as $k_{xy}((x_i, y_i), (x_j, y_j)) := k_x(x_i, x_j) k_y(y_i, y_j)$. Kernel mean embedding (KME) allows us to represent a probability distribution as a point in the kernel-induced Hilbert space (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Sriperumbudur et al., 2010), $\mu_{\mathrm{P}} := \int k(\cdot, x)\, \mathrm{P}(dx) \in \mathcal{H}_k$. Maximum mean discrepancy (MMD) is an integral probability metric that provides a measure of distance between the two probability distributions P and Q: $\mathrm{MMD}^2[\mathcal{F}, \mathrm{P}, \mathrm{Q}] := \left[\sup_{f \in \mathcal{F}} (\mathbf{E}_{x \sim \mathrm{P}} f(x) - \mathbf{E}_{x \sim \mathrm{Q}} f(x))\right]^2 = \|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|^2_{\mathcal{H}_{k_x}}$ where $\mathcal{F}$ is a unit ball in $\mathcal{H}_{k_x}$. Given a characteristic kernel, $\mathrm{MMD}(\mathrm{P}, \mathrm{Q})$ is 0 if and only if distributions $\mathrm{P} = \mathrm{Q}$ (Sriperumbudur et al., 2010). Given two independent samples of the same size, $\mathbf{x}^{(1)} = \{x_i^{(1)}\}_{i=1}^m$ and $\mathbf{x}^{(2)} = \{x_i^{(2)}\}_{i=1}^m$, let $u_i := (x_i^{(1)}, x_i^{(2)})$. Then, an empirical unbiased estimate of squared MMD

$$\mathrm{MMD}_u^2(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) := \frac{1}{m(m-1)} \sum_{i \neq j=1}^m h(u_i, u_j) \quad (1)$$

is a one-sample unbiased statistic (U-statistic) where

$$h(u_i, u_j) := k(x_i^{(1)}, x_j^{(1)}) + k(x_i^{(2)}, x_j^{(2)}) - k(x_i^{(1)}, x_j^{(2)}) - k(x_j^{(1)}, x_i^{(2)}).$$

A kernel two-sample test (Gretton et al., 2007, 2009, 2012) uses MMD as the test statistic and its distribution under the null hypothesis to test for homogeneity. The distribution of MMD can be estimated in a number of ways including moment-based approximation (Gretton et al., 2012), Gram matrix spectrum (Gretton et al., 2009), and resampling procedure. For instance, KCIPT uses a bootstrap procedure to repeatedly measure the MMD between two equal-sized random subsamples of the union of the two given samples.

3 KCIPT

We now proceed to describe KCIPT and discuss its limitations that motivate our proposal for SDCIT. Suppose we are given a sample $\Omega = \{(x_i, y_i, z_i)\}_{i=1}^n = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ of n observations drawn i.i.d. from P_{xyz}. We split Ω randomly into two subsamples of equal size: $\Omega^{(1)} = (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \mathbf{z}^{(1)})$ and $\Omega^{(2)} = (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}, \mathbf{z}^{(2)})$. The second split sample $\Omega^{(2)}$ is permuted so as to simulate sample from $\mathrm{P}_{xz}\mathrm{P}_{y|z}$.[1] Such permutation $\pi \in \boldsymbol{\pi}(n/2)$ must

[1] One can either permute X or Y. For consistency, we will stick with permuting Y as in (Doran et al., 2014).

Algorithm 1 KCIPT (Doran et al., 2014) with an *unbiased* estimator of MMD.

Input: B: the number of bootstraps, b: the size of sample for a null distribution per bootstrap, M: the size of sample for a bootstrap null distribution, Ω: a sample $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ of size n

1: **for** $i \in 1 \cdots B$ **do**
2: $\quad \Omega_i^{(1)}, \Omega_i^{(2)} \leftarrow$randomly split Ω evenly
3: $\quad \pi \leftarrow$ learn a permutation given δ and $\mathbf{z}^{(2)}$
4: $\quad \Omega_{i,\pi}^{(2)} \leftarrow$ apply π to Y values of $\Omega_i^{(2)}$
5: $\quad T_i \leftarrow \text{MMD}_u^2(\Omega_i^{(1)}, \Omega_{i,\pi}^{(2)})$
6: $\quad$ **for** $j \in 1 \cdots b$ **do**
7: $\quad\quad \Omega', \Omega'' \leftarrow$randomly split $\Omega_i^{(1)} \cup \Omega_{i,\pi}^{(2)}$ evenly
8: $\quad\quad (\mathbf{N})_{ij} \leftarrow \text{MMD}_u^2(\Omega', \Omega'')$
9: $T \leftarrow$average of $\{T_i\}_{i=1}^B$
10: $\mathbf{N} \leftarrow \{$average of $\{s_{ij}\}_{i=1}^B \mid \forall_{i=1}^B s_{ij} \sim \mathbf{N}_i\}_{j=1}^M$
11: **return** a p-value of T from $\mathbf{N}$

satisfy that $z_i^{(2)} = z_{\pi(i)}^{(2)}$ to preserve the relationship between Y and Z while breaking ties between X and Y. In many interesting cases, including $\mathcal{Z} = \mathbb{R}$, $\mathbf{z}$ often consists of unique values and such a permutation may not exist. Hence, under the assumption $\mathrm{P}(Y \mid z) \approx \mathrm{P}(Y \mid z')$ if $z \approx z'$, we relax the requirement to learn a permutation $\pi \in \boldsymbol{\pi}(n/2)$ by minimizing $\delta(\mathbf{z}^{(2)}, \pi\mathbf{z}^{(2)})$ where δ is a user-defined distortion measure. A common choice for δ is the sum of distances between permuted $\mathbf{z}$ values, $\sum_{i=1}^{n/2} d_z(z_i^{(2)}, z_{\pi(i)}^{(2)})$ where d_z is induced by k_z, i.e., $d_z^2(z', z'') := k_z(z', z') + k_z(z'', z'') - 2k_z(z', z'')$ or a regression-based distance $d_z(z', z'') := \|f(z') - f(z'')\|_2$ where f is a function relating Z and Y that can be learned, for instance, using a Gaussian process regression (GPR) (Zhang et al., 2011; Doran et al., 2014).

