A Permutation-Based Kernel Conditional Independence Test: Supplementary Materials

A PROOFS

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

Proof. Since mean and variance must be invariant for any sample, in particular it must be invariant when Y is a random variable that takes values in $\mathbb{R}$. Instead of being an operator (as in the general case when sample elements are vectors), $var(\mathbf{y})$ is a real number, and $\mathbf{P}$ must satisfy $var(\mathbf{Py}) = var(\mathbf{y})$. Furthermore, we can consider $\mathbf{y}$ a vector in $\mathbb{R}^n$, in which case the permutation $\mathbf{Py}$ corresponds to usual matrix multiplication.

Now, suppose further that $\mathbf{y}$ is a sample of all zeros except for a one in the k^{th} entry. Then invariance of the mean requires that:

$$\frac{1}{n}\sum_i \sum_j \mathbf{P}_{ij}\mathbf{y}_j = \frac{1}{n} \implies \sum_i \mathbf{P}_{ik} = 1. \tag{1}$$

Since this is true for any $1 \leq k \leq n$, all columns of $\mathbf{P}$ must sum to one.

In vector form, (biased) sample variance can be expressed as:

$$var(\mathbf{y}) = \frac{1}{n}\mathbf{y}^\intercal\mathbf{y} - mean(\mathbf{y})^2. \tag{2}$$

Since the means of the permuted and unpermuted samples are equal, equality of variance implies equality of:

$$\mathbf{y}^\intercal\mathbf{y} = (\mathbf{Py})^\intercal\mathbf{Py} = \mathbf{y}^\intercal\mathbf{P}^\intercal\mathbf{Py}. \tag{3}$$

Since this must hold for any $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{P}^\intercal\mathbf{P} = \mathbf{I}$, the identity matrix, so $\mathbf{P}$ must be an orthogonal matrix. Since the entries of $\mathbf{P}$ are nonnegative, only a single entry in each row and column of $\mathbf{P}$ can be nonzero. Since each column of $\mathbf{P}$ must sum to one, the nonzero entries of $\mathbf{P}$ must equal one. Therefore, $\mathbf{P}$ is a permutation matrix. □

Theorem. *Let $\mathbf{D}^{\mathrm{RKHS}}_{ij} = \|\phi_z(z_i) - \phi_z(z_j)\|$ be a pairwise reproducing kernel Hilbert space (RKHS) distance matrix between Z values in a sample. The proposed test statistic is asymptotically consistent if the quantity $\min_{\mathbf{P}\in\mathcal{P}} \frac{1}{n}\operatorname{Tr}(\mathbf{PD}^{\mathrm{RKHS}})$ converges in probability to zero as $n \to \infty$.*

Proof. Let $\Pr_{xyz}$ be the distribution corresponding to the original sample and $\Pr'_{xyz}$ be the distribution corresponding to the permuted sample. This distribution can be factored as $\Pr'_{xy|z}\Pr'_z$. Since the values of Y are permuted arbitrarily w.r.t. those of X and samples are independent and identically distributed (i.i.d.), any dependency between X and Y in the sample is broken. So the distribution can be factored further as $\Pr'_{x|z}\Pr'_{y|z}\Pr'_z$. Since X and Z are left unchanged, $\Pr'_{x|z} = \Pr_{x|z}$ and $\Pr'_z = \Pr_z$. Hence, the permuted distribution equals the factorized joint distribution $\Pr_{x|z}\Pr_{y|z}\Pr_z$ iff $\Pr'_{y|z} = \Pr_{y|z}$. By the invariance of $\Pr_z$, this occurs iff $\Pr'_{yz} = \Pr_{yz}$.

