A Permutation-Based Kernel Conditional Independence Test

Gary Doran[*‡] **Krikamol Muandet**[‡] **Kun Zhang**[‡] **Bernhard Schölkopf**[‡]

[*]Case Western Reserve University
Cleveland, Ohio 44106, USA
gary.doran@case.edu

[‡]Max Planck Institute for Intelligent Systems
72076 Tübingen, Germany
{krikamol,kzhang,bs}@tuebingen.mpg.de

Abstract

Determining conditional independence (CI) relationships between random variables is a challenging but important task for problems such as Bayesian network learning and causal discovery. We propose a new kernel CI test that uses a single, learned permutation to convert the CI test problem into an easier two-sample test problem. The learned permutation leaves the joint distribution unchanged if and only if the null hypothesis of CI holds. Then, a kernel two-sample test, which has been studied extensively in prior work, can be applied to a permuted and an unpermuted sample to test for CI. We demonstrate that the test (1) easily allows the incorporation of prior knowledge during the permutation step, (2) has power competitive with state-of-the-art kernel CI tests, and (3) accurately estimates the null distribution of the test statistic, even as the dimensionality of the conditioning variable grows.

1 INTRODUCTION

A distribution P_{xyz} over variables X, Y, and Z satisfies a *conditional independence* relationship $X \perp\!\!\!\perp Y \mid Z$ ("X is conditionally independent of Y given Z") when the joint distribution factorizes as $\mathrm{P}_{xyz} = \mathrm{P}_{x|z}\,\mathrm{P}_{y|z}\,\mathrm{P}_z$, assuming the existence of conditional density functions. There are several other equivalent characterizations of conditional independence (Dawid, 1979). Determining whether such conditional independence relationships hold between variables is important for problems such as Bayesian network learning, causal discovery, and counterfactual analysis. Using a conditional independence test as a subroutine, the PC algorithm (Spirtes, Glymour, and Scheines, 2000), for example, can be used to determine a set of causal graphs based on the conditional independence relationships between variables. Moreover, counterfactual analysis often requires assumptions of ignorability, which involve conditional independences among counterfactual variables (Rosenbaum and Rubin, 1983).

Numerous approaches exist to measure conditional dependence or test for conditional independence. For example, under the assumption of Gaussian variables with linear dependence relationships, *partial correlation* can be used to test for conditional independence (Baba, Shibata, and Sibuya, 2004). Another characterization of conditional independence is that $\mathrm{P}_{x|yz} = \mathrm{P}_{x|z}$. Some tests use this characterization to determine conditional independence by measuring the distance between estimates of these conditional densities (Su and White, 2008). When the conditioning variable is discrete, $X \perp\!\!\!\perp Y \mid Z$ if and only if $X \perp\!\!\!\perp Y \mid Z = z_i$ for every possible value z_i that Z takes. Permutation-based tests have been successfully applied to conditional independence testing in this discrete-variable case (Tsamardinos and Borboudakis, 2010). Other tests use this characterization by *discretizing* continuous conditioning variables and testing for independence within each discrete "bin" of Z (Margaritis, 2005).

Generally, conditional independence testing is a challenging problem (Bergsma, 2004). The "curse of dimensionality" in terms of the dimensionality of the conditioning variable Z can make the problem even more difficult to solve. To see why, first consider the case when Z takes a finite number of values $\{z_1, \ldots, z_k\}$; then $X \perp\!\!\!\perp Y \mid Z$ if and only if $X \perp\!\!\!\perp Y \mid Z = z_i$ for each value z_i. Given a sample of size n, even if the data points are evenly distributed across values of Z, we must show independence within every subset of the sample with identical Z values using only approximately n/k points within each subset. When Z is real-valued and P_z is continuous, the observed values of Z are almost surely unique. To extend the procedure to the continuous case, we must infer conditional independence using nonidentical but nearby values of Z, where "nearby" must be quantified with some distance metric. Finding nearby points becomes difficult (without additional assumptions) as the dimensionality of Z grows. To guarantee that conditional independence reduces to unconditional independence between X and Y within each subset,

we need a large number of subsets of Z. On the other hand, with many subsets, in each subset one may not have enough points to assess independence.

Recently, kernel-based tests have also been proposed for conditional as well as unconditional independence testing (see Section 3 for a more detailed discussion). Kernel functions can be used to implicitly map objects from an input space into a "feature space," or reproducing kernel Hilbert space (RKHS) (Aizerman, Braverman, and Rozoner, 1964; Schölkopf and Smola, 2002). Some tests use the kernel mean embedding, which is an embedding of distributions into an RKHS (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Sriperumbudur et al., 2010). When the kernel used is *characteristic*, the embeddings of two distributions are equal (under the distance metric imposed by the RKHS norm) if and only if the distributions are identical. For example, all *universal* kernels such as the radial basis function (RBF) kernel are characteristic (Sriperumbudur et al., 2010). The Hilbert–Schmidt independence criterion (HSIC) is an unconditional independence test that measures the distance in the RKHS between the embedding of a joint distribution and the embedding of the product of its marginal distributions. The HSIC can also be interpreted as the Hilbert–Schmidt norm of a cross-covariance operator, a generalization of the covariance matrix, between the RKHSs corresponding to the marginal distributions (Gretton et al., 2008). The intuition behind the test is that a joint distribution P_{xy} is equal to the product of its marginals if and only if $X \perp\!\!\!\perp Y$. The HSIC has been extended to the conditional independence setting using the norm of the conditional cross-covariance operator to measure conditional dependence (Fukumizu et al., 2008). However, this approach also degrades as the dimensionality of the conditioning variable increases. A more recent approach, the kernel conditional independence test (KCIT), proposed by Zhang et al. (2011), uses a characterization of conditional independence defined in terms of the partial association under all square-integrable functions relating the variables X, Y, and Z (Daudin, 1980). The test relaxes this characterization to use a smaller, but sufficiently rich class of functions from some universal RKHS. For this test, the distribution of the test statistic is known and can be estimated efficiently. However, as the dimensionality of the conditioning variable grows larger or the relationships between the variables grow more complex, the distribution of the KCIT test statistic under the null distribution becomes harder to accurately estimate in practice.

In contrast to a conditional independence test, a kernel two-sample test (Gretton et al., 2006, 2009, 2012a) merely tests whether two samples have been drawn from the same distribution. The two-sample problem is conceptually simpler than testing for conditional independence, and has been studied extensively in prior work. Thus, the behavior of the null distributions for two-sample test statistics are well-understood.