We apply the learned permutation π to $\mathbf{y}^{(2)}$ in $\Omega^{(2)}$ to obtain $\Omega_\pi^{(2)} := (\mathbf{x}^{(2)}, \pi\mathbf{y}^{(2)}, \mathbf{z}^{(2)})$. We then perform a two-sample test between $\Omega^{(1)}$ and $\Omega_\pi^{(2)}$ where we compute a p-value using $\text{MMD}_u^2(\Omega^{(1)}, \Omega_\pi^{(2)})$ as the test statistic. We obtain the empirical distribution of the p-value under the null hypothesis by repeatedly measuring MMD_u^2 between the split samples $\Omega^{(1)}$ and $\Omega_\pi^{(2)}$ (see line 6–8).

We can repeat the test to increase the power of KCIPT. Suppose we bootstrap the aforementioned two-sample test B times. Let T_i and $\mathbf{N}_i$ be the ith MMD_u^2 estimate and a corresponding null distribution (called the 'inner' null distribution). The bootstrapped test statistic is simply the average of test statistics over each of the splits, i.e., $\mu(\{T_i\}_{i=1}^B)$. The null distribution of the bootstrapped test statistic is obtained using a Monte Carlo simulation by averaging together the draws from each individual statistic's null distribution, $\{\mu(\{s_{ij}\}_{i=1}^B) \mid \forall_{i=1}^B s_{ij} \sim \mathbf{N}_i\}_{j=1}^M$, where M is the number of points drawn from each of the B null distributions.

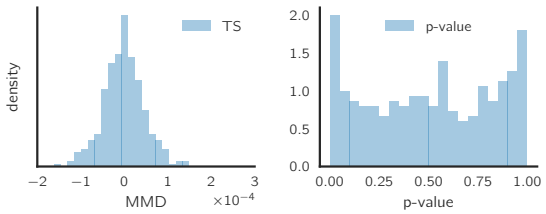

Figure 1: The distribution of test statistics of KCIPT with $B = 5000$ on 300 samples generated under the null hypothesis and a corresponding p-value distribution where B around 2000 is a good choice to balance between power and calibratedness (see Section 5.4)

Doran et al. (2014) evaluated three kernel-based CI tests, KCIT, CHSIC, and KCIPT, based on how well a test correctly rejects the null hypothesis (power) and how similar are the distribution of p-values under the null hypothesis and the uniform distribution between 0 and 1 (calibratedness). Doran et al. (2014) concluded that KCIPT "has power competitive with existing kernel-based approaches" and that it is well-calibrated.

3.1 LIMITATIONS OF KCIPT

Clearly, the larger the number *B* of bootstraps, the larger the power of KCIPT. However, we observe that the increase in the power of KCIPT comes at the expense of its calibratedness. Consider for example, the case where the expected test statistic is a small positive number. As *B* increases, the test statistic converges to its expected value (which is close to 0) and the null distribution of the test statistic will be degenerate at 0. Consequently, KCIPT with a sufficiently large *B* will reject the null hypothesis more often than it should. See Figure 1, where KCIPT with $B = 5000$ results around 10% of type I error given $\alpha = 0.05$. We will examine this phenomenon more closely later in Section 5.

Further, different runs of KCIPT with different random seeds will yield different *random* splits of a given sample (line 2, Algorithm 1) and, hence, potentially different p-values. If the given sample exhibits strong conditional dependence, the differences in random splits have little impact on the resulting p-values. However, if the sample is generated under the null hypothesis, the p-values follow a distribution close to the uniform distribution between 0 and 1 (unless B is so large that KCIPT simply loses its calibratedness). This problem makes it difficult to interpret or reproduce the p-values returned by KCIPT. This calls into question the conclusion of Doran et al. (2014) that KCIPT is well-calibrated.

In summary, KCIPT, in the absence of clear guidance on how to determine the optimal number of bootstraps, fails

Algorithm 2 MMSD

Input K_{xz}, K_y: Gram matrices for $\{(x_i, z_i)\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$; D: a pairwise distance matrix for $\mathbf{z}$

1: $\pi \leftarrow$find $\pi \in \boldsymbol{\pi}(n)$ minimizing δ given D
2: $\mathcal{I} \leftarrow \{(i,j) \mid 1 \leq i \neq j \leq n,\, i \neq \pi(j),\, j \neq \pi(i)\}$
3: $\mathrm{K}_1, \mathrm{K}_2, \mathrm{K}_3 \leftarrow \mathrm{K}_{xz} \odot \mathrm{K}_y, \mathrm{K}_{xz} \odot \mathrm{K}_y[\pi, \pi], \mathrm{K}_{xz} \odot \mathrm{K}_y[:, \pi]$
4: $T \leftarrow |\mathcal{I}|^{-1} \cdot \sum_{i,j \in \mathcal{I}} (\mathrm{K}_1 + \mathrm{K}_2 - \mathrm{K}_3 - \mathrm{K}_3^\top)_{i,j}$
5: **return** T, π, $\mathcal{I}$

to provide informative p-values when it is either underpowered (small B) or not well calibrated (large B).

4 SELF-DISCREPANCY CI TEST

We proceed to introduce Self-Discrepancy Conditional Independence Test (SDCIT), a new permutation and kernel-based CI test, using a new test statistic MMSD, which is based on an unbiased squared MMD estimate. We specify the empirical distribution of the MMSD and that under the null hypothesis. We further establish the *asymptotic consistency* of MMSD as a test statistic for CI (its convergence in probability to zero if and only if the null hypothesis holds).

4.1 MAXIMUM MEAN SELF-DISCREPANCY

One way to get around the limitations of KCIPT noted above is to ensure that the test statistic is determined by the given sample Ω, and *not* the any particular random splits of Ω, and thereby eliminate the dependence of the result of KCIPT on the choice of the random seed (line 2, Algorithm 1). We proceed to show how this can be achieved by replacing MMD_u^2 estimated from the samples obtained by randomly splitting Ω into two parts where one part is left intact and the other is permuted to break the ties that violate conditional independence by a variant of MMD_u^2 between the given sample Ω and its permuted counterpart. We learn a permutation π of Ω by minimizing distortion measure δ on $\mathbf{z}$. Let Ω_π be a sample where the permutation π is applied to $\mathbf{y}$ of Ω. However, because Ω_π is *not* obtained independently from Ω (see Equation 1), we cannot naively measure MMD_u^2 between Ω and Ω_π. Hence, we introduce a new test statistic, which we call Maximum Mean Self-Discrepancy (MMSD), which estimates the discrepancy between a sample Ω and its conditionally independent counterpart Ω_π by removing spurious correlations between elements of Ω and Ω_π arising from the dependence of Ω_π on Ω. Let

