Eric V. Strobl*, Kun Zhang, and Shyam Visweswaran

Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery

https://doi.org/10.1515/jci-2018-0017
Received February 5, 2017; revised October 10, 2018; accepted November 19, 2018

Abstract: Constraint-based causal discovery (CCD) algorithms require fast and accurate conditional independence (CI) testing. The Kernel Conditional Independence Test (KCIT) is currently one of the most popular CI tests in the non-parametric setting, but many investigators cannot use KCIT with large datasets because the test scales at least quadratically with sample size. We therefore devise two relaxations called the Randomized Conditional Independence Test (RCIT) and the Randomized conditional Correlation Test (RCoT) which both approximate KCIT by utilizing random Fourier features. In practice, both of the proposed tests scale linearly with sample size and return accurate p-values much faster than KCIT in the large sample size context. CCD algorithms run with RCIT or RCoT also return graphs at least as accurate as the same algorithms run with KCIT but with large reductions in run time.

Keywords: Conditional Independence Test, Random Fourier Features, Causal Discovery, Non-Parametric

1 The problem

Constraint-based causal discovery (CCD) algorithms such as Peter-Clark (PC) and Fast Causal Inference (FCI) infer causal relations from observational data by combining the results of many conditional independence (CI) tests [1]. In practice, a CCD algorithm can easily request p-values from thousands of CI tests even with a sparse underlying graph. Developing fast and accurate CI tests is therefore critical for maximizing the usability of CCD algorithms across a wide variety of datasets.

Investigators have developed many fast parametric methods for testing CI. For example, we can use partial correlation to test for CI under the assumption of Gaussian variables [2, 3]. We can also consider testing for unconditional independence $X \perp\!\!\!\perp Y|Z = z$ for each constant value z when Z is discrete and $\mathbb{P}(Z = z) > 0$. The chi-squared test for instance utilizes this strategy when both X and Y are also discrete [4]. Another permutation-based test generalizes the same strategy even when X and Y are not necessarily discrete [5].

Testing for CI in the non-parametric setting generally demands a more sophisticated approach. One strategy involves discretizing continuous conditioning variables Z as $\check{Z}$ in some optimal fashion and assessing unconditional independence $\forall \check{Z} = \check{z}$ [6, 7]. Discretization however suffers severely from the curse of dimensionality because consistency arguments demand smaller bins with increasing sample size, but the number of cells in the associated contingency table increases exponentially with the conditioning set size. A second method involves measuring the distance between estimates of the conditional densities $f(X|Y, Z)$ and $f(X|Z)$, or their associated characteristic functions, by observing that $f(X|Y, Z) = f(X|Z)$ when $X \perp\!\!\!\perp Y|Z$ [8, 9]. However, the power of these tests also deteriorates quickly with increases in the dimensionality of Z.

Article note: R implementation at github.com/ericstrobl/RCIT. We recommend that users install Microsoft R Open for fast matrix computations.

***Corresponding author: Eric V. Strobl,** University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, United States, e-mail: ericvonstrobl@gmail.com
Kun Zhang, Carnegie Mellon University, Department of Philosophy, Pittsburgh, United States, e-mail: kunz1@cmu.edu
Shyam Visweswaran, University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, United States, e-mail: shv3@pitt.edu

Several investigators have since proposed reproducing kernel-based CI tests in order to tame the curse of dimensionality. Indeed, kernel-based methods in general are known for their strong empirical performance in the high dimensional setting. The Kernel Conditional Independence Test (KCIT) for example assesses CI by capitalizing on a characterization of CI in reproducing kernel Hilbert spaces (RKHSs; [10]). Intuitively, KCIT works by testing for vanishing regression residuals among functions in RKHSs. Another kernel-based CI test called the Permutation Conditional Independence Test (PCIT) reduces CI testing to two-sample kernel-based testing via a carefully chosen permutation found at the solution of a convex optimization problem [11].

The aforementioned kernel-based CI tests unfortunately suffer from an important drawback: both tests scale at least quadratically with sample size and therefore take too long to return a p-value in the large sample size setting. In particular, KCIT's bottleneck lies in the eigendecomposition as well as the inversion of large kernel matrices [10], and PCIT must solve a linear program that scales cubicly with sample size in order to obtain its required permutation [11]. As a general rule, it is difficult to develop exact kernel-based methods which scale sub-quadratically with sample size, since the computation of kernel matrices themselves scales at least quadratically.

Many investigators have nonetheless utilized *random Fourier features* in order to quickly approximate kernel methods in linear time with respect to the number of Fourier features. For example, Lopez-Paz and colleagues developed an unconditional independence test using statistics obtained from canonical correlation analysis with random Fourier features [12]. Zhang and colleagues have also utilized random Fourier features for unconditional independence testing but took a different approach by approximating the kernel cross-covariance operator [13]. The authors further analyzed block and Nyström-based kernel approximations to the unconditional independence testing problem. The authors ultimately concluded that the random Fourier feature and the Nyström-based approaches both outperformed the block-based approach on average. Others have analyzed the use of random Fourier features for predictive modeling (e. g., [14, 15]) or dimensionality reduction [16]. In practice, investigators have observed that methods which utilize random Fourier features often scale linearly with sample size and achieve comparable accuracy to exact kernel methods [12, 13, 14, 15, 16].

In this paper, we also use random Fourier features to design two fast tests called the Randomized Conditional Independence Test (RCIT) and the Randomized conditional Correlation Test (RCoT) which approximate the solution of KCIT. Simulations show that RCIT, RCoT and KCIT have comparable accuracy, but both RCIT and RCoT scale linearly with sample size in practice. As a result, RCIT and RCoT return p-values several orders of magnitude faster than KCIT in the large sample size context. Moreover, experiments demonstrate that the causal structures returned by CCD algorithms using either RCIT, RCoT or KCIT have nearly identical accuracy.

2 High-level summary

We now provide a high-level summary for investigators who simply wish to perform CCD with RCIT or RCoT. We want to test whether we have $X \perp\!\!\!\perp Y|Z$ in a fast and accurate manner without resorting to parametric assumptions. Previously, Zhang and colleagues introduced a non-parametric CI test called KCIT which can test whether we have $X \perp\!\!\!\perp Y|Z$ in an accurate but not fast manner by analyzing the partial cross-covariance (operator) using the kernel method [10]. Kernels however are expensive to compute because they scale at least quadratically with sample size. Another line of work has fortunately shown that we can approximate kernels by averaging over a variety of non-linear transformations called random Fourier features (e. g., [14, 15, 17]). We therefore propose to approximate KCIT by utilizing random Fourier features, specifically by analyzing the partial cross-covariance matrix of $\{X, Z\}$ and Y (RCIT) or X and Y (RCoT) after subjecting the variable sets to the non-linear transformations and then non-linearly regressing out the effect of Z. Simulations show that RCIT and RCoT return p-values in a much shorter time frame while also matching or outperforming KCIT in approximating the null distribution. RCoT in particular also returns the most accurate p-values when the conditioning set size Z is large (≥ 4).

3 Preliminaries on kernels

We will deal with kernels in this paper, so we briefly review the corresponding theory; see [18] for a more extensive discussion of similar concepts. Capital letters X, Y, Z denote sets of random variables with codomains $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. Let $\mathcal{H}_\mathcal{X}$ correspond to a Hilbert space of functions mapping $\mathcal{X}$ to $\mathbb{R}$. We say that $\mathcal{H}_\mathcal{X}$ is more specifically a *reproducing kernel Hilbert space*, if the Dirac delta operator $\delta_x : \mathcal{H}_\mathcal{X} \mapsto \mathbb{R}$ (which maps $f \in \mathcal{H}_\mathcal{X}$ to $f(x) \in \mathbb{R}$) is a bounded linear functional [19]. We can associate $\mathcal{H}_\mathcal{X}$ with a unique positive definite *kernel* $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ in which the *feature map* $\phi(x) : \mathcal{X} \mapsto \mathcal{H}_\mathcal{X}$ satisfies the *reproducing property* $\langle f, \phi(x) \rangle_{\mathcal{H}_\mathcal{X}} = f(x), \forall f \in \mathcal{H}_\mathcal{X}, \forall x \in \mathcal{X}$ by the Moore-Aronszajn Theorem. We will require that $\mathcal{H}_\mathcal{X}$ be *separable* (i. e., it must have a complete orthonormal system) throughout this paper [18]. Hein and Bousquet [20] showed that any continuous kernel on a separable space $\mathcal{X}$ (e. g., $\mathbb{R}^d$) induces a separable RKHS. We will likewise consider other kernels such as $k_\mathcal{Y}$ on the separable space $\mathcal{Y}$.

We have the following norm defined on linear operators between RKHSs:

Definition. *[21] Denote by $\Sigma : \mathcal{H}_\mathcal{X} \mapsto \mathcal{H}_\mathcal{Y}$ a linear operator. Then, provided that the sum converges, the Hilbert Schmidt (HS) norm of Σ is defined as:*

$$\|\Sigma\|_{HS}^2 = \sum_{i,j} \langle u_i, \Sigma v_j \rangle_{\mathcal{H}_\mathcal{Y}}^2, \tag{1}$$

where u_i and v_j are orthonormal bases of $\mathcal{H}_\mathcal{Y}$ and $\mathcal{H}_\mathcal{X}$, respectively.