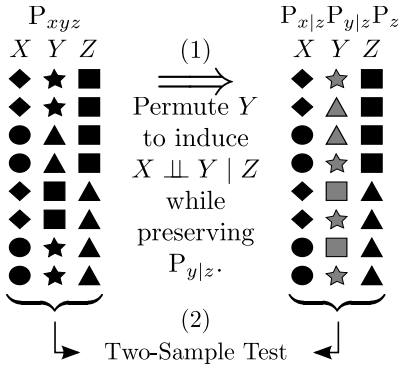

Figure 1: An overview of the proposed approach. First, we observe a sample from the joint distribution (left), and permute the sample to simulate a sample from the factorized distribution (right). The permutation is chosen to induce conditional independence while preserving $\mathrm{P}_{x|z}$, $\mathrm{P}_{y|z}$, and P_z. Then, a two-sample test is used to compare the permuted sample to an independent, unpermuted sample from the joint distribution.

We propose a new approach to test for conditional independence that uses a permutation to reduce the problem to a two-sample test. An overview of the approach is illustrated in Figure 1. First, a single, carefully chosen permutation is applied to a sample to simulate a sample from the factorized distribution $\mathrm{P}_{x|z}\,\mathrm{P}_{y|z}\,\mathrm{P}_z$, which equals the underlying joint distribution if and only if the null hypothesis holds. Then, a kernel two-sample test (Gretton et al., 2012a) is performed between the permuted sample and an independent, unpermuted sample from the original distribution to determine whether the null hypothesis of conditional independence should be rejected. The approach permits various strategies for "learning" an appropriate permutation given prior knowledge about relationships between X, Y, and Z. The p-values for our test can be accurately, efficiently approximated using the approaches studied previously for kernel two-sample tests. We show using synthetic datasets that the proposed test has power competitive with state-of-the-art approaches, and can accurately estimate the distribution of the test statistic under the null hypothesis as the dimensionality of Z grows to produce a well-calibrated test. We also illustrate using a real-world dataset the practicability of the test for inferring conditional independence relationships.

2 DESCRIPTION OF THE TEST

In testing for *unconditional* independence, we observe an independent and identically distributed (i.i.d.) sample $\Omega = \{(x_i, y_i)\}_{i=1}^n$ drawn from P_{xy}. The variables X and Y are independent if and only if the joint distribution factorizes as

$\mathrm{P}_{xy} = \mathrm{P}_x \mathrm{P}_y$. Here, P denotes a density function, but we use the same notation to represent the distribution itself. If we managed to draw a sample Ω' from $\mathrm{P}_x \mathrm{P}_y$, we could use a two-sample test between Ω and Ω' to determine whether to reject the null hypothesis $\mathcal{H}_0 : X \perp\!\!\!\perp Y$. Since we do not have access to the underlying joint distribution, but only a sample Ω, we must "simulate" a sample from the factorized distribution. By the i.i.d. assumption, the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a product of identical factors $\mathrm{P}_{xy}(X_i, Y_i)$. Hence for all i, X_i and Y_i have the same marginals P_x and P_y, respectively. Moreover, for $i \neq j$, we have $X_i \perp\!\!\!\perp Y_j$. If π is a permutation satisfying $\pi(i) \neq i$, we thus have $X_i \perp\!\!\!\perp Y_{\pi(i)}$, and the distribution of $(X_i, Y_{\pi(i)})$ must be $\mathrm{P}_x(X_i) \mathrm{P}_y(Y_{\pi(i)})$. Therefore, the permuted sample $(x_i, y_{\pi(i)})_{i=1}^n$ approximately simulates an i.i.d. sample from $\mathrm{P}_x \mathrm{P}_y$.[1]

Below, we first discuss a way to extend the use of permutations to the conditional independence setting. Then, we show how to apply a kernel two-sample test to a permuted and an unpermuted sample to test for conditional independence. We describe how bootstrapping can be used to improve the power of the test. Given the two-sample test, we describe a kernel-based approach for learning an appropriate permutation.

2.1 PERMUTING FOR CONDITIONAL INDEPENDENCE

In this paper, for a joint distribution P_{xyz} over the variables X, Y, and Z, we are interested in determining whether $X \perp\!\!\!\perp Y \mid Z$, which occurs if and only if $\mathrm{P}_{xyz} = \mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$. We observe an i.i.d. sample $\Omega = \{(x_i, y_i, z_i)\}_{i=1}^n$ drawn according to P_{xyz}. As above, if we were able to draw an independent sample Ω' from the factorized distribution $\mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$, we could use a two-sample test between Ω and Ω' to determine whether to reject the null hypothesis $\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid Z$. Given the distributions $\mathrm{P}_{x|z}$, $\mathrm{P}_{y|z}$, and P_z, we could sample from the factorized distribution by first drawing $z_i \sim \mathrm{P}_z$, and then $x_i \sim \mathrm{P}_{x|z_i}$ and $y_i \sim \mathrm{P}_{y|z_i}$. However, since we are given only a sample, we must "simulate" a sample from $\mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$. If the null hypothesis holds, one can consider each x_i and y_i in Ω as independently drawn from the conditional distributions $\mathrm{P}_{x|z_i}$ and $\mathrm{P}_{y|z_i}$, respectively. Suppose for some $i \neq j$, we find that $z_i = z_j$. In that case, we can proceed analogously to the unconditional case: we can swap the corresponding values y_i and y_j, breaking the dependence between X and Y, to obtain a joint observations (x_i, y_j, z_i) (and, likewise, (x_j, y_i, z_j)) drawn from the distribution $\mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$. Therefore, if we were able to (nontrivially) permute every y_i value so that the same permutation leaves the values of z_i in the sample invariant, then this would simulate i.i.d. draws from $\mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$. Unfortunately, when Z is continuous, the observed values of Z in Ω will be almost surely unique. In this case, we must "approximately" simulate a sample from $\mathrm{P}_{x|z} \mathrm{P}_{y|z} \mathrm{P}_z$.