$$h(i,j) := k_{xyz}((\Omega)_i, (\Omega)_j) + k_{xyz}((\Omega_\pi)_i, (\Omega_\pi)_j) - k_{xyz}((\Omega)_i, (\Omega_\pi)_j) - k_{xyz}((\Omega)_j, (\Omega_\pi)_i)$$

where $(\cdot)_i$ represents the ith observation of the argument. We count $h(i,j)$ only if two triples from the two observations in each of four terms are independently obtained. For example, we exclude the case $i = j$ since $(x_i, y_i, z_i) \not\perp\!\!\!\perp (x_j, y_j, z_j)$ if $i = j$ based on the first term. Similarly, we exclude the case $i = \pi(j)$ since $(x_i, y_i, z_i) \not\perp\!\!\!\perp (x_j, y_{\pi(j)}, z_j)$ based on the third term. Now, given a permutation π, we denote by

$$\mathcal{I} := \{(i,j) \mid i \neq \pi(j),\, j \neq \pi(i)\}_{i \neq j = 1}^n$$

a set of pairs of indices of independent observations conditioning on π. Based on the extent to which the permutation is reciprocal (i.e., $i = (\pi \circ \pi)(i)$), the size of $\mathcal{I}$ ranges from $n^2 - 3n$ to $n^2 - 2n$. We proceed to estimate MMSD as follows:

$$T := \frac{1}{|\mathcal{I}|} \sum_{i,j \in \mathcal{I}} h(i,j). \tag{2}$$

MMSD is closely related to the expectation of the test statistic of KCIPT. However, MMSD depends only on a single learned permutation π and eliminates the need for bootstrapping. We now proceed to prove the asymptotic consistency of MMSD by extending a theorem in (Doran et al., 2014).

Theorem 1. *Let the kernel k_{xyz} be universal and the sample space be compact. Given $\max_{\Omega_i \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})} \|\phi_{xyz}(\Omega_i)\| \leq C$ for some constant C, the test statistic MMSD is asymptotically consistent if the distortion measure based on RKHS distance, $\frac{1}{n}\sum_{i=1}^n \|\phi(z_i) - \phi(z_{\pi(i)})\|$, converges in probability to zero as $n \to \infty$.*

Proof. Based on Theorem 1 (Doran et al., 2014), the embedding of permuted sample Ω_π converges to the embedding of $\mathrm{P}_{xz}\mathrm{P}_{y|z}$. Thus, $\mathrm{MMD}_u^2(\Omega, \Omega_\pi) \xrightarrow{\mathrm{P}} 0$ under the null hypothesis. Since $h(i,j) \leq 4C^2$ and there are at most $2n$ excluded values in $\mathcal{I}$ except the diagonal elements, $\left|\mathrm{MMD}_u^2(\Omega, \Omega_\pi) - \mathrm{MMSD}(\Omega)\right| \leq \frac{2n \cdot 4C^2}{n(n-1)} = \frac{8C^2}{n-1}$. Hence, $\mathrm{MMSD}(\Omega) \xrightarrow{\mathrm{P}} 0$. □

Note that C is equal to 1 if a Gaussian kernel is used. We describe the pseudocode for MMSD estimate in Algorithm 2 where $\mathbf{M}[\mathbf{a}, \mathbf{b}]$ denotes a submatrix of $\mathbf{M}$ where its rows and columns are selected and ordered by $\mathbf{a}$ and $\mathbf{b}$, respectively, that is, $(\mathbf{M}[\mathbf{a}, \mathbf{b}])_{i,j} = \mathrm{M}_{a_i, b_j}$. A colon ':' represents entire rows or columns. An operator $\odot$ denotes a Hadamard product (i.e., element-wise multiplication).

4.2 EMPIRICAL DISTRIBUTION OF MMSD

We now turn our attention to obtaining the empirical distribution of MMSD, the test statistic used by SDCIT. Although the distribution of MMD_u^2 has been thoroughly

Algorithm 3 EMPMMSD

Input K_{xz}, K_y: Gram matrices for $\{(x_i, z_i)\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$; D: a pairwise distance matrix for $\mathbf{z}$; b: a number of samples to be generated.

$\mathbf{T}' \leftarrow$initialize a vector of size b.
for b times **do**
  $\mathbf{i} \leftarrow$randomly choose $\frac{n}{2}$ unique integers from $(1, \ldots, n)$
  $\mathbf{T}'_i, \cdot, \cdot \leftarrow \textsc{mmsd}(\mathrm{K}_{xz}[\mathbf{i}, \mathbf{i}], \mathrm{K}_y[\mathbf{i}, \mathbf{i}], \mathrm{D}[\mathbf{i}, \mathbf{i}])$
return $\frac{1}{2} \cdot (\mathbf{T}' - \mu(\mathbf{T}')) + \mu(\mathbf{T}')$

Algorithm 4 SDCIT

Input K_{xz}, K_y: Gram matrices for $\{(x_i, z_i)\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$; D: a pairwise distance matrix for $\mathbf{z}$; b: the size of empirical null distribution.

$T, \pi, \mathcal{I} \leftarrow \textsc{mmsd}(\mathrm{K}_{xz}, \mathrm{K}_y, \mathrm{D})$
$\mathrm{D}_{\mathcal{I}} \leftarrow$D with $\mathrm{D}_{i,j} = \infty$ for every $(i,j) \in \mathcal{I}$
$\cdot, \pi_2, \mathcal{I}_2 \leftarrow \textsc{mmsd}(\mathrm{K}_{xz}, \mathrm{K}_y, \mathrm{D}_{\mathcal{I}})$
$\mathrm{D}_{\mathcal{I}_2} \leftarrow$D with $\mathrm{D}_{i,j} = \infty$ for every $(i,j) \in \mathcal{I}_2$
$\mathbf{N} \leftarrow \textsc{empmmsd}(\mathrm{K}_{xz}, \mathrm{K}_y[\pi_2, \pi_2], \mathrm{D}_{\mathcal{I}_2}, b)$
return a p-value with T and $\mathbf{N} - \mu(\mathbf{N})$.

analyzed (Gretton et al., 2012), it is not at all immediately obvious how the analysis of MMD_u^2, which is based on two independent distributions, is applicable to the case of MMSD, which is defined with respect to two *dependent* distributions. Unlike other point estimates such as mean, MMSD on a bootstrap sample can be problematic: it is possible that pairs of the repeated observations in a bootstrap sample are permuted to each other thereby violating the conditional independence desired in the permuted sample. The resulting test statistic from such bootstrap sample will be closer to 0 than it should be.