We denote the probability distribution of X as $\mathbb{P}_X$ and that of Y as $\mathbb{P}_Y$. We may then define the *mean elements* $\mathbb{E}[\phi(X)] = \mu_X \in \mathcal{H}_\mathcal{X}$ and $\mathbb{E}[\psi(Y)] = \mu_Y \in \mathcal{H}_\mathcal{Y}$ with respect to the probability distributions; here, ϕ denotes the feature map from $\mathcal{X}$ to $\mathcal{H}_\mathcal{X}$, and ψ likewise denotes the feature map from $\mathcal{Y}$ to $\mathcal{H}_\mathcal{Y}$. We also define the quantity $\|\mu_X\|_{\mathcal{H}_\mathcal{X}}^2$ by applying the expectation twice:

$$\|\mu_X\|_{\mathcal{H}_\mathcal{X}}^2 = \mathbb{E}_{XX'}[\langle \phi(X), \phi(X') \rangle_{\mathcal{H}_\mathcal{X}}] = \mathbb{E}_{XX'}[k_\mathcal{X}(X, X')], \tag{2}$$

where X' is an independent copy of X which follows the same distribution. The mean elements μ_X and μ_Y exist so long as their respective norms in $\mathcal{H}_\mathcal{X}$ or $\mathcal{H}_\mathcal{Y}$ are bounded; this is true so long as the kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ are also bounded so that $\mathbb{E}_{XX'}[k_\mathcal{X}(X, X')] < \infty$ and $\mathbb{E}_{YY'}[k_\mathcal{Y}(Y, Y')] < \infty$ [18].

The *cross-covariance operator* associated with the joint probability distribution $\mathbb{P}_{XY}$ over (X, Y) is a linear operator $\Sigma_{XY} : \mathcal{H}_\mathcal{Y} \mapsto \mathcal{H}_\mathcal{X}$ defined as follows:

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)], \tag{3}$$

for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. Gretton et al. [18] showed that the HS-norm of the cross covariance operator, $\|\Sigma_{XY}\|_{HS}^2$, can be written in terms of kernels as follows:

$$\|\Sigma_{XY}\|_{HS}^2 = \mathbb{E}_{XX'YY'}[k_\mathcal{X}(X, X')k_\mathcal{Y}(Y, Y')], \tag{4}$$

provided that $X \perp\!\!\!\perp Y$, and $k_\mathcal{X}$ as well as $k_\mathcal{Y}$ are *centered* (i. e., $\mathbb{E}_{XX'}[k_\mathcal{X}(X, X')] = 0$ and $\mathbb{E}_{YY'}[k_\mathcal{Y}(Y, Y')] = 0$). We can therefore consider the following empirical estimate of the $\|\Sigma_{XY}\|_{HS}^2$ using n i. i. d. samples $\boldsymbol{x}, \boldsymbol{y}$ as follows, if we assume that $X \perp\!\!\!\perp Y$ [10, 18]:

$$\mathcal{T}_{\boldsymbol{xy}} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [\widetilde{K}_X]_{ij} [\widetilde{K}_Y]_{ij} = \frac{1}{n^2} \mathrm{tr}[\widetilde{K}_X \widetilde{K}_Y], \tag{5}$$

where $\mathcal{T}_{\boldsymbol{xy}} \xrightarrow{p} \|\Sigma_{XY}\|_{HS}^2$ [18]; the dependent case can be found in Lemma 1 of the same paper for the interested reader. The notations $\widetilde{K}_X$ and $\widetilde{K}_Y$ correspond to *centralized kernel matrices* such that:

$$\widetilde{K}_X = HK_XH, \tag{6}$$

and likewise for $\widetilde{K}_Y$. Here, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, I denotes an $n \times n$ identity matrix, and $\mathbf{1}$ denotes a vector of ones. The notation K_X denotes a *kernel matrix* such as the RBF kernel where $[K_X]_{ij} = \exp\Big(-\frac{\|x_i - x_j\|_2^2}{\sigma}\Big)$ with $x_i, x_j \in \boldsymbol{x}$. The transformation in Equation 6 ensures that $\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n [\widetilde{K}_X]_{ij} = 0$ similar to the centered kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ in Equation 4.

4 Characterizations of conditional independence

We denote the probability distribution of X as $\mathbb{P}_X$ and the joint probability distribution of (X, Z) as $\mathbb{P}_{XZ}$. Let L^2_X denote the space of square integrable functions of X, and L^2_{XZ} that of (X, Z). Here, $L^2_X = \{s(X) \mid \mathbb{E}_X(|s|^2) < \infty\}$ and likewise for L^2_{XZ}. Next consider a dataset of n i. i. d. samples drawn according to $\mathbb{P}_{XYZ}$.

We use the notation $X \perp\!\!\!\perp Y|Z$ when X and Y are conditionally independent given Z. Perhaps the simplest characterization of CI reads as follows: $X \perp\!\!\!\perp Y|Z$ if and only if $\mathbb{P}_{XY|Z} = \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}$. Equivalently, we have $\mathbb{P}_{X|YZ} = \mathbb{P}_{X|Z}$ and $\mathbb{P}_{Y|XZ} = \mathbb{P}_{Y|Z}$.

4.1 Characterization by RKHSs

A second characterization of CI is given in terms of the cross-covariance operator Σ_{XY} on RKHSs [22]. Recall the cross-covariance operator from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}_\mathcal{X}$ in Equation 3. We may then define the partial cross-covariance operator of (X, Y) given Z by:

$$\Sigma_{XY\cdot Z} = \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}, \tag{7}$$

where we use the right inverse instead of the inverse, if Σ_{ZZ} is not invertible (see Corollary 3 in Fukumizu et al. [22]). Notice the similarity of the partial cross-covariance operator to the linear partial cross-covariance matrix (as well as the conditional cross-covariance matrix in the Gaussian case).[1] We can interpret the above equation as the partial covariance between $\{f(X), \forall f \in \mathcal{H}_\mathcal{X}\}$ and $\{g(Y), \forall g \in \mathcal{H}_\mathcal{Y}\}$ given $\{h(Z), \forall h \in \mathcal{H}_\mathcal{Z}\}$ (i. e., the partial covariance of X and Y given Z after passing these three variable sets through the functions in the RKHSs $\mathcal{H}_X, \mathcal{H}_Y$ and $\mathcal{H}_Z$).

A kernel k_X is characteristic if $\mathbb{E}_{X\sim\mathbb{P}_X}[f(X)] = \mathbb{E}_{X\sim\mathbb{Q}_X}[f(X)]$, $\forall f \in \mathcal{H}_\mathcal{X}$ implies $\mathbb{P}_X = \mathbb{Q}_X$, where $\mathbb{P}_X$ and $\mathbb{Q}_X$ are two probability distributions of X [23]; alternatively, a kernel is characteristic if equality in the mean elements under the two distributions $\mu_{\mathbb{P}_X} = \mu_{\mathbb{Q}_X}$ implies equality of the distributions. Two examples of characteristic kernels include the Gaussian kernel and the Laplacian kernel. Now if we use characteristic kernels in (7), then the partial cross-covariance operator is related to the CI relation via the following conclusion:

Proposition 1. *[22, 23] Let $\ddot{X} = (X, Z)$ and $k_{\ddot{\mathcal{X}}} = k_\mathcal{X} k_\mathcal{Z}$. Also let $\mathcal{H}_{\ddot{\mathcal{X}}}$ represent the RKHS corresponding to $k_{\ddot{\mathcal{X}}}$. Assume $\mathbb{E}[k_\mathcal{X}(X, X)] < \infty$ and $\mathbb{E}[k_\mathcal{Y}(Y, Y)] < \infty$.[2] Further assume that $k_{\ddot{\mathcal{X}}} k_\mathcal{Y}$ is a characteristic kernel on $(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Z}$, and that $\mathcal{H}_\mathcal{Z} + \mathbb{R}$ (the direct sum of the two RKHSs) is dense in L^2_Z. Then*

$$\Sigma_{\ddot{X}Y\cdot Z} = 0 \iff X \perp\!\!\!\perp Y|Z. \tag{8}$$

Here, $\Sigma_{\ddot{X}Y\cdot Z} = 0$ means that $\langle f, \Sigma_{\ddot{X}Y\cdot Z} g\rangle_{\mathcal{H}_{\ddot{\mathcal{X}}}} = 0$ for all $f \in \mathcal{H}_{\ddot{\mathcal{X}}}$ and all $g \in \mathcal{H}_\mathcal{Y}$. Recall further that $\Sigma_{\ddot{X}Y\cdot Z} = 0 \iff \|\Sigma_{\ddot{X}Y\cdot Z}\|^2_{HS} = 0$ because $\|\Sigma_{\ddot{X}Y\cdot Z}\|^2_{HS} = \sum_{i,j}\langle u_i, \Sigma_{\ddot{X}Y\cdot Z} v_j\rangle^2_{\mathcal{H}_{\ddot{\mathcal{X}}}}$, where u_i and v_j are orthonormal bases of $\mathcal{H}_{\ddot{\mathcal{X}}}$ and $\mathcal{H}_\mathcal{Y}$, respectively, by Definition 1.

4.2 Characterization by L^2 spaces

We also consider a different characterization of CI which is intuitively more appealing because it allows to use to directly view CI as the uncorrelatedness of functions in certain spaces rather than a norm of the partial cross-covariance operator. In particular, consider the following constrained L^2 spaces:

$$\mathcal{F}_{XZ} \triangleq \{\widetilde{f} \in L^2_{XZ} \mid \mathbb{E}(\widetilde{f}|Z) = 0\},$$

1 Recall that the partial cross-covariance of X and Y given Z is defined as $\mathbb{E}[(X - \mathbb{E}(X|Z))(Y - \mathbb{E}(Y|Z))]$; in other words, it is equivalent to the cross-covariance of X and Y given Z. In contrast, the conditional cross-covariance of X and Y given Z is defined as $\mathbb{E}[(X - \mathbb{E}(X|Z))(Y - \mathbb{E}(Y|Z))|Z]$ (notice the extra conditioning).