The procedure described above can be formalized as follows. Let the sample be expressed as $\Omega = (\mathbf{x}, \mathbf{y}, \mathbf{z})$, where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ denote tuples of length n holding the sample elements for each of the variables (which might be multivariate), with ranges $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$. For the moment, we assume that $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are equipped with addition and scalar multiplication. When we introduce the kernelization of the sample in Section 2.2, this assumption holds even when the sample elements are nonvectorial structured objects. Let $\mathbf{P}$ be a linear transformation, represented as a matrix with nonnegative entries, that is defined to act on a sample as in: $\mathbf{P}\Omega \triangleq (\mathbf{x}, \mathbf{P}\mathbf{y}, \mathbf{z})$, where $\mathbf{P}\mathbf{y}$ is a tuple whose i^{th} element contains $\sum_j \mathbf{P}_{ij} y_j$. To preserve statistical properties of the sample, we cannot use a general linear transformation $\mathbf{P}$; it must be a permutation matrix:

Proposition 1. *Let $\mathcal{T}$ be the set of transformation such that for any $\mathbf{P} \in \mathcal{T}$ and sample $\mathbf{y}$ of size n, mean$(\mathbf{P}\mathbf{y}) =$ mean$(\mathbf{y})$ and $\|var(\mathbf{P}\mathbf{y})\|_{\mathrm{HS}} = \|var(\mathbf{y})\|_{\mathrm{HS}}$. Then $\mathcal{T}$ is a set of permutation matrices of size n.*[2]

Essentially, the matrix $\mathbf{P}$ must be stochastic to preserve the mean and orthogonal to preserve the variance, and these properties combined imply that it is a permutation. Given that $\mathbf{P}$ is a permutation, we additionally require that $\mathrm{Tr}(\mathbf{P}) = 0$, so that no element in the sample $\mathbf{y}$ is permuted with itself (i.e., left unchanged). Otherwise, some dependence between x_i and y_i might remain. We use $\mathcal{P}$ to denote the set of zero-trace permutations.

Ideally, we would further constrain $\mathcal{P}$ so that all $\mathbf{P} \in \mathcal{P}$ satisfy $\mathbf{z} = \mathbf{P}\mathbf{z}$; that is, the values of $\mathbf{z}$ are invariant under each permutation $\mathbf{P}$, or equivalently, we only permute the values of y_i that correspond to the same value of z_i. In the unconditional case, we can consider Z to be some constant variable, in which case any permutation is permitted. However, in the conditional case, this requirement is too restrictive so that the set of valid permutations is empty because each value of Z appears only once almost surely in the sample with continuous Z. Accordingly, we relax the problem to finding a permutation that enforces $\mathbf{z} \approx \mathbf{P}\mathbf{z}$. In particular, given some *distortion measure* $\delta : \mathcal{Z}^n \times \mathcal{Z}^n \to [0, \infty)$ that quantifies the discrepancy between permuted and unpermuted values of Z, we seek to optimize $\min_{\mathbf{P} \in \mathcal{P}} \delta(\mathbf{z}, \mathbf{P}\mathbf{z})$. For general classes of distortion measures, this optimiza-

[1] In the finite sample setting, this is only an approximation: while the permutation removes the dependence between X and its corresponding Y, it introduces a dependence to one of the other Y variables. In the limit $n \to \infty$, this becomes negligible (Janzing et al., 2013); moreover, in the limit we could waive the constraint $\pi(i) \neq i$ which we do not do in the present work since it is easy to implement.

[2] A proof can be found in the supplementary materials, available at `http://engr.case.edu/doran_gary/publications.html`.

tion problem is straightforward to solve. For example, let $d(z_i, z_j)$ be any symmetric pairwise distortion measure (e.g., a distance metric) and $\delta(\mathbf{z}, \mathbf{Pz}) = \sum_i d(\mathbf{z}_i, (\mathbf{Pz})_i)$. Let $\mathbf{D}$ be a matrix of pairwise distances between sample elements ($\mathbf{D}_{ij} = d(z_i, z_j)$). Since $\mathbf{P}_{ij} = 1$ if and only if z_i is permuted to z_j, $\delta(\mathbf{z}, \mathbf{Pz}) = \sum_{ij} \mathbf{P}_{ij}\mathbf{D}_{ij} = \mathrm{Tr}(\mathbf{PD})$ and the distortion measure can be minimized using:

$$\min_{\mathbf{P} \in \mathcal{P}} \mathrm{Tr}(\mathbf{PD}). \tag{1}$$

Relaxing $\mathcal{P}$ to the set of doubly stochastic matrices (matrices whose rows and columns sum to one) with zero trace, the feasible region becomes the convex hull of permutation matrices, by the Birkhoff–von Neumann theorem (Birkhoff, 1946; von Neumann, 1953), subject to a linear constraint. Therefore, the simplex algorithm applied to Equation 1 returns a solution corresponding to a vertex of the feasible region, which is a permutation. The formulation in Equation 1 gives a general approach to permuting a sample, where the choice of distance metric d can encode some assumptions about the properties of the distributions $\mathrm{P}_{x|z}$ or $\mathrm{P}_{y|z}$. Below we discuss using $\mathbf{P}$ to construct the test statistic and possible choices for d.

2.2 TEST STATISTIC AND NULL DISTRIBUTION

After learning an appropriate permutation, a two-sample test between a permuted and an unpermuted sample can be used to test the null hypothesis of conditional independence. A well-studied kernel-based two-sample test uses the maximum mean discrepancy (MMD) test statistic (Gretton et al., 2012a). The MMD employs *mean embeddings* of the two samples into some RKHS. Before describing the mean embedding, we introduce the notation used to "kernelize" the sample. Given the ranges $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ of the random variables X, Y, and Z, let $k_x(\cdot,\cdot)$, $k_y(\cdot,\cdot)$, and $k_z(\cdot,\cdot)$ be positive-definite *kernel* functions defined on these spaces ($k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, etc.). Corresponding to each kernel k_x is some *feature map* $\phi_x : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ such that $k_x(x, x') = \langle \phi_x(x), \phi_x(x') \rangle$, where $\mathcal{H}_\mathcal{X}$ is the RKHS or *feature space* of k_x. We use a product of individual kernels to define the kernel on joint spaces; e.g., $k_{xyz}\big((x,y,z),(x',y',z')\big) = k_x(x,x')k_y(y,y')k_z(z,z')$ is a kernel over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with feature map $\phi_x \otimes \phi_y \otimes \phi_z$, where $\otimes$ denotes the tensor product. Given that k_x, k_y, and k_z are translation-invariant characteristic kernels, the product kernel is also characteristic under mild assumptions (Sriperumbudur et al., 2010).