Hence, we consider an alternative approach to estimating MMSD using sampling without replacement. We exploit the observation that half-sampling without replacement yields a very similar result to bootstrapping (Buja and Stuetzle, 2006; Friedman and Hall, 2007). Let $\boldsymbol{\Omega'}$ be b half-samples via sampling without replacement procedure, $\boldsymbol{\Omega'} := \{\Omega'_i \mid \Omega'_i \sim \binom{\Omega}{n/2}\}_{i=1}^b$. We compute the test statistic on each of randomly chosen half-samples. Since the estimate is based on the average of less than $n^2/4$ values, the expected distribution will be about twice as wide as the empirical distribution of test statistic, which is the average of fewer than n^2 values. Hence, the empirical distribution of MMSD is obtained by shrinking the width of the distribution by half: $\mathbf{T} := \frac{1}{2}(\mathbf{T}' - \mu(\mathbf{T}')) + \mu(\mathbf{T}')$ where $\mathbf{T}' := \{\textsc{mmsd}(\Omega', \mathrm{D}) \mid \Omega' \in \boldsymbol{\Omega'}\}$. A pseudocode is given in Algorithm 3. Now, we provide an approximate null distribution based on the analysis in this section.

4.3 APPROXIMATE NULL DISTRIBUTION OF MMSD

An approximate null distribution of test statistic can often be obtained by applying the statistic on many samples generated under the null hypothesis ($\mathrm{P}_{xz}\mathrm{P}_{y|z}$ in this case). Since we do have access to neither $\mathrm{P}_{xz}\mathrm{P}_{y|z}$ nor the model for the distribution, we will apply half-sampling without replacement on the permuted sample Ω_π to approximate the null distribution of MMSD. As described in the previous section, we generate $\boldsymbol{\Omega'_\pi}$, b half-samples of Ω_π via sampling without replacement. However, we need to be careful in obtaining the test statistic from the resulting half samples. Let a half-sample $\Omega'_\pi = \{(x'_c, y'_c, z'_c)\}_{c=1}^{n/2} \in \boldsymbol{\Omega'_\pi}$, where $\Omega'_\pi \subset \Omega_\pi$, contain the following two observations at its ath and bth indices, $(x'_a, y'_a, z'_a) = \left(x_i, y_{\pi(i)}, z_i\right) = (x_i, y_j, z_i)$ and $(x'_b, y'_b, z'_b) = \left(x_k, y_{\pi(k)}, z_k\right) = (x_k, y_i, z_k)$ where indices i, j, and k are originally from Ω. We can infer that $z_j \approx z_i \approx z_k$. Hence, a permutation π' to be learned for Ω'_π will likely permute y's in the two observations. If that is the case (i.e., $\pi'(a) = b$), then the permuted half-sample $(\mathbf{x}', \pi'\mathbf{y}', \mathbf{z}')$ will contain the observation $(x'_a, y'_{\pi'(a)}, z'_a) = (x_i, y'_b, z_i) = (x_i, y_i, z_i)$. That is, applying the permutation π' recovers an original observation in Ω failing to simulate conditional independence in $\Omega'_{\pi\pi'}$. Hence, the following constraints should be imposed in computing a permutation for Ω'_π, $\forall_{1 \le c \le \frac{n}{2}} (f^{-1} \circ \pi \circ f \circ \pi')(c) \neq c$ where f is a mapping between the index of an observation in a subsample to that in Ω_π. In the preceding example, $\pi'(a) = b$, $f(b) = k$, $\pi(k) = i$, and $f^{-1}(i) = a$. In other words, simply disallow permutations that associate y_i back to $(x_i, \cdot, z_i)$. In addition to constraints on the allowed permutations, identifying independent observation pairs is more complicated by the need for additional bookkeeping for tracking the composition of every observation in the half-samples of Ω_π.

Hence, we consider an alternative approach to using the permuted sample Ω_π in MMSD by modifying the distance on $\mathbf{z}$. If we have a pre-computed pairwise distance matrix from the previous step, setting the entries of the matrix at index pairs that appear in $\mathcal{I}$ to infinity will do the trick. We denote by $\mathrm{D}_{\mathcal{I}}$ a distance matrix D with values at each $(i,j) \in \mathcal{I}$ replaced with infinity. Then, $\textsc{empmmsd}(\mathrm{K}_{xz}, \mathrm{K}_y[\pi, \pi], \mathrm{D}_{\mathcal{I}}, b)$ is a b sample approximation to the distribution of test statistics under the null hypothesis. We additionally correct the biases by ensuring the average of the distribution is 0. Consequently, the approximate null distribution is given by $\mathbf{N} := \frac{1}{2}(\mathbf{N}' - \mu(\mathbf{N}'))$ where $\mathbf{N}' := \left\{\textsc{mmsd}(\Omega'_\pi, \mathrm{D}_{\mathcal{I}}) \mid \Omega'_\pi \in \boldsymbol{\Omega'_\pi}\right\}$.

4.4 ALGORITHM

We combine the new test statistic MMSD and its empirical null distribution to compute a p-value for testing

$X \perp\!\!\!\perp Y \mid Z$ given a sample Ω in the form of Gram matrices $\mathrm{K}_{xz} = \mathrm{K}_x \odot \mathrm{K}_z$ and K_y (see Algorithm 4). We introduce a simple heuristic to improve the quality of approximate null distribution. We say a permutation π for Ω is of good quality if Ω_π closely approximates to the factorized distribution $\mathrm{P}_{xz}\mathrm{P}_{y|z}$. Consider a permutation π' to be learned from $\Omega'_\pi = (\mathbf{x}', \mathbf{y}', \mathbf{z}')$, a half-sample of Ω_π, with the penalized distance matrix $\mathrm{D}_\jmath$ (as shown in the previous section). Since the size of sample is half the size of the original sample, $\mathbf{z}'$ is less dense than $\mathbf{z}$, and the quality of π' becomes worse than that of π. Furthermore, since penalized distance is used, two *originally* close z values, say z'_i and z'_j in Ω'_π, will not be considered in computing π' if the pair is permuted in Ω_π. For these reasons, permutations to be learned for half-samples will not be of as high quality as that of π, which is learned from a full sample without any restrictions. We can rectify this situation using a simple trick. In line 3, permutation π_2 is learned from Ω with a penalized matrix $\mathrm{D}_\jmath$. Although Ω_{π_2} will not be as good as Ω_π in approximating $\mathrm{P}_{xz}\mathrm{P}_{y|z}$, permutations to be learned for half-samples of Ω_{π_2} will be better than those of Ω_π. We empirically observed that the resulting null distributions are more consistent when we used this trick. In Appendix, a further adjustment to the test statistic and null distribution is provided. The key differences between SDCIT and KCIPT are also summarized in Table 2 in Appendix.