2 This assumption ensures that $\mathcal{H}_\mathcal{X} \subset L^2_X$ and $\mathcal{H}_\mathcal{Y} \subset L^2_Y$.

$$\begin{aligned}\mathcal{F}_{YZ} &\triangleq \{\widetilde{g} \in L^2_{YZ} \mid \mathbb{E}(\widetilde{g}|Z) = 0\},\\ \mathcal{F}_{Y\cdot Z} &\triangleq \{\widetilde{h}' \mid \widetilde{h}' = h'(Y) - \mathbb{E}(h'|Z), h' \in L^2_Y\}.\end{aligned} \tag{9}$$

Here, we write $\mathbb{E}(\widetilde{f}|Z = z)$ as shorthand for $\mathbb{E}(\widetilde{f}(\cdot, z))$ when $\widetilde{f} \in L^2_{XZ}$, and likewise for $\widetilde{g} \in L^2_{YZ}$. Notice that we can construct the three spaces listed above using nonlinear regression. For instance, for any $f \in L^2_{XZ}$ in $\mathcal{F}_{XZ}$, we have:

$$\begin{aligned}\widetilde{f}(\ddot{X}) &= f(\ddot{X}) - \mathbb{E}(f|Z)\\ &= f(\ddot{X}) - h^*(Z),\end{aligned} \tag{10}$$

where $h^*(Z) \in L^2_Z$ is the regression function of $f(\ddot{X})$ on Z.

We then have the following result:

Proposition 2. *[24] The following conditions are equivalent:*

1. $X \perp\!\!\!\perp Y|Z$,
2. $\mathbb{E}(\widetilde{f}\widetilde{g}) = 0$, $\forall \widetilde{f} \in \mathcal{F}_{XZ}$ and $\forall \widetilde{g} \in \mathcal{F}_{YZ}$,
3. $\mathbb{E}(\widetilde{f}g) = 0$, $\forall \widetilde{f} \in \mathcal{F}_{XZ}$ and $\forall g \in L^2_{YZ}$,
4. $\mathbb{E}(\widetilde{f}\widetilde{h}') = 0$, $\forall \widetilde{f} \in \mathcal{F}_{XZ}$ and $\forall \widetilde{h}' \in \mathcal{F}_{Y\cdot Z}$,
5. $\mathbb{E}(\widetilde{f}g') = 0$, $\forall \widetilde{f} \in \mathcal{F}_{XZ}$ and $\forall g' \in L^2_Y$.

The second condition means that any "residual" function of (X, Z) given Z is uncorrelated with that of (Y, Z) given Z. The equivalence also represents a generalization of the case when (X, Y, Z) is jointly Gaussian; here, $X \perp\!\!\!\perp Y|Z$ if and only if any residual function of X given Z is uncorrelated with that of Y given Z; i. e., the linear partial correlation coefficient $\rho_{XY\cdot Z}$ is zero.

We also encourage the reader to observe the close relationship between Proposition 1 and claim 4 of Proposition 2. Notice that claim 4 of Proposition 2 implies that we have $\Sigma_{\ddot{X}Y\cdot Z} = 0$ in Proposition 1. Moreover, Proposition 1 only considers functions in RKHSs, while claim 4 of Proposition 2 considers functions in L^2 spaces. We find Proposition 1 more useful than claim 4 of Proposition 2 because the RKHS of a characteristic kernel might be much smaller than the corresponding L^2 space.

5 Kernel conditional independence test

We now consider the following hypotheses:

$$\begin{aligned}H_0 &: X \perp\!\!\!\perp Y|Z,\\ H_1 &: X \not\perp\!\!\!\perp Y|Z.\end{aligned} \tag{11}$$

We may equivalently rewrite the above null and alternative more explicitly using Proposition 1 as follows, provided that the kernels are chosen such that the premises of that proposition are satisfied (e. g., when the kernels are Gaussian or Laplacian [22]):

$$\begin{aligned}H_0 &: \|\Sigma_{\ddot{X}Y\cdot Z}\|^2_{HS} = 0,\\ H_1 &: \|\Sigma_{\ddot{X}Y\cdot Z}\|^2_{HS} > 0.\end{aligned} \tag{12}$$

The above hypothesis implies that we can test for conditional independence by testing for uncorrelatedness between functions in reproducing kernel Hilbert spaces.

Zhang et al. [10] exploited the equivalence between 11 and 12 in the Kernel Conditional Independence Test (KCIT), which we now describe. Consider the functional spaces $f \in \mathcal{H}_{\ddot{\mathcal{X}}}$, $g \in \mathcal{H}_{\mathcal{Y}}$, and $h \in \mathcal{H}_{\mathcal{Z}}$. We can compute the corresponding centered kernel matrices $\widetilde{K}_{\ddot{X}}$, $\widetilde{K}_Y$ and $\widetilde{K}_Z$ from n i. i. d. samples $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ as in 6. We can then use these matrices to perform kernel ridge regression in order to estimate the function $h^* \in \mathcal{H}_{\mathcal{Z}}$ in 10 as follows: $\widehat{h}^*(\boldsymbol{z}) = \widetilde{K}_Z(\widetilde{K}_Z + \lambda I)^{-1} f(\ddot{\boldsymbol{x}})$, where λ denotes the ridge regularization parameter and $f \in \mathcal{H}_{\ddot{\mathcal{X}}}$. Consequently, we can estimate the residual function $\widetilde{f}$ as $\widetilde{f}(\ddot{\boldsymbol{x}}) = f(\ddot{\boldsymbol{x}}) - \widehat{h}^*(\boldsymbol{z}) = R_{\mathcal{Z}} f(\ddot{\boldsymbol{x}})$ where:

$$\begin{aligned} R_{\mathcal{Z}} &= I - \widetilde{K}_Z(\widetilde{K}_Z + \lambda I)^{-1} \\ &= \lambda(\widetilde{K}_Z + \lambda I)^{-1}. \end{aligned} \tag{13}$$

Next consider the eigenvalue decomposition $\widetilde{K}_{\ddot{X}} = V_{\ddot{X}}\Lambda_{\ddot{X}}V_{\ddot{X}}$. Let $\phi_{\ddot{X}} = V_{\ddot{X}}\Lambda_{\ddot{X}}^{1/2}$. Then the *empirical* kernel map of $\mathcal{H}_{\ddot{X}|Z}$ is given by $\widetilde{\phi}_{\ddot{X}} = R_Z\phi_{\ddot{X}}$ because $\widetilde{K}_{\ddot{X}|Z} = \widetilde{\phi}_{\ddot{X}}\widetilde{\phi}_{\ddot{X}}^T$. We may similarly write $\widetilde{K}_{Y|Z} = \widetilde{\phi}_Y\widetilde{\phi}_Y^T$. We can thus write the centralized kernel matrices corresponding to $\widetilde{f}$ and $\widetilde{g}$ as follows:

$$\begin{aligned} \widetilde{K}_{\ddot{X}|Z} &= R_Z\widetilde{K}_{\ddot{X}}R_Z, \\ \widetilde{K}_{Y|Z} &= R_Z\widetilde{K}_Y R_Z. \end{aligned} \tag{14}$$

We can use the above two centered kernel matrices to compute an empirical estimate of the HS norm of the partial cross covariance operator similar to how we computed the quantity $\mathcal{T}_{XY}$ in Equation 5:

$$\mathcal{T}_{\ddot{\boldsymbol{x}}\boldsymbol{y}\cdot\boldsymbol{z}} = \frac{1}{n^2}\mathrm{tr}(\widetilde{K}_{\ddot{X}|Z}\widetilde{K}_{Y|Z}). \tag{15}$$

Note that $\mathcal{T}_{\ddot{\boldsymbol{x}}\boldsymbol{y}\cdot\boldsymbol{z}}$ denotes an empirical estimate of $\|\Sigma_{\ddot{X}Y\cdot Z}\|_{HS}^2$. KCIT uses the statistic $\mathcal{S}_K = n\mathcal{T}_{\ddot{\boldsymbol{x}}\boldsymbol{y}\cdot\boldsymbol{z}}$ in order to determine whether or not to reject H_0; we multiply the empirical estimate of the HS norm by n in order to ensure convergence to a non-degenerate distribution.

6 Proposed test statistic & its asymptotic distribution

Observe that computing $\mathcal{S}_K$ requires the inversion of large kernel matrices which scales strictly greater than quadratically with sample size; the exact complexity depends on the algorithm used to compute the matrix inverse (e. g., $O(n^{2.376})$ if we use the Coppersmith-Winograd algorithm [25]). CCD algorithms run with KCIT therefore take too long to complete. In this report, we will propose an inexpensive hypothesis test that almost always rejects or fails to reject the null whenever conditional dependence or independence holds, respectively, and even outperforms 12 as illustrated in our experimental results in Section 7. We will use a strategy that has already been successfully adopted in the context of unconditional dependence testing in doing so [12, 13].