In prior work, Sriperumbudur et al. present conditions under which the maximum mean discrepancy (MMD) metrizes the weak topology on probability measures (2010). For example, it is sufficient that underlying kernel be universal and the sample space be compact. When the MMD metrizes the weak topology, the MMD between the permuted distribution $\Pr'_{yz}$ converges in distribution to $\Pr_{yz}$ iff the MMD between permuted and unpermuted joint samples $(\mathbf{Py}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ converge to zero. For any particular sample and permutation π on indices, this MMD can be written as:

$$\begin{aligned}&\mathrm{MMD}((\mathbf{Py}, \mathbf{z}), (\mathbf{y}, \mathbf{z}))\\ &= \left\| \tfrac{1}{n}\sum_i \phi_y(y_i) \otimes \phi_z(z_i) - \tfrac{1}{n}\sum_j \phi_y(y_{\pi(j)}) \otimes \phi_z(z_j) \right\|\end{aligned}$$

$$
\begin{aligned}
&= \Bigg\| \frac{1}{n} \sum_i \phi_y(y_i) \otimes \phi_z(z_i) \\
&\quad - \frac{1}{n} \sum_{i=\pi(j)} \phi_y(y_i) \otimes \phi_z(z_{\pi^{-1}(i)}) \Bigg\| \\
&= \left\| \frac{1}{n} \sum_i \phi_y(y_i) \otimes \phi_z(z_i) - \phi_y(y_i) \otimes \phi_z(z_{\pi^{-1}(i)}) \right\| \\
&= \left\| \frac{1}{n} \sum_i \phi_y(y_i) \otimes (\phi_z(z_i) - \phi_z(z_{\pi^{-1}(i)})) \right\| \\
&\leq \frac{1}{n} \sum_i \left\| \phi_y(y_i) \otimes (\phi_z(z_i) - \phi_z(z_{\pi^{-1}(i)})) \right\| \\
&= \frac{1}{n} \sum_i \left\| \phi_y(y_i) \right\| \left\| \phi_z(z_i) - \phi_z(z_{\pi^{-1}(i)}) \right\|.
\end{aligned}
$$

When $\|\phi_y(y_i)\| \leq M$ (for the Gaussian kernel, $M = 1$), this implies that:

$$
\begin{aligned}
&\mathrm{MMD}((\mathbf{P}\mathbf{y}, \mathbf{z}), (\mathbf{y}, \mathbf{z})) \\
&\qquad \leq M(\frac{1}{n} \sum_i \left\| \phi_z(z_i) - \phi_z(z_{\pi^{-1}(i)}) \right\|) \\
&\qquad = M(\tfrac{1}{n} \operatorname{Tr}(\mathbf{P}^{-1}\mathbf{D}^{\mathrm{RKHS}})) \\
&\qquad = M(\tfrac{1}{n} \operatorname{Tr}(\mathbf{P}^{\intercal}\mathbf{D}^{\mathrm{RKHS}})) \\
&\qquad = M(\tfrac{1}{n} \operatorname{Tr}(\mathbf{P}\mathbf{D}^{\mathrm{RKHS}})),
\end{aligned}
$$

where the last two steps follow from the properties that $\mathbf{P}$ is orthogonal and $\mathbf{D}^{\mathrm{RKHS}}$ is symmetric.

Therefore, 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. □

B DETAILED RESULTS

Detailed results are given in Figure 1 and Figure 2. These figures are plots of the cumulative density functions (CDFs) of p-values for tests when null hypothesis does and does not hold. These curves represent the type I error and power, respectively, of the tests as α varies. For power curves, the farther up and to the left the curve is, the more power the test has at low values of α. For type I error, curves should lie directly along the diagonal, since a test should only reject the null hypothesis with probability α when it is correct.

B.1 EVALUATING PERMUTATIONS

In the paper, we discuss several different techniques for learning the permutation $\mathbf{P}$. Below, we empirically evaluate the differences between these approaches using the post-nonlinear noise synthetic dataset described above. In this dataset, only the first conditioning variable Z_1 determines conditional independence. Therefore, as argued in Equation A, the average of pairwise RKHS distances between the permuted and unpermuted Z_1 values is an indication of ground truth for how "distorted" the permuted sample is w.r.t. the MMD.

Figure 3 shows the pairwise RKHS distance between Z_1 values for four different approaches to learning the permutation as sample size and the dimensionality of the conditioning variable Z increase. The presented distances are averaged across 30 trials. The first approach randomly choses a permutation. The resulting test effectively tests whether $Y \perp\!\!\!\perp (X, Z)$, not whether $Y \perp\!\!\!\perp X \mid Z$, since any dependence between Y and Z is broken by a random permutation. We also see from Figure 3 that the distortion of the permutation is relatively large.