5 EXPERIMENTS

We compare the performance of SDCIT with that of other kernel-based CI tests, including KCIT, CHSIC, and KCIPT, with respect to the following two criteria: (i) Area Under Power Curve (AUPC) which estimates how powerful the test is by measuring the area under the cumulative density function (CDF) of $\mathbf{p}$, a set of p-values produced by running a test on a set of samples; and (ii) Kolmogorov-Smirnov (KS) statistic (a measure of the largest discrepancy between cumulative density functions of the two distributions), applied to $\mathbf{p}$ and a uniform distribution to assess the degree to which the p-values produced by the test given samples under the null hypothesis are distributed uniformly in the interval [0,1]. We also compare the different tests with respect to their type I error rates (given a significance level $\alpha = 0.05$) and their run time. We examine the robustness of the different tests with respect to the choice of the kernel function. We contrast SDCIT with KCIPT when the latter uses a large number of bootstraps.

We implemented SDCIT and KCIPT.[2] Unlike the original implementation of KCIPT which employs a simplex algorithm to learn permutations, all permutations are *approximately* learned using minimum cost perfect matching algorithm (BLOSSOM-V, Kolmogorov, 2009) with heuristics for local improvement where three 2-cycles are transformed to two 3-cycles and a 2-cycle and a 3-cycle are transformed to a 5-cycle. Gaussian RBF kernel $k(v, w) := \exp(-\|v-w\|_2^2/(2\sigma^2))$ is used across all experiments where σ is determined by the median heuristic (Gretton et al., 2005). SDCIT uses empirical null distributions of size 10^3. We used the recommended settings for KCIPT (i.e., $B = 25$ and $b, M = 10^4$) and for CHSIC. For KCIT, we used *both* the original implementation, in which variables are normalized to unit variance and kernel parameters are set empirically (Zhang et al., 2011), and a modified implementation, where the variables are not normalized and the kernel parameters are set by median heuristic. For both implementations, GPR is used to optimize the regularization parameter.

[2]The code is available online at `https://github.com/sanghack81/SDCIT`

5.1 EXPERIMENTAL SETTING

Following previous work (Fukumizu et al., 2008; Zhang et al., 2011; Doran et al., 2014), we conducted experiments on two synthetic datasets, post-nonlinear noise and chaotic time series identical to those used in Doran et al. (2014), where each data has two modes, i.e., $X \perp\!\!\!\perp Y \mid Z$ and $X \not\perp\!\!\!\perp Y \mid Z$.

Post-nonlinear noise data is generated using the model described by Zhang and Hyvärinen (2009); Zhang et al. (2011). X and Y are constructed from functions of the form $G_X(F_X(Z_1) + E)$ and $G_Y(F_Y(Z_1) + E)$, respectively, where $G_\cdot$ and $F_\cdot$ are smooth nonlinear functions, E is a Gaussian error, and Z_1 is a random variable in an m dimension conditioning variable $Z := \{Z_i\}_{i=1}^m$. The conditioning variable Z satisfies that $\forall_{i=2}^m Z_i \perp\!\!\!\perp \{X, Y, Z_1\}$ for some m making only Z_1 relevant to X and Y. Since $X \perp\!\!\!\perp Y \mid Z_1$ by construction, we also generate samples under the alternative, $X \not\perp\!\!\!\perp Y \mid Z$, by adding identical Gaussian noise to both X and Y.

Chaotic time series is based on the coupled Hénon map. Each X_t and Y_t is four-dimensional, $X_t = (X_{i,t})_{i=1}^4$ and $Y_t = (Y_{i,t})_{i=1}^4$ where

$$X_{1,t} := 1.4 - X_{1,t-1}^2 + 0.3X_{2,t-1}$$
$$Y_{1,t} := 1.4 - \gamma X_{1,t-1}Y_{1,t-1} + (1-\gamma)\,Y_{1,t-1}^2 + 0.3Y_{2,t-1}$$

and $X_{2,t}$ and $Y_{2,t}$ inherits $X_{1,t-1}$ and $Y_{1,t-1}$, respectively. The third and fourth dimensions of X_t and Y_t correspond to Gaussian noise $\mathcal{N}(0, 0.5^2)$ to make the data more challenging. The parameter γ controls dependence. Regardless of γ, $X_{t+1} \perp\!\!\!\perp Y_t \mid X_t$ holds and given $\gamma > 0$, $Y_{t+1} \not\perp\!\!\!\perp X_t \mid Y_t$.

Following Doran et al. (2014), a set of 300 samples are

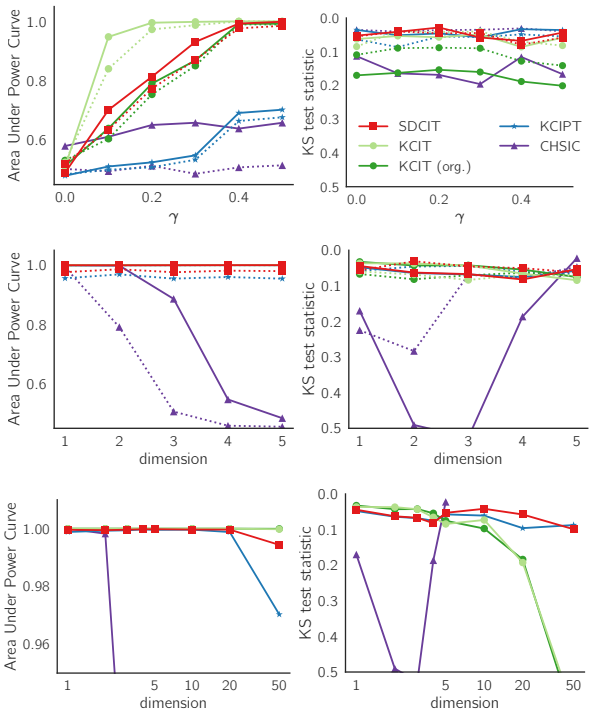

Figure 2: AUPC and KS statistics for (**Top**) Chaotic time series (**Middle**) Post-nonlinear noise data (**Bottom**) High-dimensional conditioning variables on post-nonlinear noise dataset. Dotted and solid lines are for the samples of size 200 and 400, respectively. For high-dimensional setting, CHSIC is excluded.