We will in particular also take advantage of the characterization of CI presented in Proposition 1. Recall that the Frobenius norm corresponds to the Hilbert-Schmidt norm in Euclidean space. We therefore consider the following hypotheses as approximations to those in 12 and 11:

$$\begin{aligned} H_0 &: \|\mathcal{C}_{\ddot{A}B\cdot Z}\|_F^2 = 0, \\ H_1 &: \|\mathcal{C}_{\ddot{A}B\cdot Z}\|_F^2 > 0, \end{aligned} \tag{16}$$

where $\mathcal{C}_{\ddot{A}B\cdot Z} = \mathbb{E}[(\ddot{A}_i - \mathbb{E}(\ddot{A}|Z))(B_i - \mathbb{E}(B|Z))^T]$ corresponds to the ordinary partial cross covariance matrix, where $\mathbb{E}(\ddot{A}|Z)$ and $\mathbb{E}(B|Z)$ may be non-linear functions of Z. We also have $\ddot{A} = f'(\ddot{X}) \triangleq \{f_1'(\ddot{X}), \ldots, f_m'(\ddot{X})\}$ with $f_j'(\ddot{X}) \in \mathcal{G}_{\ddot{X}}$, $\forall j$. Similarly, $B = h'(Y) \triangleq \{h_1'(Y), \ldots, h_q'(Y)\}$ with $h_k'(Y) \in \mathcal{G}_Y$, $\forall k$. The terms $\mathcal{G}_{\ddot{X}}$ and $\mathcal{G}_Y$ denote two spaces of functions, which we set to be the support of the process $\sqrt{2}\cos(W^T \cdot + B)$, $W \sim \mathbb{P}_W$, $B \sim \mathrm{Uniform}([0, 2\pi])$. In other words, we select m functions from $\mathcal{G}_{\ddot{X}}$ and q functions from $\mathcal{G}_Y$. We select these specific spaces because we can use them to approximate continuous shift-invariant kernels. A kernel k is said to be shift-invariant if and only if, for any $a \in \mathbb{R}^p$, we have $k(x - a, y - a) = k(x, y)$, $\forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^p$; examples of shift-invariant kernels include the Gaussian kernel frequently used in KCIT or the Laplacian kernel. The following result allows us to perform the approximation using the proposed spaces:

Proposition 3. *[14] For a continuous shift-invariant kernel $k(x, y)$ on $\mathbb{R}^p$, we have:*

$$k(x, y) = \int_{\mathbb{R}^p} e^{iw^T(x-y)}\, dF_w = \mathbb{E}[\zeta(x)\zeta(y)], \tag{17}$$

where F_W represents the CDF of $\mathbb{P}_W$ and $\zeta(x) = \sqrt{2}\cos(W^T x + B)$ with $W \sim \mathbb{P}_W$ and $B \sim \mathrm{Uniform}([0, 2\pi])$.

The precise form of $\mathbb{P}_W$ depends on the type of shift-invariant kernel one would like to approximate (see Figure 1 of Rahimi and Recht [14] for a list). Since investigators most frequently implement KCIT using the Gaussian kernel $k_\sigma(x,y) = \exp(-\|x-y\|^2/\sigma)$ with hyperparameter σ, we choose to approximate the Gaussian kernel by setting $\mathbb{P}_W$ to a centered Gaussian with standard deviation $\sqrt{\sigma/2}$.

We will use the squared Frobenius norm of the empirical partial cross-covariance matrix as the statistic for RCIT:

$$\mathcal{S} = n\|\widehat{\mathcal{C}}_{\ddot{A}B\cdot Z}\|_F^2, \tag{18}$$

where $\widehat{\mathcal{C}}_{\ddot{A}B\cdot Z} = \frac{1}{n-1}\sum_{i=1}^n[(\ddot{A}_i - \widehat{\mathbb{E}}(\ddot{A}|Z))(B_i - \widehat{\mathbb{E}}(B|Z))^T]$. Recall however that $\mathbb{E}(\ddot{A}|Z)$ and $\mathbb{E}(B|Z)$ may be non-linear functions of Z and therefore difficult to estimate. We would instead like to approximate $\mathbb{E}(\ddot{A}|Z)$ and $\mathbb{E}(B|Z)$ with linear functions. We therefore let $C = g(Z) \triangleq \{g_1(Z),\dots,g_d(Z)\}$ with $g_l(Z) \in \mathcal{G}_Z, \forall l$, where we also set $\mathcal{G}_Z$ to be the support of the process $\sqrt{2}\cos(W^T\cdot + B)$, $W \sim \mathbb{P}_W$, $B \sim \text{Uniform}([0,2\pi])$. We will approximate $\mathcal{C}_{\ddot{A}B\cdot Z}$ with $\widehat{\mathcal{C}}_{\ddot{A}B\cdot C} = \widehat{\mathcal{C}}_{\ddot{A}B} - \widehat{\mathcal{C}}_{\ddot{A}C}(\widehat{\Sigma}_{CC} + \gamma I)^{-1}\widehat{\mathcal{C}}_{CB}$ similar to 7, where γ denotes a small ridge parameter; recall that this is equivalent to computing the cross-covariance matrix across the residuals of $\ddot{A}$ and B given C using linear ridge regression. We can justify this procedure because the particular choice of C allows us to approximate both $\mathbb{E}(\ddot{A}|Z)$ and $\mathbb{E}(B|Z)$ with linear functions of C as described below.

Let $f_j = f_j' - \mathbb{E}(f_j'|Z)$. Then $\mathbb{E}(f_j|Z) = 0$, so $f_j \in \mathcal{F}_{XZ}$. Moreover, $h_k' - \mathbb{E}(h_k'|Z) \in \mathcal{F}_{Y\cdot Z}$. Note that we can estimate $\mathbb{E}(f_j'|Z)$ with the linear ridge regression solution $\widehat{u}_j^T g(Z)$ under mild conditions because we can guarantee the following:

Proposition 4. *(Section 3.1 of Sutherland and Schneider [15]) Consider performing kernel ridge regression of f_j' on Z. Assume that (1) $\sum_{i=1}^n f_{j,i}' = 0$ and (2) the empirical kernel matrix of Z, k_Z, only has finite entries (i. e., $\|k_Z\|_\infty < \infty$). Further assume that the range of Z, $\mathcal{Z} \subset \mathbb{R}^{d_Z}$, is compact. We then have:*

$$\mathbb{P}[|\check{\mathbb{E}}(f_j'|Z) - \widehat{u}_j^T g(Z)| \geq \varepsilon] \leq \frac{c_0}{\varepsilon^2}\mathrm{e}^{-d\varepsilon^2 c_1}, \tag{19}$$

where $\check{\mathbb{E}}(f_j'|Z)$ denotes the estimate of $\mathbb{E}(f_j'|Z)$ by kernel ridge regression, and c_0 and c_1 are both constants that do not depend on n or d.

The exponential rate with respect to d in the above proposition suggests we can approximate the output of kernel ridge regression with a small number of random Fourier features, a hypothesis which we verify empirically in Section 7. Moreover, we can estimate $\mathbb{E}(h_k'|Z)$ with $\widehat{u}_k^T g(Z)$, because we can similarly guarantee that $\mathbb{P}[|\check{\mathbb{E}}(h_k'|Z) - \widehat{u}_k^T g(Z)| \geq \varepsilon] \to 0$ for any fixed $\varepsilon > 0$ at an exponential rate with respect to d.

We can therefore consider the following spaces for $\mathcal{S}$ which are similar to the L^2 spaces used in claim 4 of Proposition 2:

$$\begin{aligned}\widehat{\mathcal{G}}_{\ddot{X}} &\triangleq \{f \mid f_j = f_j' - \mathbb{E}(f_j'|Z), f_j' \in \mathcal{G}_{\ddot{X}}\},\\ \widehat{\mathcal{G}}_{Y\cdot Z} &\triangleq \{h \mid h_k = h_k' - \mathbb{E}(h_k'|Z), h_k' \in \mathcal{G}_Y\}.\end{aligned} \tag{20}$$

We then approximate CI with $\mathcal{S}$ in the following sense:

1. We always have $X \perp\!\!\!\perp Y|Z \implies \mathbb{E}(fh) = 0, \forall f \in \widehat{\mathcal{G}}_{\ddot{X}}$ and $\forall h \in \widehat{\mathcal{G}}_{Y\cdot Z}$.
2. The reverse direction may hold for a larger subset of all possible joint distributions as the values of m and q increase; this is because at least one entry of $\mathcal{C}_{\ddot{A}B\cdot C}$ will likely be greater than zero for any given distribution as the values of m, q increase.

Note the second point refers to the population $\mathcal{C}_{\ddot{A}B\cdot C}$ as opposed to its finite sample estimate. In this paper, we only deal with the classical low dimensional scenario where m, q are fixed and the sample size $n \to \infty$. This is reasonable because nearly all CB algorithms only test for CI when X and Y each contain a single variable. We find that the second point held in all of the cases we tested in Section 7 with only $m, q = 5$, since we were always able to reject the null $H_0 : \|\mathcal{C}_{\ddot{A}B\cdot C}\|_F^2 = 0$ by generating enough samples with $m, q = 5$ when $X \not\perp\!\!\!\perp Y|Z$.

6.1 Null distribution

We now consider the asymptotic distribution of $\mathcal{S}$ under the null. Let Π refer to a positive definite covariance matrix of the vectorization of $(\ddot{A} - \mathbb{E}(\ddot{A}|C))(B - \mathbb{E}(B|C))^T$. We may denote an arbitrary entry in Π as follows:

$$\begin{aligned} &\Pi_{\ddot{A}_i B_j, \ddot{A}_k B_l} \\ &= \mathbb{E}\big[(\ddot{A}_i - \mathbb{E}(\ddot{A}_i|C))(B_j - \mathbb{E}(B_j|C))(\ddot{A}_k - \mathbb{E}(A_k|C))(B_l - \mathbb{E}(B_l|C))\big]. \end{aligned} \tag{21}$$

We have the following result:

Theorem 1. *Consider n i. i. d. samples from $\mathbb{P}_{XYZ}$. Let $\{z_1, \ldots, z_L\}$ denote i. i. d. standard Gaussian variables (thus $\{z_1^2, \ldots, z_L^2\}$ denotes i.i.d χ_1^2 variables) and λ the eigenvalues of Π. We then have the following asymptotic distribution under the null in 16:*

$$n\|\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}\|_F^2 \xrightarrow{d} \sum_{i=1}^{L} \lambda_i z_i^2, \tag{22}$$

where L refers to the number of elements in $\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}$.