By mapping our sample into a feature space as $\{(\phi_x(x_i), \phi_y(y_i), \phi_z(z_i)\}_{i=1}^n$, we can treat each sample element as a vector (which is infinite-dimensional for the characteristic kernels used below), even when the underlying distribution over X, Y, and Z is over arbitrary sets of objects on which kernels are defined. In matrix notation, we can express the mapped sample as $(\Phi_x(\mathbf{x}), \Phi_y(\mathbf{y}), \Phi_z(\mathbf{z}))$, and the permuted sample in the feature space as $\mathbf{P}\Omega = (\Phi_x(\mathbf{x}), \mathbf{P}\Phi_y(\mathbf{y}), \Phi_z(\mathbf{z}))$ with the i^{th} element of $\mathbf{P}\Phi_y(\mathbf{y})$ equal to $\sum_j \mathbf{P}_{ij}\phi_y(y_j)$. The conditions on $\mathbf{P}$ in Proposition 1 are still necessary, since the linear kernel with feature map $\phi_y : \mathcal{Y} \to \mathcal{Y}$ is a special case to which Proposition 1 applies. In prior work (Sriperumbudur et al., 2010), $mean(\Phi_y(\mathbf{y}))$ is called the *empirical kernel mean embedding*, and is expressed with the notation $\widehat{\mu}(\mathbf{y}) = \frac{1}{n}\sum_{y \in \mathbf{y}} \phi_y(y)$. Given the constraints on $\mathbf{P}$, $\Phi_y(\mathbf{Py}) = \mathbf{P}\Phi_y(\mathbf{y})$ for any Φ_y, and the mean embedding is invariant under $\mathbf{P}$: $\widehat{\mu}(\mathbf{y}) = \widehat{\mu}(\mathbf{Py})$. The notation $\widehat{\mu}(\Omega)$ will be used to denote the mean embedding of an entire sample using the product kernel defined on the joint space.

Given the kernelization of the sample, the test statistic is computed as follows. The original sample Ω of n elements is randomly split in half to form the samples $\Omega^{(1)}$ and $\Omega^{(2)}$, to ensure independence between the permuted and unpermuted samples (a condition required by the two-sample test). Using the formulation in Equation 1, we learn a permutation that induces conditional independence in the second subsample $\Omega^{(2)}$. Finally, we compute the (biased) MMD test statistic as follows:

$$\begin{aligned}\mathrm{MMD}(\Omega^{(1)}, \mathbf{P}\Omega^{(2)}) &= \left\| \widehat{\mu}(\Omega^{(1)}) - \widehat{\mu}(\mathbf{P}\Omega^{(2)}) \right\|_\mathcal{H}^2 \\ &= \tfrac{4}{n^2}\mathbf{1}^\intercal(\mathbf{K}^{(1)} + \mathbf{K}^{(2)} - 2\mathbf{K}^{(12)})\mathbf{1}.\end{aligned} \tag{2}$$

Here, $\mathbf{1}$ is a vector of ones of an appropriate size, the matrices $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$ are pairwise kernel matrices within the permuted and unpermuted samples, respectively, and $\mathbf{K}^{(12)}$ is the "cross" kernel matrix between the unpermuted and permuted samples. Since we use product kernels, the matrices can be expressed in terms of a Hadamard product between the original kernel matrices for each variable:

$$\begin{aligned}\mathbf{K}^{(1)} &= \mathbf{K}_x^{(1)} \odot \mathbf{K}_y^{(1)} \odot \mathbf{K}_z^{(1)} \\ \mathbf{K}^{(2)} &= \mathbf{K}_x^{(2)} \odot (\mathbf{P}\mathbf{K}_y^{(2)}\mathbf{P}^\intercal) \odot \mathbf{K}_z^{(2)} \\ \mathbf{K}^{(12)} &= \mathbf{K}_x^{(12)} \odot (\mathbf{K}_y^{(12)}\mathbf{P}^\intercal) \odot \mathbf{K}_z^{(12)},\end{aligned}$$

where $(\mathbf{K}_x)_{ij} = k_x(x_i, x_j)$, and likewise for $\mathbf{K}_y$ and $\mathbf{K}_z$.

The behavior of the MMD test statistic has been extensively studied in prior work (Gretton et al., 2006, 2009, 2012a), and there are numerous approaches to estimating the null distribution and computing a p-value for the test statistic. For example, the null distribution can be estimated via a bootstrapping approach in which (1) $\Omega^{(1)}$ and $\mathbf{P}\Omega^{(2)}$ are randomly shuffled together and then split into two again, and then (2) the test statistic is recomputed between the shuffled samples (Gretton et al., 2009). Steps (1) and (2) are repeated b times to obtain an empirical estimate of the null distribution. The null distribution can also be approximated using a Gamma distribution. This estimate is computationally more efficient to obtain, but can also be less accurate in some scenarios (Gretton et al., 2009). As we are interested in the small-sample case, we choose to use the more robust bootstrap estimate at the expense of more computation.

Algorithm 1 KCIPT: Kernel Conditional Independence Permutation Test

Require: Sample $\Omega = (\mathbf{x}, \mathbf{y}, \mathbf{z})$, Distortion measure δ, Significant level α, Outer bootstrap iterations B, Inner bootstrap iterations b, Monte Carlo iterations M

for Outer Bootstrap $1 \leq i \leq B$ **do**
  Split sample evenly into $\Omega^{(1)}$, $\Omega^{(2)}$
  Find permutation matrix $\mathbf{P}$ for $\Omega^{(2)}$ using δ to compute $\mathbf{D}$ and solving Equation 1.
  $\mathrm{MMD}[i] \leftarrow \mathrm{MMD}(\Omega^{(1)}, \mathbf{P}\Omega^{(2)})$
  for Inner Bootstrap $1 \leq j \leq b$ **do**
    Shuffle, re-split $\Omega^{(1)}$, $\mathbf{P}\Omega^{(2)}$ to Ω', Ω''.
    $\mathrm{inner\_null}[i, j] \leftarrow \mathrm{MMD}(\Omega', \Omega'')$
  end for
end for
$\mathrm{statistic} \leftarrow mean_{1 \leq i \leq B}(\mathrm{MMD}[i])$
for Monte Carlo Iteration $1 \leq k \leq M$ **do**
  for Outer Bootstrap $1 \leq i \leq B$ **do**
    $r \leftarrow \mathrm{random\_integer}(1, b)$
    $\mathrm{samples}[i] \leftarrow \mathrm{inner\_null}[i, r]$
  end for
  $\mathrm{outer\_null}[k] \leftarrow mean_{1 \leq i \leq B}(\mathrm{samples}[i])$
end for
p-value $\leftarrow 1 - \mathrm{percentile}(\mathrm{statistic}, \mathrm{outer\_null})$
if p-value $\geq \alpha$ **then**
  Fail to Reject $\mathcal{H}_0$ ($X \perp\!\!\!\perp Y \mid Z$)
else
  Reject $\mathcal{H}_0$, Conclude $X \not\perp\!\!\!\perp Y \mid Z$
end if

A characteristic kernel must be used to ensure that the MMD test statistic is consistent (convergent to zero if and only if the two samples are drawn from the same distribution). A kernel k_{xyz} is said to be *characteristic* if the corresponding mean map is injective (Sriperumbudur et al., 2010). Several popular kernels are characteristic, including the Gaussian RBF kernel $k(x, x') = \exp(-\|x - x'\|_2^2 / 2\sigma^2)$, with bandwidth parameter σ. Given that the RBF kernel is used for the test, there is still a question of how to select the bandwidth parameter. We set σ to be the median pairwise distance between sample elements, which prior work shows to be an effective heuristic (Gretton et al., 2012a). Other strategies existing for selecting σ to improve the power of the test statistic (Sriperumbudur et al., 2009; Gretton et al., 2012b).