generated for post-nonlinear noise data per each condition based on the combination of: two variants ($X \perp\!\!\!\perp Y \mid Z$ and $X \not\perp\!\!\!\perp Y \mid Z$), different dimensions $m \in \{1, 2, 3, 4, 5, 10, 20, 50\}$, and different sizes (200 and 400). Similarly, we generated sets of 300 samples of chaotic time series data for two variants, two sample sizes, and γ ranging from 0 to 0.5 by 0.1. Datasets under the null hypothesis are used to report KS statistic and datasets under the alternative hypothesis are used to report AUPCs. SDCIT and KCIPT used RKHS distance for chaotic time series data and, for post-nonlinear noise data, regression-based distance is used where functions are learned based on GPR with automatic relevance determination.

5.2 COMPARISONS AMONG KERNEL-BASED CI TESTS

Figure 2 illustrates the performance of the four tests on the two datasets under various conditions. Note that the original implementation of KCIT with recommended setting is labeled as KCIT (org.) in the figure. Plots in the left column represent AUPCs and those in the right column show KS statistic which measures the degree to which

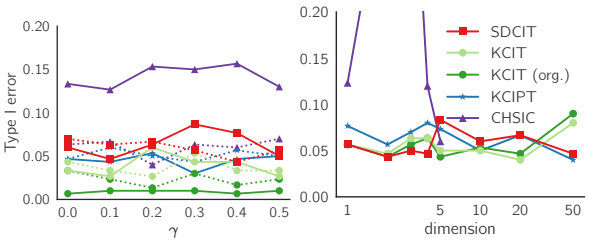

Figure 3: Type I error rates with $\alpha = 0.05$ (**Left**) chaotic time series data (**Right**) post-nonlinear noise data

the distribution of p-values under the null hypothesis deviates from the uniform distribution. Note that the plots for the different tests may differ in the range of values plotted on the y-axis and the values on the y-axis for plots of KS statistic are in descending order.

With respect to power, SDCIT is ranked right below KCIT for both data although SDCIT is more powerful than the implementation of KCIT used in the experiments reported by Doran et al. (2014) for chaotic time series. On post-nonlinear noise data, all tests except CHSIC showed comparable power although the permutation-based methods (SDCIT and KCIPT) show a slight loss in power when $n = 200$. When the number of conditioning variables increases ($m = 50$), SDCIT slightly loses its power (AUPC=0.9944). In the case of KCIPT, we observed a far smaller loss in power (AUPC=0.9703) compared to that reported (around 0.79, Doran et al., 2014). We conclude that SDCIT achieves better power that is comparable to or better than that of all other CI tests except KCIT. We conjecture that the observed difference in power of KCIT relative to SDCIT may be due to the differences in the respective hypotheses, $\mathrm{P}_{xyz} = \mathrm{P}_{x|z}\mathrm{P}_{y|z}\mathrm{P}_z$ versus $\mathrm{P}_{xy|z} = \mathrm{P}_{x|z}\mathrm{P}_{y|z}$.

Next, we compare the tests based on their calibratedness. For both datasets, SDCIT and KCIPT have very consistent null distributions with the KS statistics generally below 0.1. Also the results vary little with sample size. KCIT demonstrates difficulty obtaining accurate null distributions for CI tests with high-dimensional conditional variables. We also report the type I error rates on the two datasets in Figure 3. All tests except CHSIC and the KCIT (org.) reject around 5% of samples under the null hypothesis given a significance level α set to 0.05.

We report the run time of SDCIT and KCIT in Table 1. Based on a set of 300 samples of chaotic time series data, we measured the running time on an iMac with Core i7 3.5Ghz CPU allowing only a single thread. In the case of SDCIT, we expect the learning of permutations will dominate the run time since BLOSSOM-V has $O(nm \log n)$ time complexity where $m = \frac{n(n-1)}{2}$ is the number of edges. However, we see that the run time of

n	SDCIT $b = 500$	SDCIT $b = 1000$	KCIT
200	0.50±0.02	0.96±0.02	1.17±0.19
400	2.41±0.07	4.60±0.07	4.68±0.75

Table 1: Running time in seconds averaged over 300 samples of chaotic time series data (including time for kernel and distance matrices computation)

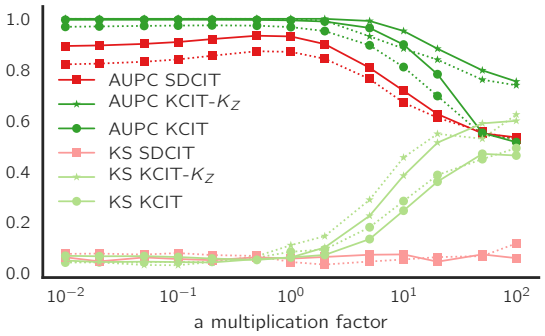

Figure 4: Changes of performance of KCIT and SDCIT with different choice of kernel parameters (relative to the default median heuristic). Solid and dotted lines represent $n = 400$ and 200, respectively.

SDCIT increased by a factor less than 5 when the size of data is doubled. The run time is proportional to b, the size of empirical null distribution. Since SDCIT is trivially parallelizable, the run time can be significantly reduced making the use of modern processors. Note that KCIPT is about 10 times slower even when $B = 25$ due to $B\,(b+1)$ MMD computations.

5.3 ROBUSTNESS WITH RESPECT TO KERNEL CHOICE

The median heuristic to determine the kernel parameter for an RBF kernel works well for both SDCIT and KCIT in previous experiments. However, the kernel choice is entirely at the user's discretion. A well-designed kernel-based test should be robust with respect to the choice of kernel parameters and should have a consistent null distribution. Consider a parametrized RBF kernel, $k^{(C)}(x, x') = \exp\left(-C\|x - x'\|^2/2\sigma^2\right)$, where σ^2 is chosen based on the median heuristic. We explore how KCIT and SDCIT behave as C varies from 10^{-2} to 10^2. We report AUPC and KS statistics based on a set of 300 chaotic time series datasets with 'dependent' and 'independent' modes, respectively, both with $\gamma = 0.3$.