Proof. We may first write:

$$\begin{aligned} &n\|\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}\|_F^2 \\ &= n * \mathrm{tr}(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}^T) \\ &= n * v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})^T v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}), \\ &= [\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})]^T [\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})], \end{aligned} \tag{23}$$

where $v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})$ stands for the vectorization of $\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}$. By CLT of the sample covariance matrix (see Lemma 1 in the Appendix A) combined with the continuous mapping theorem and the null, we know that $\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C}) \xrightarrow{d} \mathcal{N}(0, \Pi)$. Here, we write an arbitrary entry $\Pi_{\ddot{A}_i B_j, \ddot{A}_k B_l}$ under the null as follows:

$$\begin{aligned} &\Pi_{\ddot{A}_i B_j, \ddot{A}_k B_l} \\ &= \mathrm{Cov}\big[(\ddot{A}_i - \mathbb{E}(\ddot{A}_i|C))(B_j - \mathbb{E}(B_j|C)), \\ &\qquad\qquad (\ddot{A}_k - \mathbb{E}(\ddot{A}_k|C))(B_l - \mathbb{E}(B_l|C))\big] \\ &= \mathbb{E}\big[(\ddot{A}_i - \mathbb{E}(\ddot{A}_i|C))(B_j - \mathbb{E}(B_j|C)) * \\ &\qquad\qquad (\ddot{A}_k - \mathbb{E}(\ddot{A}_k|C))(B_l - \mathbb{E}(B_l|C))\big]. \end{aligned} \tag{24}$$

Now consider the eigendecomposition of Π written as $\Pi = E\Lambda E^T$. Then, we have $E^T[\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})] \xrightarrow{d} \mathcal{N}(0, \Lambda)$ by the continuous mapping theorem. Note that:

$$\begin{aligned} &[\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})]^T [\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})] \\ &= (E^T[\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})])^T (E^T[\sqrt{n}v(\widehat{\mathcal{C}}_{\ddot{A}B\cdot C})]) \\ &\xrightarrow{d} \sum_{i=1}^{L} \lambda_i z_i^2. \end{aligned} \tag{25}$$

□

We conclude that the null distribution of the test statistic is a positively weighted sum of i. i. d. χ_1^2 random variables.

Note that multiple methods exist for estimating the conditional expectations in $\mathcal{S}$ and Π in the above theorem. In this report, we will obtain consistent estimates of the conditional expectations by using kernel ridge regressions with the RBF kernel; here, consistency holds so long as the conditional expectations are continuous because the RBF kernel is dense in the space of continuous functions mapping $\mathcal{Z}$ to $\mathbb{R}$ [26]. We

therefore have $\check{C}_{\ddot{A}B\cdot C} \xrightarrow{p} C_{\ddot{A}B\cdot C}$, where $\check{C}_{\ddot{A}B\cdot C} = \frac{1}{n-1}\sum_{i=1}^{n}[(\ddot{A}_i - \check{\mathbb{E}}(\ddot{A}|C))(B_i - \check{\mathbb{E}}(B|C))^T]$, by the continuous mapping theorem and weak law of large numbers assuming continuity of the conditional expectations. We can similarly approximate any arbitrary entry in Π because we may write the following:

$$\begin{aligned}&\frac{1}{n}\sum_{r=1}^{n}(\ddot{A}_{i,r} - \check{\mathbb{E}}(\ddot{A}_i|C))(B_{j,r} - \check{\mathbb{E}}(B_j|C))* \\ &\qquad (\ddot{A}_{k,r} - \check{\mathbb{E}}(\ddot{A}_k|C))(B_{l,r} - \tilde{\mathbb{E}}(B_l|C)) \\ \xrightarrow{p}\ &\mathbb{E}\big[(\ddot{A}_i - \mathbb{E}(\ddot{A}_i|C))(B_j - \mathbb{E}(B_j|C))* \\ &\qquad (\ddot{A}_k - \mathbb{E}(\ddot{A}_k|C))(B_l - \mathbb{E}(B_l|C))\big].\end{aligned} \tag{26}$$

Kernel ridge regressions however scale (strictly greater than) quadratically with sample size due to the inversion of the kernel matrix, so they may not be practical in the large sample size regime. Fortunately, we will not need to perform the kernel ridge regressions directly, because we can approximate the output of kernel ridge regression to within an arbitrary degree of accuracy for any fixed sample size n using linear ridge regression with enough random Fourier features as highlighted previously in Proposition 4. In particular, Proposition 4 implies that $\tilde{u}^T g(Z) \equiv \tilde{\mathbb{E}}(\ddot{A}|C) \xrightarrow{p} \check{\mathbb{E}}(\ddot{A}|C)$ at rate exponential in d for any fixed n. We can also conclude that $\tilde{\Sigma}_{\ddot{A}B\cdot C} \xrightarrow{p} \check{\Sigma}_{\ddot{A}B\cdot C}$ as $d \to \infty$ for any fixed n. We can finally approximate the kernel ridge regression estimate of Π because we may write the following for an arbitrary entry in Π as $d \to \infty$:

$$\begin{aligned}&\frac{1}{n}\sum_{r=1}^{n}(\ddot{A}_{i,r} - \tilde{\mathbb{E}}(\ddot{A}_i|C))(B_{j,r} - \tilde{\mathbb{E}}(B_j|C))* \\ &\qquad (\ddot{A}_{k,r} - \tilde{\mathbb{E}}(\ddot{A}_k|C))(B_{l,r} - \tilde{\mathbb{E}}(B_l|C)) \\ \xrightarrow{p}\ &\frac{1}{n}\sum_{r=1}^{n}(\ddot{A}_{i,r} - \check{\mathbb{E}}(\ddot{A}_i|C))(B_{j,r} - \check{\mathbb{E}}(B_j|C))* \\ &\qquad (\ddot{A}_{k,r} - \check{\mathbb{E}}(\ddot{A}_k|C))(B_{l,r} - \check{\mathbb{E}}(B_l|C)).\end{aligned} \tag{27}$$

We conclude that, for a dataset of fixed sample size n, we can substitute the conditional expectation estimates of kernel ridge regression with those of linear regression with random Fourier features when estimating $\mathcal{S}$ as well as Π for applying Theorem 1.

Unfortunately though, a closed form CDF of a positively weighted sum of chi-squared random variables does not exist in general for applying Theorem 1. We can approximate the CDF by Imhof's method which inverts the characteristic function numerically [27]. We should consider Imhof's method as exact, since it provides error bounds and can be used to compute the distribution at a fixed point to within a desired precision [28, 29]. However, Imhof's method is too computationally intensive for our purposes. We can nonetheless utilize several fast methods which approximate the null by moment matching.

6.2 Approximating the null distribution by moment matching

We write the cumulants of a positively weighted sum of i. i. d. χ_1^2 random variables as follows:

$$c_r = 2^{r-1}(r-1)!\sum_{i=1}^{L}\lambda_i^r, \tag{28}$$

where $\lambda = \{\lambda_1, \ldots, \lambda_L\}$ denotes the weights. We may for example derive the first three cumulants:

$$m_1 = \sum_{i=1}^{L}\lambda_i, \quad m_2 = 2\sum_{i=1}^{L}\lambda_i^2, \quad m_3 = 8\sum_{i=1}^{L}\lambda_i^3. \tag{29}$$

We then recover the moments from the cumulants as follows:

$$m_r = c_r + \sum_{i=1}^{r-1}\binom{r-1}{i-1}c_i m_{r-i}, \quad r = 2, 3, \ldots \tag{30}$$

Now the Satterthwaite-Welch method [30, 31, 32] represents perhaps the simplest and earliest moment matching method. The method matches the first two moments of the sum with a gamma distribution $\Gamma(\widehat{g}, \widehat{\theta})$. Zhang and colleagues adopted a similar strategy in their paper introducing KCIT [10]. Here, we have:

$$\widehat{g} = \frac{1}{2}c_1^2/c_2, \quad \widehat{\theta} = c_2/c_1. \tag{31}$$

We however find the above gamma approximation rather crude. We therefore also consider applying more modern methods to estimating the distribution of a sum of positively weighted chi-squares. Improved methods such as the Hall-Buckley-Eagleson [33, 34] and the Wood F [35] methods match the first three moments of the sum to other distributions in a similar fashion. On the other hand, the Lindsay-Pilla-Basak method [36] matches the first $2L$ moments to a mixture distribution.

We will focus on the Lindsay-Pilla-Basak method in this paper, since Bodenham & Adams have already determined that the Lindsay-Pilla-Basak method performs the best through extensive experimentation [37, 38]. We therefore choose to use the method as the default method for RCIT. Briefly, the method approximates the CDF under the null $F_{\mathcal{H}_0}$ using a finite mixture of L Gamma CDFs $F_{\Gamma(g,\theta_i)}$:

$$F_{\mathcal{H}_0} = \sum_{i=1}^{L} \pi_i F_{\Gamma(g,\theta_i)}, \tag{32}$$

where $\pi \geq 0$, $\sum_{i=1}^{L} \pi_i = 1$, and we seek to determine the $2L + 1$ parameters g, $\theta_1, \ldots, \theta_L$, and $\pi_1, \ldots, \pi_L$. The Lindsay-Pilla-Basak method computes these parameters by a specific sequence of steps that makes use of results concerning moment matrices (see Appendix II in Uspensky [39]). The sequence is complicated and beyond the scope of this paper, but we refer the reader to [36] for details.