2.3 BOOTSTRAPPING THE TEST STATISTIC

As defined above, the test statistic relies on splitting a sample randomly in half, which reduces the power of the two-sample test. However, if we randomly split the sample many times to compute many test statistics, we can bootstrap the MMD statistic itself to recover some of the power lost due to splitting. An overview of the test with bootstrapping is given in Algorithm 1.

Let $\{(\Omega_i^{(1)}, \Omega_i^{(2)})\}_{i=1}^B$ be a set of random splits of the dataset, where B denotes the number of random splits. The bootstrapped test statistic is the average of individual MMD test statistics for each split: $\mathrm{MMD}_{\mathrm{boot}}(\Omega) = \frac{1}{B}\sum_{i=1}^B \mathrm{MMD}(\Omega_i^{(1)}, \mathbf{P}_i\Omega_i^{(2)})$, where $\mathbf{P}_i$ is the permutation learned for the i^{th} split. The null distribution of $\mathrm{MMD}_{\mathrm{boot}}$ can be estimated via a Monte Carlo simulation by repeatedly averaging together the draws from each individual test statistic's null distribution. Specifically, the null distribution N_i is first estimated for each test statistic $\mathrm{MMD}_b(\Omega_i^{(1)}, \mathbf{P}_i\Omega_i^{(2)})$. Then, M points are drawn from each of the B null distributions: $s_{ij} \sim N_i$, for $1 \leq i \leq B$, $1 \leq j \leq M$. The points are averaged so that the resulting sample $\{\frac{1}{B}\sum_{i=1}^B s_{ij}\}_{j=1}^M$ is used to estimate the null distribution of $\mathrm{MMD}_{\mathrm{boot}}(\Omega)$.

Since we are combining many tests, any systematic error in estimating the null distribution will be compounded. Accordingly, we choose to use the robust bootstrapping approach described in Section 2.2 with a large number of draws b to estimate each null distribution N_i. Note that the statistic bootstrapping procedure (the "outer" bootstrap) is separate from the bootstrapping used to estimate the null hypothesis (the "inner" bootstrap). In the first case, a new permutation is learned for each split to compute the test statistic. In the second case, the learned permutation for the given split is left fixed, and the permuted and unpermuted subsamples are shuffled together randomly to simulate the null hypothesis. Since each N_i is an empirical estimate using a set of observed test statistics, we draw from N_i by sampling with replacement from the underlying set. If desired, the inner bootstrap shown in Algorithm 1 can be replaced with some other estimate of the null distribution of each MMD test statistic.

2.4 LEARNING THE PERMUTATION

Given the description of our test procedure, we now return to the issue of learning a permutation. Intuitively, since the test statistic uses an RKHS distance between samples, we would like our distortion measure to also utilize the RKHS distance. Therefore, we choose $d(z_i, z_j) = \|\phi_z(z_i) - \phi_z(z_j)\|$. In fact, we show that minimizing Equation 1 with respect to the RKHS distortion measure leads to a consistent test statistic when the distortion converges to zero. That is, we would like for the MMD between permuted and unpermuted samples to converge to zero if and only if the null hypothesis $\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid Z$ holds.

Definition 1. *A test statistic is asymptotically consistent if it converges in probability to zero if and only if the null hypothesis holds.*

Theorem 1. *Let $\mathbf{D}_{ij}^{\mathrm{RKHS}} = \|\phi_z(z_i) - \phi_z(z_j)\|$ be a pairwise RKHS distance matrix between Z values in a sample. The proposed test statistic (Equation 2) is asymptot-*

ically consistent if the quantity $\min_{\mathbf{P}\in\mathcal{P}} \frac{1}{n}\operatorname{Tr}(\mathbf{P}\mathbf{D}^{\mathrm{RKHS}})$ *converges in probability to zero as* $n \to \infty$.

Proof. Intuitively, minimizing $\frac{1}{n}\operatorname{Tr}(\mathbf{P}\mathbf{D}^{\mathrm{RKHS}})$ minimizes a majorant of the MMD between the permuted and unpermtued joint samples $(\mathbf{P}\mathbf{y}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$. When this value converges to zero in probability, then so does the MMD, which implies that the permuted sample embedding converges to the embedding of the factorized joint distribution.[3] □

The optimal choice of the distance metric $d(z_i, z_j)$ should depend on how Z influences X and Y. Consider an extreme case where all dimensions of Z except Z_1 are irrelevant to (independent from) X and Y given Z_1. We aim to find the nearby points along Z_1, which are not necessarily neighbors when all dimensions are included. In other words, we should exclude all those irrelevant dimensions of Z when calculating the distances between z_i. An example is shown in Figure 2, where Y is some linear function of only the first component of Z, plus some Gaussian noise. Sample elements within the "level sets" of the hyperplane (indicated in the figure) are approximately exchangeable.

Generally speaking, given prior knowledge about structure in the relationships between variables, better measures of distance can be employed when learning the permutation. For example, a well-studied assumption in causal discovery is that X and Y are continuous functions of Z plus some independent Gaussian noise (Hoyer et al., 2008). When this is true, $\mathrm{P}_{y|z} = \mathcal{N}(f(z), \Sigma)$, where f is some continuous function relating Z and Y, and Σ is a covariance matrix. In this case, $\mathrm{P}_{y|z_i} \approx \mathrm{P}_{y|z_j}$ if $f(z_i) \approx f(z_j)$, so it makes sense to use the distance metric $d(z_i, z_j) = \|f(z_i) - f(z_j)\|_2$ when learning the permutation. Although f is unknown, it can be learned from the data; e.g., by using Gaussian Process (GP) regression (Rasmussen and Williams, 2006). Of course, the consistency of the test statistic when heuristics are used depends upon whether the assumptions made by the heuristics are satisfied by the underlying joint distribution. In our experimental results, described below, we find that the function-based distance heuristic adds power to the test for synthetic datasets in which X and Y are in fact noisy functions of Z.