The original implementation of KCIT uses GPR to learn a regularization parameter only with a real vector Z. Hence, we also provide a result where empirical kernel map for Z is approximately inferred from K_Z (labeled as KCIT-K_Z in Figure 4). Our results demonstrate that SDCIT is robust with respect to the choice of kernel without sacrificing well-calibratedness whereas KCIT degrades in calibratedness as C increases.

5.4 COMPARISON OF SDCIT AND KCIPT WITH INCREASING NUMBER OF BOOTSTRAPS

Unlike other tests, the power of KCIPT can be adjusted by controlling B. Hence, we compare the performance of KCIPT with that of SDCIT as we increase the number of bootstraps used by KCIPT. For computational reasons, we replace the bootstrap procedure for the null distribution of KCIPT by an analytic one to cope with a large number of bootstraps (e.g., $B \geq 1000$). Since the inner null distributions are very similar to each other, we aggregated inner null distributions $\{\mathbf{N}_i\}_{i=1}^B$ where each distribution consists of a relatively small number of observations (100). We approximate the null distribution by $\mathcal{N}(0, \mathrm{Var}(\bigcup_{i=1}^B \mathbf{N}_i)/B)$, which is nearly identical to the bootstrap-based null distribution (line 10 in Alg. 1).

We used 300 samples of chaotic time series data of $n = 400$ and $\gamma = 0$. We inferred the appropriate value of B to be 2145 for KCIPT to obtain power comparable to that of SDCIT by matching the variance of null distribution of KCIPT to that of SDCIT based on the first of the 300 samples. The rationale behind the matching is that 1) given a sample, the test statistic of SDCIT and expected test statistic of KCIPT (i.e., $B = \infty$) are close to each other since both are based on the RKHS distance to measure the effect of applying permutation(s), and 2) similar null distributions mean similar critical values for rejecting the null hypothesis. Note, however, that the appropriate B may vary by a given sample.

With the first sample, we compared empirical null distributions of SDCIT and KCIPT together with a test statistic for SDCIT and a distribution of test statistics for KCIPT through repeated trials with varying random seeds on the sample (recall p-value inconsistency of KCIPT) (see top of Figure 5). The null distribution of SDCIT are positively-skewed (0.80), which is similar to the null distribution of MMD_u^2 in a two-sample test setting (Gretton et al., 2012) whereas KCIPT demonstrates normality because of large B. Since KCIPT test statistic is stochastic, the p-value is also stochastic. In this example, p-value is nearly uniformly distributed between 0 and 1. For other samples, we observed distributions of either negative or positive skew, shaped like staircases.

In bottom of Figure 5, we plot test statistics of KCIPT and SDCIT on 300 samples. We observe that the p-values of SDCIT are fairly uniformly distributed, ensuring that it

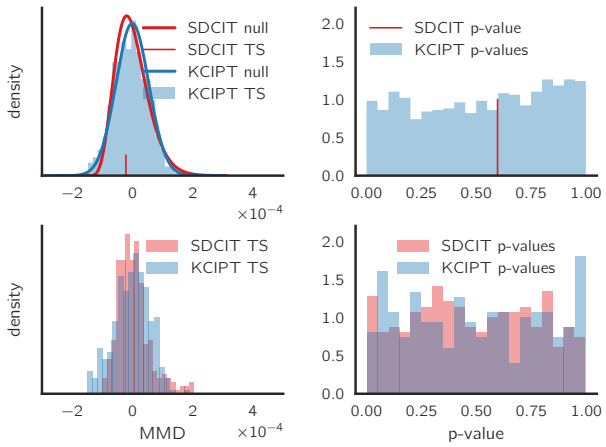

Figure 5: Empirical distributions and p-value distributions of SDCIT and KCIPT with $B = 2145$ based on the first sample (**Top**) and all 300 samples (**Bottom**) with $\gamma = 0.0$ and $n = 400$.

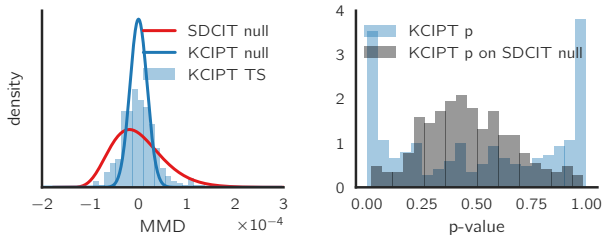

Figure 6: Null distributions of SDCIT and KCIPT, and an empirical distribution of test statistics of KCIPT with $B = 20000$ under the null hypothesis.

is well-calibrated. However, KCIPT starts to lose its calibratedness as the number of bootstraps is increased, resulting in more p-values concentrated around the two extremes, 0 and 1. The analysis presented in Section 3.1 offers an explanation of the loss of calibratedness of KCIPT that is different from that conjectured by Doran et al. (2014), namely, an approximation error in representing $\mathrm{P}_{xz}\mathrm{P}_{y|z}$ due to imperfect permutations.

In Figure 6, with the 300 samples, we illustrate the null distribution of KCIPT and an empirical distribution of its statistics with $B = 20000$. Given a very large B, each test statistic quite accurately reflects conditional (in)dependence of the given sample. That is, the empirical distribution will not shrink as B increases unlike a null distribution will. On the right side, we plot two p-value distributions of KCIPT against null distributions from KCIPT and SDCIT. The dome-shaped p-value distribution is similar to what we have observed, in another simulation, with SDCIT using a bagged statistic of MMSD (i.e., the average of empirical distribution of MMSD) as a test statistic without modifying the null distribution.

Our results suggest that SDCIT is not simply equivalent to KCIPT with some unknown but sufficiently large number of bootstraps B with respect to its power and calibratedness. Further, the run time of SDCIT is generally smaller than that of KCIPT because KCIPT requires a large, data-dependent number (say greater than 1000) to achieve power that is comparable to that of SDCIT even if we factor in the savings achieved in run time by replacing the empirical null distribution by an analytic approximation.