6.3 Testing for conditional un-correlatedness

Strictly speaking, we must consider the extended variable set $\ddot{X}$ to test for conditional independence according to Proposition 1. However, we have two observations: (1) we can substitute a test for non-linear conditional uncorrelatedness with tests for conditional independence in almost all cases encountered in practice because most conditionally dependent variables are correlated after some functional transformations, and (2) using the extended variable set $\ddot{X}$ makes estimating the null distribution more difficult compared to using the unextended variable set X. The first observation coincides with the observations of others who have noticed that Fisher's z-test performs well (but not perfectly) in ruling out conditional independencies with non-Gaussian data. We can also justify the first observation with the following result using the cross-covariance operator $\Sigma_{XY \cdot Z}$:

Proposition 5. *[22, 23] Assume* $\mathbb{E}[k_{\mathcal{X}}(X, X)] < \infty$ *and* $\mathbb{E}[k_{\mathcal{Y}}(Y, Y)] < \infty$. *Further assume that* $k_{\mathcal{X}} k_{\mathcal{Y}}$ *is a characteristic kernel on* $\mathcal{X} \times \mathcal{Y}$, *and that* $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ *(the direct sum of the two RKHSs) is dense in* L^2_Z. *Then*

$$\Sigma_{XY \cdot Z} = 0 \iff \mathbb{E}_Z[\mathbb{P}_{XY|Z}] = \mathbb{E}_Z[\mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}]. \tag{33}$$

In other words, we have:

$$\begin{aligned} \Sigma_{XY \cdot Z} = 0 &\implies \mathbb{P}_{XY} = \int \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}\, d\mathbb{P}_Z, \\ \Sigma_{XY \cdot Z} = 0 &\impliedby \mathbb{E}_Z[\mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}] = \mathbb{E}_Z[\mathbb{P}_{XY|Z}] \impliedby X \perp\!\!\!\perp Y|Z. \end{aligned} \tag{34}$$

Notice that $\Sigma_{XY \cdot Z} = 0$ is almost equivalent to CI, in the sense that $\Sigma_{XY \cdot Z} = 0$ just misses those rather contrived distributions where $\mathbb{P}_{XY} = \int \mathbb{P}_{XY|Z}\, d\mathbb{P}_Z = \int \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}\, d\mathbb{P}_Z$ when $X \perp\!\!\!\perp Y|Z$. In other words, if $\mathbb{P}_{XY} \neq \int \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}\, d\mathbb{P}_Z$ when $X \perp\!\!\!\perp Y|Z$, then we have $\Sigma_{XY \cdot Z} = 0 \iff \Sigma_{\ddot{X}Y \cdot Z} = 0$ (under the corresponding additional assumptions of Propositions 1 and 5).

Table 1: Example of a situation where $\int \mathbb{P}_{XY|Z}\, d\mathbb{P}_Z = \int \mathbb{P}_{X|Z}P_{Y|Z}\, d\mathbb{P}_Z$ when $X \not\perp\!\!\!\perp Y|Z$ using binary variables.

	$\mathbb{P}_{X\|Z=0}$	$\mathbb{P}_{X\|Z=1}$
$X=0$	0.5	0.3
$X=1$	0.5	0.7

(a)

	$\mathbb{P}_{Y\|Z=0}$	$\mathbb{P}_{Y\|Z=1}$
$Y=0$	0.3	0.4
$Y=1$	0.7	0.6

(b)

	$\mathbb{P}_{XY\|Z=0}$	$\mathbb{P}_{XY\|Z=1}$
$X=0, Y=0$	0.2	0.1075
$X=0, Y=1$	0.3	0.1925
$X=1, Y=0$	0.1	0.2925
$X=1, Y=1$	0.4	0.4075

(c)

	$\mathbb{P}_{XY}$
$X=0, Y=0$	0.126
$X=0, Y=1$	0.214
$X=1, Y=0$	0.254
$X=1, Y=1$	0.406

(d)

Let us now consider an example of a situation where $\int \mathbb{P}_{XY|Z}\, d\mathbb{P}_Z \neq \int \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}\, d\mathbb{P}_Z$ when $X \not\perp\!\!\!\perp Y|Z$. Take three binary variables $X, Y, Z \in \{0,1\}$. Let $\mathbb{P}_{Z=0} = 0.2$ and $\mathbb{P}_{Z=1} = 0.8$. Also consider the four probability tables in Table 1. Here, we have chosen the probabilities in the tables carefully by satisfying the following equation:

$$\begin{aligned} &\int \mathbb{P}_{XY|Z}\, d\mathbb{P}_Z = \int \mathbb{P}_{X|Z}P_{Y|Z}\, d\mathbb{P}_Z \\ \iff &\mathbb{P}_{Z=0}(\mathbb{P}_{X|Z=0}\mathbb{P}_{Y|Z=0}) + \mathbb{P}_{Z=1}(\mathbb{P}_{X|Z=1}\mathbb{P}_{Y|Z=1}) \\ &= \mathbb{P}_{Z=0}\mathbb{P}_{XY|Z=0} + \mathbb{P}_{Z=1}\mathbb{P}_{XY|Z=1}. \end{aligned} \tag{35}$$

Of course, the equality holds when we have conditional independence $\mathbb{P}_{XY|Z} = \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}$. We are however interested in the case when conditional dependence holds. We therefore instantiated the values of Tables 1a and 1b as well as the second column in Table 1c ($\mathbb{P}_{XY|Z=0}$) such that $\mathbb{P}_{XY|Z=0} \neq \mathbb{P}_{X|Z=0}\mathbb{P}_{Y|Z=0}$. We then solved for $\mathbb{P}_{XY|Z=1}$ using Equation 35 in order to complete Table 1c. This ultimately yielded Table 1d.

Notice that we obtain a unique value for $\mathbb{P}_{XY|Z=1}$ by solving Equation 35. Hence, $\mathbb{P}_{XY|Z=1}$ has Lebesgue measure zero on the interval $[0,1]$, once we have defined all of the other variables in the equation. Thus, $\Sigma_{XY\cdot Z} = 0$ is not always equivalent to $X \perp\!\!\!\perp Y|Z$, but satisfying the condition $\int \mathbb{P}_{XY|Z}\, d\mathbb{P}_Z = \int \mathbb{P}_{X|Z}P_{Y|Z}\, d\mathbb{P}_Z$ when $X \not\perp\!\!\!\perp Y|Z$ requires a very particular setup which is probably rarely encountered in practice.

The aforementioned argument motivates us to also consider a different empirical estimate of the squared Hilbert-Schmidt norm of the partial cross covariance operator:

$$\mathcal{S}'_K = n\mathcal{T}_{\boldsymbol{xy\cdot z}} = \frac{1}{n}\mathrm{tr}(\widetilde{K}_{X|Z}\widetilde{K}_{Y|Z}), \tag{36}$$

where we have replaced $\ddot{X}$ with X. We can approximate the null distribution of $\mathcal{S}'_K$ by utilizing the strategies presented in Sections 3.3 and 3.4 of Zhang et al. [10]. Here, we utilize the hypotheses:

$$\begin{aligned} H_0 &: \|\Sigma_{XY\cdot Z}\|^2_{HS} = 0, \\ H_1 &: \|\Sigma_{XY\cdot Z}\|^2_{HS} > 0. \end{aligned} \tag{37}$$

We similarly consider a corresponding finite dimensional partial cross-covariance matrix:

$$\mathcal{S}' = n\|\widehat{\mathcal{C}}_{AB\cdot C}\|^2_F, \tag{38}$$

where we have replaced $\ddot{A}$ with A. The above statistic is a generalization of linear partial correlation, because we consider uncorrelatedness of the residuals of non-linear functional transformations after performing non-linear regression. The asymptotic distribution for $\mathcal{S}$ in Theorem 1 also holds for $\mathcal{S}'$, when we replace $\ddot{A}$ with A. Here, we use the hypotheses:

$$\begin{aligned} H_0 &: \|\mathcal{C}_{AB\cdot C}\|^2_F = 0, \\ H_1 &: \|\mathcal{C}_{AB\cdot C}\|^2_F > 0. \end{aligned} \tag{39}$$

In practice, the test which uses $\mathcal{S}'$, which we now call the Randomized conditional Correlation Test (RCoT), usually rivals or outperforms RCIT and KCIT, because (1) nearly all conditionally dependent variables encountered in practice are also conditionally correlated after at least one functional transformation, and (2) we can easily calibrate the null distribution of the test using $\mathcal{S}'$ even when Z has large cardinality. We will therefore find this test useful for replacing RCIT when we have large conditioning set sizes (≥ 4).

6.4 Time complexity

We now show that RCIT and RCoT have linear time complexity with respect to the sample size n. Let d denote the number of random Fourier features in C.

Proposition 6. *If we have $n > d$, then RCIT and RCoT have time complexity $\mathcal{O}(d^2 n)$.*

Proof. Wlog, we will prove the claim for RCIT (the proof for RCoT will follow analogously). Note that it suffices to show that all sub-procedures of RCIT have time complexity $\mathcal{O}(d^2 n)$ or $o(d^2 n)$.

The first step of RCIT computes the random Fourier features on the support of the process $\sqrt{2}\cos(W^T \cdot + B)$, $W \sim \mathbb{P}_W$, $B \sim \text{Uniform}([0, 2\pi])$. Let a_X denote the number of dimensions in X, and likewise for a_Y and a_Z. Let m, q and d denote the number of random Fourier features in $\ddot{A}, B$ and C, respectively. Note that computing the samples of $\ddot{A}$ requires $\mathcal{O}((a_X + a_Z)mn)$ time because W is a $(a_X + a_Z) \times m$ matrix. As a result, computing the samples of $\ddot{A}$ requires $\mathcal{O}(n)$ time with m, a_Z and a_X fixed. Similarly, computing the samples of B requires $\mathcal{O}(n)$ time with a_Y, q fixed and that of C requires $\mathcal{O}(dn)$ time with only a_Z fixed.