3 RELATED WORK

A previous approach, the conditional HSIC (CHSIC), uses the Hilbert–Schmidt norm of the conditional cross-covariance operator, which is a measure of conditional covariance of the images of X and Y under functions f and g from RKHSs corresponding to some kernels defined on X and Y. When the RKHSs correspond to characteristic kernels, the operator norm is zero if and only if $X \perp\!\!\!\perp Y \mid Z$ (Fukumizu et al., 2008). Since it is unknown how to analytically compute the null distribution of the CHSIC, the distribution is estimated using a bootstrapping approach. As described above for the MMD, a null distribution can be estimated by shuffling and recomputing the test statistic numerous times. In the conditional case, X and Y should only be shuffled when the corresponding Z values are near each other. Therefore, the values of Z are partitioned using a clustering algorithm, and bootstrap estimates are obtained by permuting Y values only within clusters (Fukumizu et al., 2008). Compared to our approach, the CHSIC has several disadvantages. The CHSIC requires *many* permutations to estimate the null distribution, whereas our approach only requires *one* carefully chosen permutation (per outer bootstrap iteration). Since the CHSIC clusters Z values to generate permutations, the permuted data points within each cluster have more widely varying values for Z, causing larger approximation errors. Finally, for high-dimensional datasets, finding an appropriate clustering algorithm becomes difficult, and the approximation quickly breaks down.

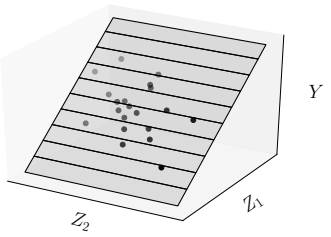

Figure 2: If Y is a function of Z plus noise, then many dimensions of Z might be irrelevant for determining conditional independence. In this example, Y is a noisy function of Z_1, so sample elements within the level sets of the hyperplane are approximately exchangeable.

Other previous approaches to conditional independence testing use the partial association of regression functions relating X, Y, and Z (Huang, 2010; Zhang et al., 2011). In particular, the kernel-based KCIT (Zhang et al., 2011) is based on the following characterization of conditional independence: for any $f \in L^2_{XZ}$, and $g \in L^2_Y$, define $\tilde{f}(X, Z) = f(X, Z) - h_f(Z)$ and $\tilde{g}(Y, Z) = g(Y) - h_g(Z)$, where $h_f, h_g \in L^2_Z$ are regression functions of the values of f and g using only the variable Z. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if for all $f \in L^2_{XZ}$, $g \in L^2_Y$, and $\tilde{f}$, $\tilde{g}$ defined as above, $\mathrm{E}[\tilde{f}\tilde{g}] = 0$ (Daudin, 1980). The KCIT relaxes the spaces of functions L^2_{XZ}, L^2_Y, and L^2_Z to be RKHSs corresponding to kernels defined on these variables. A universal kernel is required so that the RKHS for Z is dense in corresponding L^2_Z space. By contrast, the HSIC and our approach only require characteristic kernels, which need not be universal (Sriperumbudur et al., 2010).

[3] See supplementary materials for the full proof.

4 EMPIRICAL EVALUATION

Our analysis suggests that by using a single permutation to compute the test statistic, our approach, the kernel conditional independence permutation test (KCIPT) will be more powerful than the CHSIC, which requires clustering the values of Z and many permutations in each cluster to estimate the null distribution. These permutations become difficult to find as the dimensionality of Z grows, as shown in prior work (Zhang et al., 2011). Furthermore, by using an MMD-based test statistic, the KCIPT can better estimate the null distribution than the KCIT in scenarios that require a careful choice of parameters. Finally, the outer bootstrapping procedure should improve the power of the KCIPT.

To empirically support our analysis, we implement KCIPT in MATLAB,[4] and compare it to implementations of CHSIC and KCIT used in prior work (Zhang et al., 2011). We use two criteria for performance evaluation, type I error (the fraction of the time the null hypothesis $\mathcal{H}_0$ is incorrectly rejected), and power (the fraction of the time $\mathcal{H}_0$ is correctly rejected). Rather than choosing a specific significance level α at which to evaluate power and type I error, we record the p-values resulting from each test and analyze the behavior of the tests as α varies. For KCIPT, we use an RBF kernel $k(x, x') = \exp(-\|x - x'\|_2^2 / 2\sigma^2)$ for each variable, with bandwidth parameters σ_x, σ_y, and σ_z chosen using the "median" heuristic (Gretton et al., 2012a). For bootstrapping, we use parameters $B = 25$, $b = 10^4$, and $M = 10^4$. CHSIC and KCIT use the recommended parameters set in their implementations.

In order to characterize the power and type I error of the tests, we must evaluate the tests across many samples from the same underlying distribution. We use synthetic datasets from prior work for this purpose (Fukumizu et al., 2008; Zhang et al., 2011). Each dataset has a variant where the null hypothesis holds, for testing type I error, and where the null hypothesis does not hold, for testing power. We perform 300 tests for each condition, for each dataset described below.

Post-nonlinear Noise. The first dataset we use generates X and Y as functions of Z using a post-nonlinear noise model (Zhang and Hyvärinen, 2009; Zhang et al., 2011). In this generative process, the dimensionality of the conditioning variable Z grows, but only the first dimension Z_1 is relevant to the conditional independence of X and Y. Each of X and Y are determined using $G(F(Z_1) + E)$, where G and F are arbitrary smooth, nonlinear functions and E is a Gaussian noise variable. All dimensions of Z are i.i.d. Gaussian random variables. Since $X \perp\!\!\!\perp Y \mid Z$ by default, identical Gaussian noise is added to X and Y to produce a variant of the dataset for which $X \not\perp\!\!\!\perp Y \mid Z$. Because only one conditioning variable is relevant to the problem, we expect that at least the KCIT and KCIPT with the function-distance distortion measure will be robust to increasing dimensionality, but that performance will degrade eventually.

Chaotic Times Series. The second dataset we use is a chaotic time series based on the Hénon map (Hénon, 1976). The two-dimensional variables $X = (X_t^{(1)}, X_t^{(2)})$ and $Y = (Y_t^{(1)}, Y_t^{(2)})$ are computed using only the values from the previous time step as follows:

$$X_t^{(1)} = 1.4 - {X_{t-1}^{(1)}}^2 + 0.3X_{t-1}^{(2)}$$
$$Y_t^{(1)} = 1.4 - \left[\gamma X_{t-1}^{(1)} Y_{t-1}^{(1)} + (1-\gamma){Y_{t-1}^{(1)}}^2\right] + 0.3Y_{t-1}^{(2)}$$
$$X_t^{(2)} = X_{t-1}^{(1)}, \quad Y_t^{(2)} = Y_{t-1}^{(1)}.$$

The parameter γ controls the effect that previous values of X have on Y. To increase the difficulty of this task, two additional independent Gaussian noise variables with zero mean and standard deviation $\sigma = 0.5$ are concatenated to both X and Y. Here, conditional dependence and independence characterizes the causal influence from X to Y when $\gamma > 0$. Namely, in this dataset, $X_{t+1} \perp\!\!\!\perp Y_t \mid X_t$, but $Y_{t+1} \not\perp\!\!\!\perp X_t \mid Y_t$.