6 SUMMARY AND DISCUSSION

Doran et al. (2014) introduced KCIPT, which has the advantage of relying on a single learned permutation to reduce the conditional independence (CI) test to an easier two-sample test. Based on their experimental comparison of KCIPT with other kernel-based CI tests (CHSIC and KCIT), they concluded that KCIPT has power competitive with that of KCIT and is well-calibrated compared to other alternatives. We observed the previously unobserved behavior of KCIPT when the number of bootstraps is increased: KCIPT suffers from a loss of calibratedness as its power increases with increase in the number of bootstraps. Careful analysis of the behavior of KCIPT leads us to propose SDCIT, a novel kernel-based CI test with a new test statistic based on modified MMD estimate, called Maximum Mean Self-Discrepancy. The resulting test statistic is parameter-free for a given choice of kernel and distortion measure whereas KCIPT requires the user to specify the number of bootstraps. The results of our experiments with two benchmark datasets demonstrate several advantages of SDCIT over other existing kernel-based CI tests. The main drawback of SDCIT is the absence of analytic method to approximate the null distribution. In addition, when there are many permutations minimizing given distortion measure, our test statistic can become non-deterministic. In particular, SDCIT when used to test unconditional independence might yield inconsistent p-values under the null hypothesis despite maintaining its calibratedness.

Acknowledgements

The authors are grateful to UAI 2017 anonymous reviewers for their thorough reviews. This research was supported by the Edward Frymoyer Endowed Professorship, the Center for Big Data Analytics and Discovery Informatics at the Pennsylvania State University, and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science at the Indian Institute of Science.

References

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

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, Boston, MA.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 16(2):323–351.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A Permutation-Based Kernel Conditional Independence Test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 132–141, Corvallis, Oregon. AUAI Press.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Friedman, J. H. and Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel Measures of Conditional Dependence. In *NIPS 2007*, pages 489–496.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773.

Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two sample problem. In *NIPS 2006*, pages 513–520. MIT Press.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Algorithmic Learning Theory. ALT 2005*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A Fast, Consistent Kernel Two-Sample Test. In *NIPS 2009*, pages 673–681.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, volume 2009. MIT Press.

Kolmogorov, V. (2009). Blossom V: a new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation*, 1(1):43–67.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, second edition.

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

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813, Corvallis, Oregon. AUAI Press.

A ALGORITHMIC COMPARISON

	SDCIT	KCIPT
H_0	$\Omega = \Omega_\pi$	$\{\Omega_i^{(1)} = \Omega_{\pi,i}^{(2)}\}_{i=1}^B$
T	modified MMD	averaged MMDs
null distribution	half-sampling without replacement	aggregated bootstrap null distributions
# of permutations	1 for T $b+1$ for null.	B for T
# of MM(S)Ds	1 for T $b+1$ for null.	B for T Bb for null.

Table 2: Comparison of SDCIT and KCIPT

B TAKING PERMUTATION ERROR INTO ACCOUNT

Our test statistic measures the distance between the original sample (representing P_{xyz}) and a pseudo-null sample (representing $\mathrm{P}_{xz}\mathrm{P}'_{y|z}$), where $\mathrm{P}'_{y|z}$ approximates $\mathrm{P}_{y|z}$. Ideally, the test statistic and its null distribution will be reliably estimated if permutation error is small and, hence, $\mathrm{P}'_{y|z}$ approximates $\mathrm{P}_{y|z}$ well.

We first relate an MMSD estimate and its corresponding permutation error during the estimate, and provide a means to adjust MMSD estimates. Let T be an MMSD estimate given K_{xz}, K_y, and D (see Algorithm 2). Let τ be an MMSD estimate assuming $\mathrm{K}_x = \mathbf{1}_{n\times n}$, that is $\tau = \mathrm{MMSD}(\mathrm{K}_z, \mathrm{K}_y, \mathrm{D})$. In other words, τ is the MMSD estimate between $\mathrm{P}'_{y|z}\mathrm{P}_z$ and $\mathrm{P}_{y|z}\mathrm{P}_z$. While T is, roughly, about the conditional dependence between X and Y given Z, τ measures permutation error, i.e., discrepancy between $(\mathbf{y}, \mathbf{z})$ and $(\pi\mathbf{y}, \mathbf{z})$. We illustrate a null distribution $\{T_i\}_{i=1}^b$ and its associated $\{\tau_i\}_{i=1}^b$ in Figure 7. We can clearly observe that the distribution of τ is centered at 0 but still there are lots of null samples associating non-negligible errors.

We then formulate T (under a permutation error) is the function of unknown T^* (under zero permutation error) and τ. We assume a linear model $T = T^* + \beta\tau + \epsilon$ where ϵ is *assumed* a zero-mean Gaussian noise. Given a null distribution $\{(T_i, \tau_i)\}_{i=1}^b$, we can learn β by fitting a linear model. Then, the null distribution $\{T_i\}_{i=1}^b$ is adjusted to $\{T_i - \beta\tau_i\}_{i=1}^b$ and our test statistic is also adjusted similarly. Such adjustment yields a null distribution with smaller variance as shown in Figure 8. The adjustment slightly improves both power and calibratedness.

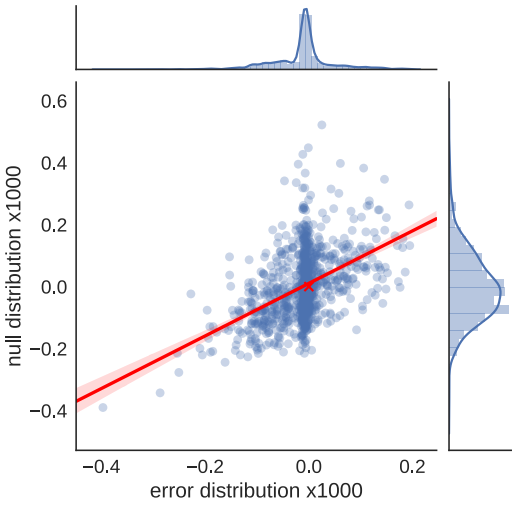

Figure 7: A null distribution and its corresponding errors measured with MMSD. A red cross near origin indicates the test statistic and its corresponding error. A red line indicates a fitted linear model.

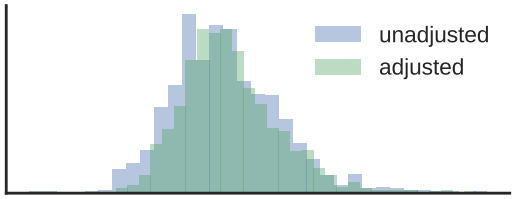

Figure 8: Unadjusted and adjusted null distributions.