The second step of RCIT estimates $\widehat{E}(\ddot{A}|C)$ and $\widehat{E}(B|C)$ using linear ridge regressions. Recall that linear ridge regression scales $\mathcal{O}(d^2 n)$ in time, where d corresponds to the number of features (assuming $n > d$). Now RCIT requires $m + q$ linear ridge regressions in order to compute $\widehat{E}(\ddot{A}|C)$ and $\widehat{E}(B|C)$. We therefore conclude that estimating $\widehat{E}(\ddot{A}|C)$ and $\widehat{E}(B|C)$ can be done in $\mathcal{O}(d^2 n)$ time with m and q fixed.

Next, computing the covariance $\widehat{\Sigma}_{\ddot{A}B \cdot C}$ requires $\mathcal{O}(mqn)$ time. Finally, the time complexity of all of the methods used to approximate the null distribution do not depend on d or n once given $\widehat{\Sigma}_{\ddot{A}B \cdot C}$; we thus conclude that all of the null distribution approximation methods have time complexity $\mathcal{O}(1)$ when m and q are fixed.

We have shown that all sub-procedures of RCIT scale $\mathcal{O}(d^2 n)$ or $o(d^2 n)$. We therefore conclude that RCIT has time complexity $\mathcal{O}(d^2 n)$. □

The above proposition implies that RCIT and RCoT have time complexity $\mathcal{O}(n)$ because d is set to a fixed number regardfless of sample size. Recall that d is fixed because we have an exponential convergence rate with respect to d as highlighted previously in Proposition 4; in other words, a fixed d is reasonable so long as n does not become extremely large. In practice, we have found that setting $m = 5$, $q = 5$ and $p = 25$ works well for a variety of sample sizes and dimensions of Z as highlighted in the next section. The statement holds so long as we choose a very small regularization constant λ (e. g., 1E–10); note that this is different from the standard prediction regime, where we must carefully tune λ to prevent overfitting. We can utilize a small regularization constant in the proposed CI test setting because the entries of $\widehat{C}_{\ddot{A}B \cdot C}$ are never seen during training time.

7 Experiments

We carried out experiments to compare the empirical performance of the following tests:
- RCIT: uses $\mathcal{S}$ with the Lindsay-Pilla-Basak approximation,
- RCoT: uses $\mathcal{S}'$ with the Lindsay-Pilla-Basak approximation,
- KCIT: uses $\mathcal{S}_K$ with a simulated null by bootstrap.
- KCoT: uses $\mathcal{S}'_K$ with a simulated null by bootstrap.

We estimated the conditional expectations for $\mathcal{S}$ and $\mathcal{S}'$ using linear ridge regressions with random Fourier features. We also compared RCIT and RCoT against permutation tests and list the results in the Appendix A.

We present the results of KCoT in Appendix A.3 as well. Note that KCIT with the gamma approximation performs *slightly* faster than KCIT with bootstrap,[3] but the bootstrap results in a significantly better calibrated null distribution. We focus on large sample size (≥ 500) scenarios because we can just apply KCIT with bootstrap otherwise. We ran all experiments using the R programming language (Microsoft R Open) on a laptop with 2.60 GHz of CPU and 16 GB of RAM.

7.1 Hyperparameters

We used the same hyperparameters for RCIT and RCoT. Namely, we used the median Euclidean distance heuristic across the first 500 samples of $\ddot{X}, X, Y$ and Z for choosing the $\sigma_{\ddot{X}}, \sigma_X, \sigma_Y$, and σ_Z hyperparameters for the Gaussian kernels $k_\sigma(x,y) = \exp(-\|x-y\|^2/\sigma)$, respectively[4] [16, 40]. We also fixed the number of Fourier features for $\ddot{X}, X$ and Y to 5 and the number of Fourier features for Z to 25. We standardized all original and Fourier variables to mean zero unit variance in order to help ensure numerically stable computations. Finally, we set γ to 1E–10 in order to keep bias minimal. These parameters are reasonable because we designed both RCIT and RCoT for the purposes of causal discovery, where the variable set Z is small with a sparse underlying causal graph. We can therefore utilize a relatively small number of random Fourier features as compared to the sample size n. Authors who wish to apply RCIT or RCoT in the high dimensional scenario should consider utilizing more Fourier features for Z and choosing the lambda values carefully (e. g., through cross-validation or information criteria).

With KCIT, we set σ to the squared median Euclidean distance between (X, Y) using the first 500 samples times double the conditioning set size; the hyperparameters as described in the original paper, the hyperparameters in the author-provided MATLAB implementation and the hyperparameters of RCIT/RCoT all gave worse performance.

7.2 Type I error

We analyzed the Type I error rates of the three CI tests as a function of sample size and conditioning set size. We evaluated the algorithms using the Kolmogorov-Smirnov (KS) test statistic. Recall that the KS test uses the following statistic:

$$\mathcal{K} = \sup_{x \in \mathbb{R}} |\widehat{F}(x) - F(x)| = \|\widehat{F}_X - F_X\|_\infty, \tag{40}$$

where $\widehat{F}_X$ denotes the empirical CDF, and F_X some comparison CDF. If the sample comes from $\mathbb{P}_X$, then $\mathcal{K}$ converges to 0 almost surely as $n \to \infty$ by the Glivenko-Cantelli theorem.

Now a good CI test controls the Type I error rate at any α value, when we have a uniform sampling distribution of the p-values over $[0,1]$. Therefore, a good CI test should have a small KS statistic value, when we set F_X to the uniform distribution over $[0,1]$.

To compute the KS statistic values, we generated data from 1000 post non-linear models [10, 11]. We can describe each post non-linear model as follows: $X = g_1(Z + \varepsilon_1)$, $Y = g_2(Z + \varepsilon_2)$, where $Z, \varepsilon_1, \varepsilon_2$ have jointly independent standard Gaussian distributions, and g_1, g_2 denote smooth functions. We always chose g_1, g_2 uniformly from the following set of functions: $\{(\cdot), (\cdot)^2, (\cdot)^3, \tanh(\cdot), \exp(-\|\cdot\|_2)\}$. Thus, we have $X \perp\!\!\!\perp Y|Z$ in any case. Notice also that this situation is more general than the additive noise models proposed in Ramsey [41], where we have $X = g_1(Z) + \varepsilon_1$, $Y = g_2(Z) + \varepsilon_2$. The post non-linear models allow us to simulate heteroskedastic noise which is commonly encountered in real scenarios but not captured with additive noise models.

3 KCIT with the gamma approximation specifically completes 66.78 ms faster on average with an SEM of 4.29 ms (or 1.013 times faster with an SEM of 0.067 times) at 2000 samples over 500 of our experiments.

4 We also tried setting σ_Z to the median distance divided by 1.5, 2 or 3. However, these values gave progressively worse performance on average.

7.2.1 Sample size

We first assess the Type I error rate as a function of sample size. We used sample sizes of 500, 1000, 2000, 5000, ten thousand, one hundred thousand and one million. A good CI test should control the Type I error rate across all α values at any sample size. Figure 1a summarizes the KS statistic values for the three different CI tests. Observe that all tests have similar KS statistic values across different sample sizes. We conclude that all three tests perform comparably in controlling the Type I error rate with a single conditioning variable at different sample sizes.

The run time results however tell a markedly different story. Both RCIT and RCoT output a p-value much more quickly than KCIT at different sample sizes (Figure 1b). Moreover, KCIT ran out of memory at 5000 samples while RCIT and RCoT handled one million samples in a little over 6 seconds. RCIT and RCoT also completed more than two orders of magnitude faster than KCIT on average at a sample size of 2000 (Figure 1c). We conclude that RCIT and RCoT are more scalable than KCIT. Moreover, the experimental results agree with standard matrix complexity theory; RCIT and RCoT scale linearly with sample size (see Proposition 6), while KCIT scales strictly greater than quadratically with sample size.

7.2.2 Conditioning set size

CCD algorithms request p-values from CI tests using large conditioning set sizes. In fact, algorithms which do not assume causal sufficiency, such as FCI, often demand very large conditioning set sizes (> 5). We should however also realize that CCD algorithms search for *minimal* conditioning sets in order to establish ancestral relations. This means that we must focus on testing for cases where $X \not\perp\!\!\!\perp Y|Z$, but we have either $X \perp\!\!\!\perp Y|Z \cup A$ or $X \not\perp\!\!\!\perp Y|Z \cup A$, where $|A| = 1$.

We therefore evaluated the Type I error rates of the CI tests as a function of conditioning set size by fixing the sample size at 1000 and then adding 1 to 10 standard Gaussian variables into the conditioning set so that $X = g_1(\frac{1}{k}\sum_{j=1}^{k} Z_j + \epsilon_1)$, $Y = g_2(\frac{1}{k}\sum_{j=1}^{k} Z_j + \epsilon_2)$, $k = \{1, \ldots, 10\}$ in 1000 models. Note that this situation corresponds to 1 to 10 common causes.

Figure 1d summarizes the KS statistic values in the aforementioned scenario. We see that the KS statistic values for RCoT remain the smallest for nearly all conditioning set sizes, followed by RCIT and then KCIT. This implies that RCoT best approximates the null distribution out of the three CI tests. We also provide the histograms of the p-values across the 1000 post non-linear models at a conditioning set size of 10 for KCIT, RCIT, and RCoT in Figures 1e–1g. Notice that the histograms become progressively more similar to a uniform distribution. We conclude that RCoT controls its Type I error rate the best even with large conditioning set sizes while KCIT controls its rate the worst.

Now the run times of all three tests only increased very slightly with the conditioning set size (Figure 1h). However, both RCIT and RCoT still completed 40.91 times faster than KCIT on average (95 % confidence interval: ±0.44). These results agree with standard matrix complexity theory, as we expect all tests to scale linearly with dimensionality.