For each underlying joint distribution, we generate the cumulative density function (CDF) of the p-values obtained for each test across the 300 random samples. While the CDFs of the tests' p-values are useful for understanding the global behavior of the tests,[5] it is more succinct to summarize each curve with a single statistic. In prior work, the powers and type I errors at a particular, fixed value of α are used to summarize results (Fukumizu et al., 2008; Gretton et al., 2012a; Zhang et al., 2011). However, presenting results in this way can be misleading if one of the tests has an advantage at a particular value of α. Therefore, we use two statistics to summarize the power and type I error across values of α. When the test has high power, it correctly rejects the null hypothesis even when α is small. Therefore, the area under the CDF, or *power curve* is close to 1.0. On the other hand, when the null hypothesis is true, a *well-calibrated* test will produce uniformly-distributed p-values so that the type I error rate is equal to α. In this case, the CDF is a diagonal line with slope 1. To measure calibratedness, the Kolmogorov test can be used to quantify the difference between the empirically observed CDF and that for the uniform distribution. Since sample sizes are finite and null distributions are only approximately estimated, the null hypothesis of perfect calibratedness will likely be rejected by the Kolmogorov test after enough tests are performed. However, the relative ($\log$) p-values corresponding to the Kolmogorov test can be used to compare calibratedness; larger p-values roughly correspond to better calibration.

[4] The code is available online at `http://engr.case.edu/doran_gary/code.html`

[5] See the supplementary materials for details.

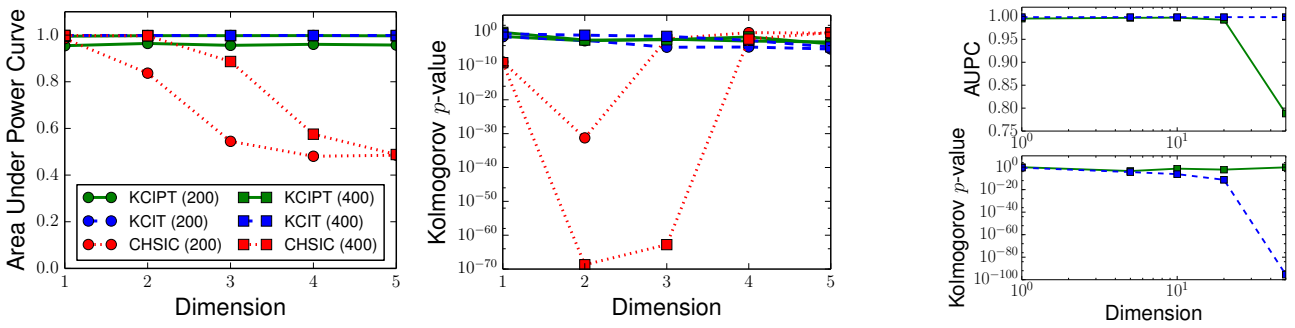

Figure 3: (**Left**) Summarized results for the post-nonlinear noise dataset. These results clearly show how the power of HSIC decreases as noise is added to the conditioning variable. (**Right**) A comparison of KCIT and KCIPT for high-dimensional datasets. The performance of the tests begin to degrade in different ways, with the power of KCIPT falling to chance levels while the KCIT becomes poorly calibrated between $D = 10$ and $D = 50$.

Figure 3 (left) shows results for the post-nonlinear noise dataset as the dimensionality D of the conditioning variable increases. Since Y is a function of Z, the function-distance distortion measure is used, as described in Section 2.4.[6] Gaussian process regression is used to find the function f relating Z and Y. As observed in prior work (Zhang et al., 2011), the CHSIC approach is sensitive to the dimensionality of the conditioning variable, so power quickly decreases as D increases. With a dataset of size 200, KCIT is slightly more powerful than KCIPT, but the performance converges as the sample size increases to 400. Furthermore, the performance of KCIPT is preserved as dimensionality increases, since the regression-based distance effectively serves as dimensionality reduction on the conditioning variable.

Figure 3 (right) shows what happens to both KCIPT and KCIT as the dimensionality of the dataset continues to increase to $D = 50$; both tests fail by this point, but in different ways. KCIT becomes very poorly calibrated between $D = 10$ and $D = 50$, while the power of KCIPT degrades around the same dimensionality. We conjecture that the observed behavior is due to the differences in kernel parameter selection for each approach. The kernel values for KCIT are chosen heuristically depending on dataset size, but the test is only evaluated on low-dimensional datasets (Zhang et al., 2011). As dimensionality increases, the heuristic is less effective, and the test poorly estimates the null distribution. By contrast, the KCIPT uses the median heuristic, which automatically adjusts the kernel parameter as dataset size and dimensionality increase. Thus, the null distribution is correctly estimated, but the test statistic becomes less powerful on this dataset.

The results for the chaotic time series are shown in Figure 4. For this test, the RKHS distance is used as the distortion measure to learn the permutation. The behavior of the

[6]We compare this distortion measure with other choices in the supplementary materials.

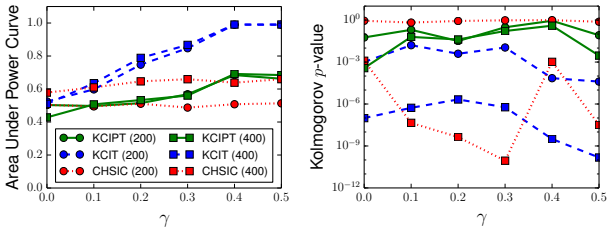

Figure 4: Results for the chaotic time series. As expected, the power of these tests increases as the conditional dependence controlled by γ increases. The KCIT is not well-calibrated on this dataset, and HSIC becomes less well-calibrated as sample size increases.

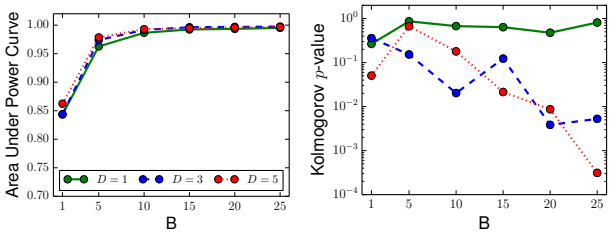

Figure 5: Effects of bootstrapping the test statistic with B iterations on the post-nonlinear noise dataset with $n = 400$. Bootstrapping increases the power of the test, but also decreases the calibration when dimensionality D of the conditioning variable increases. The observed effect is likely due to the approximation errors induced by the permutation leading to an over-rejection of the null hypothesis.

tests is shown as γ increases. In this noisy chaotic dataset, conditional dependence is more difficult to detect, and none of the techniques perform very well when γ is small. Although KCIT has the best performance in terms of power, it is poorly-calibrated as the sample size increases. In fact, both the CHSIC and KCIT become *less* well-calibrated as

sample size increases, suggesting systematic errors in null distribution estimation. For CHSIC, it appears that there are difficulties in finding permutations to estimate the null distribution, and for KCIT, the chaotic nature of the dataset might violate its assumption that variables are related by continuous, well-behaved functions.