The second approach is to directly minimize the RKHS distance between Z values. As expected, this approach works well when the dimension of Z is only 1. However, as D increases, the regression function approaches perform better. There are two variants of using the regression function distance as a distortion measure. The first finds a function $f(Z) = Y$ using Gaussian process regression, and uses $\|f(z_i) - f(z_j)\|_2$ as a distance metric. The second uses the fact that for this dataset, both X and Y are functions of Z, so a function $f(Z) = [XY]$, where $[XY]$ is the concatenation of X and Y, is found using the same process, which is used in a distance metric. The intuition is that a better distance metric can be found by incorporating information about both variables. Note that this last approach is very problem-specific, since it assumes that the same, though unknown, set of variable in Z determines both X and Y. It is included as an example of how domain-specific knowledge might be incorporated into the distortion measure.

When sample sizes are small, it is difficult to learn a regression function $f(Z) = Y$, so the RKHS distance initially leads to better performance. However, as sample size increases, an accurate regression function is learned, and performance improves much more quickly with sample size. These results justify the use of the regression distance metric for sample sizes between 200–400, as we did above.

References

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.

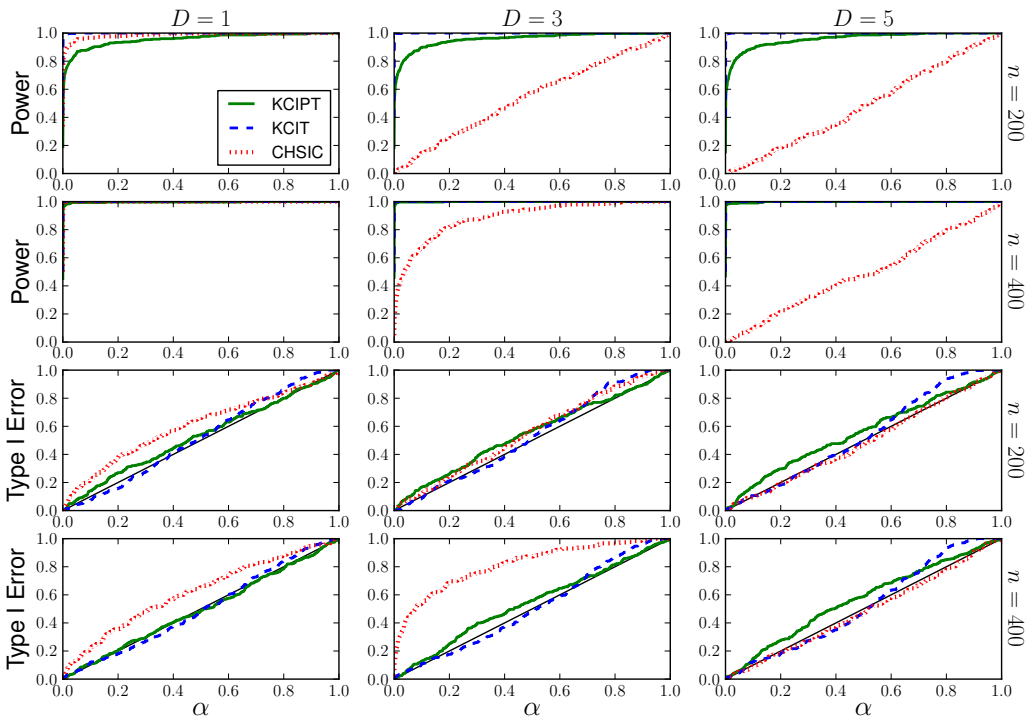

Figure 1: Performance of each test across sample sizes n and conditioning variable dimensionality D on the post-nonlinear noise dataset. The power and type I error are shown here as α varies between 0 and 1. The CHSIC approach loses power as D increases. Both KCIT and KCIPT maintain power as the conditioning variable increases. When sample sizes are small, KCIT has slightly more power than KCIPT does. However, at a sample size of $n = 400$, both KCIT and KCIPT almost perfectly reject the null hypothesis. The ideal type I error behavior is shown by the thin diagonal line. In terms of type I error, CHSIC is occasionally very poorly calibrated.