Using the post-nonlinear noise dataset with $n = 400$, we also quantify the extent to which the outer bootstrapping procedure described in Section 2.3 improves the power of the test. Figure 5 shows the power and calibration of the test as the number of bootstraps B increases; $B = 1$ corresponds to no bootstrapping of the test statistic, and $B = 25$ is used in the previous experiments. Bootstrapping the test statistic does in fact increase power for this dataset. However, when the dimensionality of Z grows, the calibration of the test decreases. We believe that this behavior is a result of the approximation error induced by the permutation; as dimensionality increases, it becomes harder to find an appropriate permutation with a fixed sample size. However, we observe in Figure 3 (left) that the other tests also tend to be poorly calibrated on this dataset as the dimensionality of Z increases. These results do not suggest a general procedure for selecting B, but they illustrate that at least for the post-nonlinear noise data, there is a power–calibration trade-off involved in the use of bootstrapping.

Medical Data. Finally, we explore the application of the KCIPT to a real-world dataset used in prior work (Fukumizu et al., 2008). The data consists of three variables, creatinine clearance (C), digoxin clearance (D), and urine flow (U), measured on 35 patients. The ground truth, that $D \perp\!\!\!\perp U \mid C$, is known for this dataset. We try to recover this relationship using the PC algorithm (Spirtes, Glymour, and Scheines, 2000), with the KCIPT and $\alpha = 0.05$ used as a test for conditional independence. We choose $B = 10$, since it appears to be an effective setting that reduces the overall computation time (Figure 5). The output of the PC algorithm is the Markov equivalence class D—C—U, which contains the only causal structures (either $D \leftarrow C \leftarrow U$, $D \rightarrow C \rightarrow U$, or $D \leftarrow C \rightarrow U$) consistent with the ground truth conditional independence relationship and pairwise dependence relationships, assuming there are no unobserved confounding variables.

5 DISCUSSION

In relation to existing kernel-based conditional independence tests, a major advantage of KCIPT observed in our empirical analysis is its ability to accurately estimate the null distribution. Hence, we observe that KCIPT is well-calibrated across the synthetic datasets we study, even under the more extreme scenarios when the dimensionality of the conditioning variable is large or there are complex, nonlinear relationships between variables in the joint distribution. Our results align with those observed in prior work, in which permutation-based conditional independence tests for datasets with discrete values were found to be well-calibrated with respect to asymptotic tests (Tsamardinos and Borboudakis, 2010). Additionally, using a well-calibrated conditional independence test produces more robust solutions in Bayesian network learning.

The need for conditional independence testing is ubiquitous in the sciences. Unfortunately, performing the test in practice is known to be very challenging. This work not only simplifies the problem, but also present a general framework for conditional independence testing which can be extended immediately to numerous settings. Thus, there remain many interesting extensions and questions to study in future work, such as applications to non-i.i.d. data, different approaches for learning a permutation, and deciding which variable to permute in the asymmetric test statistic. Furthermore, we look forward to applying KCIPT to real-world datasets with more complex conditional dependence relationships.

6 CONCLUSION

In this work, we propose a new conditional independence test that employs a permutation to generate an artificial sample from a joint distribution for which the null hypothesis of the test holds. Effectively, we transform the conditional independence test into a two-sample test problem, which is easier to solve, well-studied, and scales to high-dimensional datasets. We use a kernel-based two sample test between an original sample and a permuted sample, which share the same distribution if and only if the conditional independence relationship holds. Prior knowledge about the joint distribution can be incorporated into the process of finding an appropriate permutation. The resulting test has power competitive with existing kernel-based approaches for conditional independence testing and better estimates the null distribution on the datasets used for evaluation. In future work, we will explore theoretical relationships between our approach and those using partial association and further investigate the use of our test for applications in causal discovery.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions.

References

Aizerman, A.; Braverman, E. M.; and Rozoner, L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control* 25:821–837.

Baba, K.; Shibata, R.; and Sibuya, M. 2004. Partial correlation and conditional correlation as measures of condi-

tional independence. *Australian & New Zealand Journal of Statistics* 46(4):657–664.

Bergsma, W. 2004. Testing conditional independence for continuous random variables. EURANDOM-report 2004-049.

Berlinet, A., and Thomas-Agnan, C. 2004. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Springer.

Birkhoff, G. 1946. Three observations on linear algebra. *Univ. Nac. Tucumán. Revista A.* 5:147–151.

Daudin, J. 1980. Partial association measures and an application to qualitative regression. *Biometrika* 67(3):581–590.

Dawid, A. P. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–31.

Fukumizu, K.; Gretton, A.; Sun, X.; and Schölkopf, B. 2008. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, 489–496.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, 513–520.

Gretton, A.; Fukumizu, K.; Teo, C. H.; Song, L.; Schölkopf, B.; and Smola, A. J. 2008. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, 585–592.

Gretton, A.; Fukumizu, K.; Sriperumbudur, B. K.; et al. 2009. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, 673–681.

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

Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; and Sriperumbudur, B. K. 2012b. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, 1214–1222.

Hénon, M. 1976. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics* 50(1):69–77.

Hoyer, P.; Janzing, D.; Mooij, J.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, 689–696.

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

Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2013. Supplement to: Quantifying causal influences. *The Annals of Statistics*. DOI: 10.1214/13-AOS1145SUPP.

Margaritis, D. 2005. Distribution-free learning of Bayesian network structure in continuous domains. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, 825.

Rasmussen, C., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts, USA: MIT Press.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Schölkopf, B., and Smola, A. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Smola, A.; Gretton, A.; Song, L.; and Schölkopf, B. 2007. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, 13–31. Springer.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, prediction, and search*, volume 81. The MIT Press.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Lanckriet, G. R.; and Schölkopf, B. 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, 1750–1758.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Schölkopf, B.; and Lanckriet, G. R. 2010. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* 99:1517–1561.

Su, L., and White, H. 2008. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* 24(4):829.

Tsamardinos, I., and Borboudakis, G. 2010. Permutation testing improves Bayesian network learning. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 322–337.

von Neumann, J. 1953. A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games* 2:5–12.

Zhang, K., and Hyvärinen, A. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Annual Conference on Uncertainty in Artificial Intelligence*, 804–813.