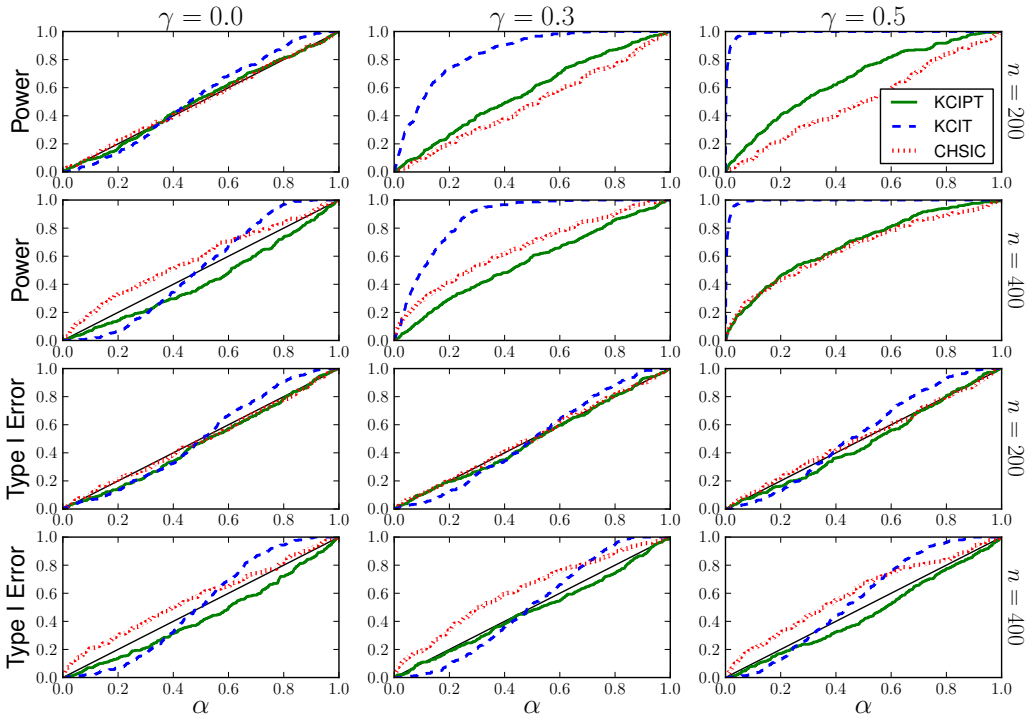

Figure 2: Performance of each test across sample sizes n and "coupling" parameter γ on the Hénon chaotic time series. The power and type I error are shown here as α varies between 0 and 1. When $\gamma = 0$, the variables are conditionally independent, so the power curve is ideally a diagonal line in this case. Otherwise, the conditional dependence, and test power, tends to increase with γ.

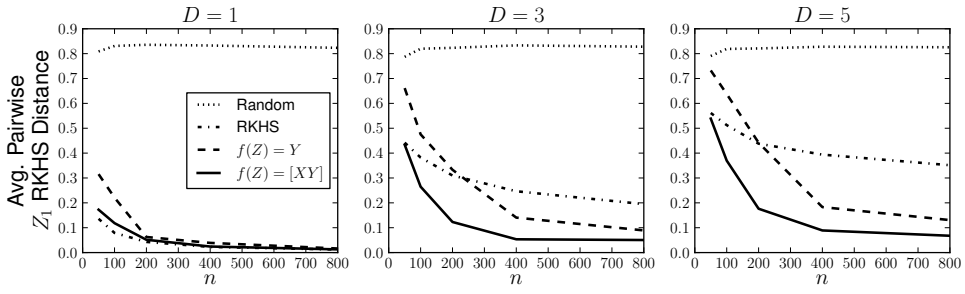

Figure 3: A comparison of permutation learning approaches on the post-nonlinear noise synthetic dataset. The y-axis measures the "distortion" of the relevant conditioning variable as additional noise variables are added to the conditioning set. Theoretically, a randomly chosen permutation is not sufficient to test for conditional independence, and empirically such permutations cause high distortion. The RKHS distance works well when the size of the conditioning set is small, but when the dimensionality of Z grows and the sample size is large enough, the regression-based distances provide better permutations.