7.3 Power

We next evaluated test power (i. e., 1-(Type II error rate)) by computing the area under the power curve (AUPC). The AUPC corresponds to the area under the empirical CDF of the p-values returned by a CI test when the null does not hold. A CI test has higher power when its AUPC is closer to one. For example, observe that if a CI test always returns a p-value of 0 in the perfect case, then its AUPC corresponds to 1.

We examined the AUPC by adding the same small error $\varepsilon_b \sim \mathcal{N}(0, 1/16)$ to both X and Y in 1000 post non-linear models as follows: $X = g_1(\varepsilon_b + \varepsilon_1)$, $Y = g_2(\varepsilon_b + \varepsilon_2)$, $Z \sim \mathcal{N}(0, 1)$. Here, we do not allow the CI tests to condition on ε_b, so we always have $X \not\perp\!\!\!\perp Y|Z$; this situation therefore corresponds to a hidden common cause.

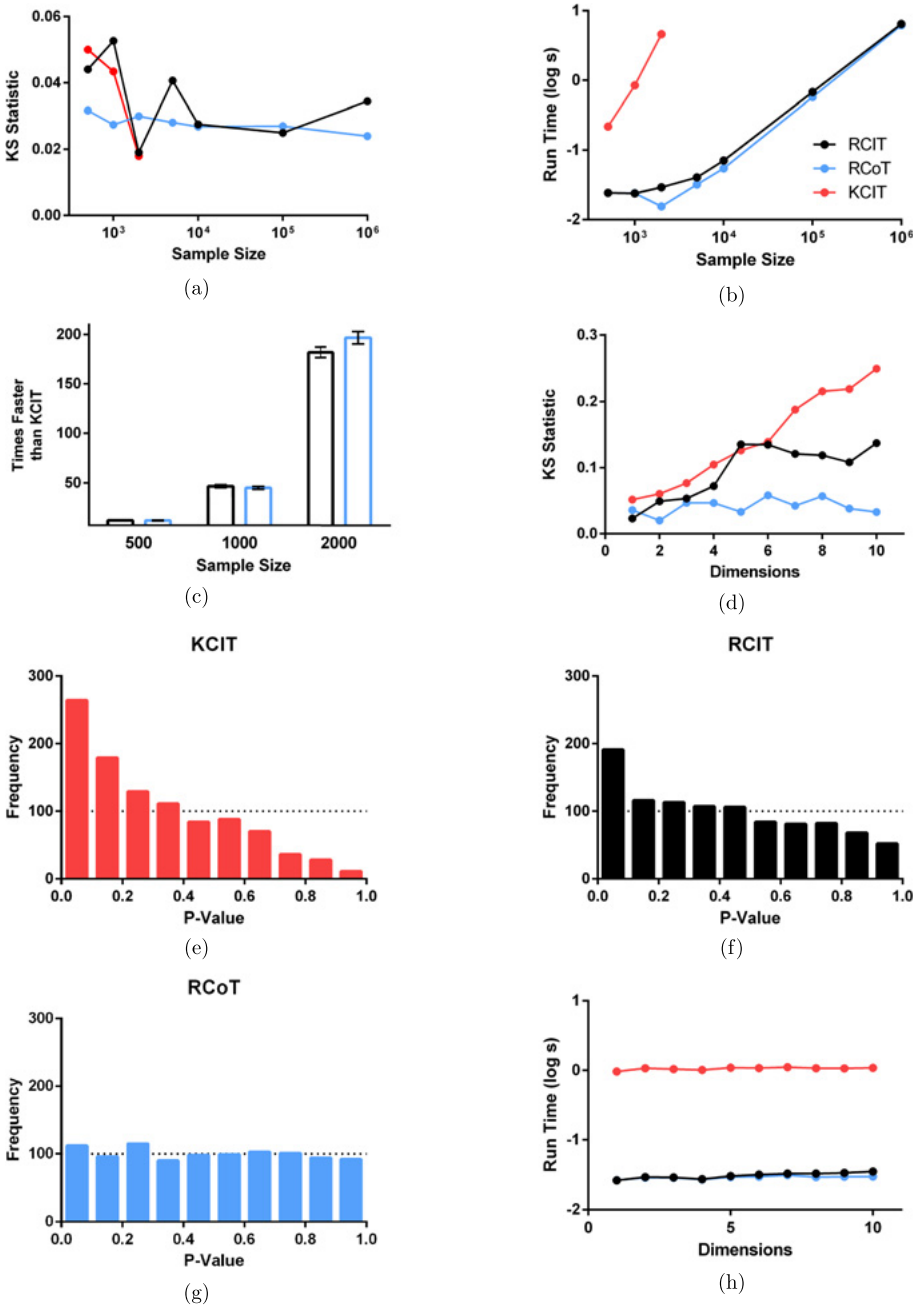

Figure 1: Experimental results of RCIT, RCoT and KCIT when conditional independence holds. (a) All tests have comparable KS statistic values as a function of sample size. (b) However, both RCIT and RCoT complete much faster than KCIT. (c) The relative difference in speed between RCIT vs. KCIT and RCoT vs. KCIT grows with increasing sample size. (d) RCoT maintains the lowest KS statistic value with increases in dimensionality. (e–g) Histograms with a conditioning set size of 10 show that KCIT, RCIT and RCoT obtain progressively more uniform null distributions. (h) Run times of all three tests scale linearly with dimensionality of the conditioning set.

7.3.1 Sample size

We first examine power as a function of sample size. We again tested sample sizes of 500, 1000, 2000, 5000, ten thousand, one hundred thousand, and one million. We have summarized the results in Figure 2a. Both RCIT and RCoT have comparable AUPC values to KCIT with sample sizes of 500, 1000 and 2000. At larger sam-

ple sizes, KCIT again did not scale due to insufficient memory, but the AUPC of both RCIT and RCoT continued to increase at similar values. We conclude that all three tests have similar power.

The run time results mimic those of Section 7.2.1; RCIT and RCoT completed orders of magnitude faster than KCIT (Figures 2b and 2c).

7.3.2 Conditioning set size

We next examined power as a function of conditioning set size. To do this, we fixed the sample size at 1000 and set $Z = (Z_1, \ldots, Z_k)$ with $Z \sim \mathcal{N}(0, I_k)$, $k = \{1, \ldots, 10\}$ in the 1000 post non-linear models. We therefore examined how well the CI tests reject the null under an increasing conditioning set size with uninformative variables. A good CI test should either (1) maintain its power or, more realistically, (2) suffer a graceful decline in power with an increasing conditioning set size because none of the variables in the conditioning set are informative for rendering conditional independence by design.

We have summarized the results in Figure 2d. Notice that all tests have comparable AUPC values with small conditioning set sizes (between 1 and 3), but the AUPC value of KCIT gradually increases with increasing conditioning set sizes; the AUPC value should not increase under the current setup with a well-calibrated null because the extra variables are uninformative. To determine the cause of the unexpected increase in power, we permuted the values of X in each run in order to assess the calibration of the null distribution. Figure 2f summarizes the results, where we can see that only KCIT's KS statistic grows with an increasing conditioning set size. We can therefore claim that the increasing AUPC value of KCIT holds because of a badly calibrated null distribution with larger conditioning set sizes. We conclude that both RCIT and RCoT maintain steady power under an increasing conditioning set size while KCIT does not.

The run times in Figures 2e and 2g again mimic those in Section 7.2.2 with RCIT and RCoT completing in a much shorter time frame than KCIT.

7.4 Causal structure discovery

We next examine the accuracy of graphical structures as recovered by PC [1], FCI [42] and RFCI [43] when run using RCIT, RCoT or KCIT.

We used the following procedure in Colombo et al. [43] to generate 250 different Gaussian DAGs with an expected neighborhood size $\mathbb{E}(N) = 2$ and $v = 20$ vertices. First, we generated a random adjacency matrix $\mathcal{A}$ with independent realizations of Bernoulli$(\mathbb{E}(N)/(v-1))$ random variables in the lower triangle of the matrix and zeroes in the remaining entries. Next, we replaced the ones in $\mathcal{A}$ by independent realizations of a Uniform$([-1, -0.1] \cup [0.1, 1])$ random variable. We interpret a nonzero entry $\mathcal{A}_{ij}$ as an edge from X_i to X_j with coefficient $\mathcal{A}_{ij}$ in the following linear model:

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_i &= \sum_{r=1}^{v-1} \mathcal{A}_{ir} X_r + \varepsilon_i. \end{aligned} \tag{41}$$

for $i = 2, \ldots, v$ where $\varepsilon_1, \ldots, \varepsilon_v$ are mutually independent standard Gaussian random variables. The variables $\{X_1, \ldots, X_v\} = \boldsymbol{X}$ then have a multivariate Gaussian distribution with mean 0 and covariance matrix $\Sigma = (I_v - \mathcal{A})^{-1}(I_v - \mathcal{A})^{-T}$, where I_v is the $v \times v$ identity matrix. To introduce non-linearities, we passed each variable in $\boldsymbol{X}$ through a non-linear function g again chosen uniformly from the set $\{(\cdot), (\cdot)^2, (\cdot)^3, \tanh(\cdot), \exp(-\|\cdot\|_2)\}$.

For FCI and RFCI, we introduced latent and selection variables using the following procedure. For each DAG, we first randomly selected a set of 0–3 latent common causes L. From the set $X \setminus L$, we then selected a set of 0–3 colliders as selection variables S. For each selection variable in S, we subsequently eliminated the bottom q percentile of samples, where we drew q according to independent realizations of a Uniform$([0.1, 0.5])$ random variable. We finally eliminated all of the latent variables from the dataset.

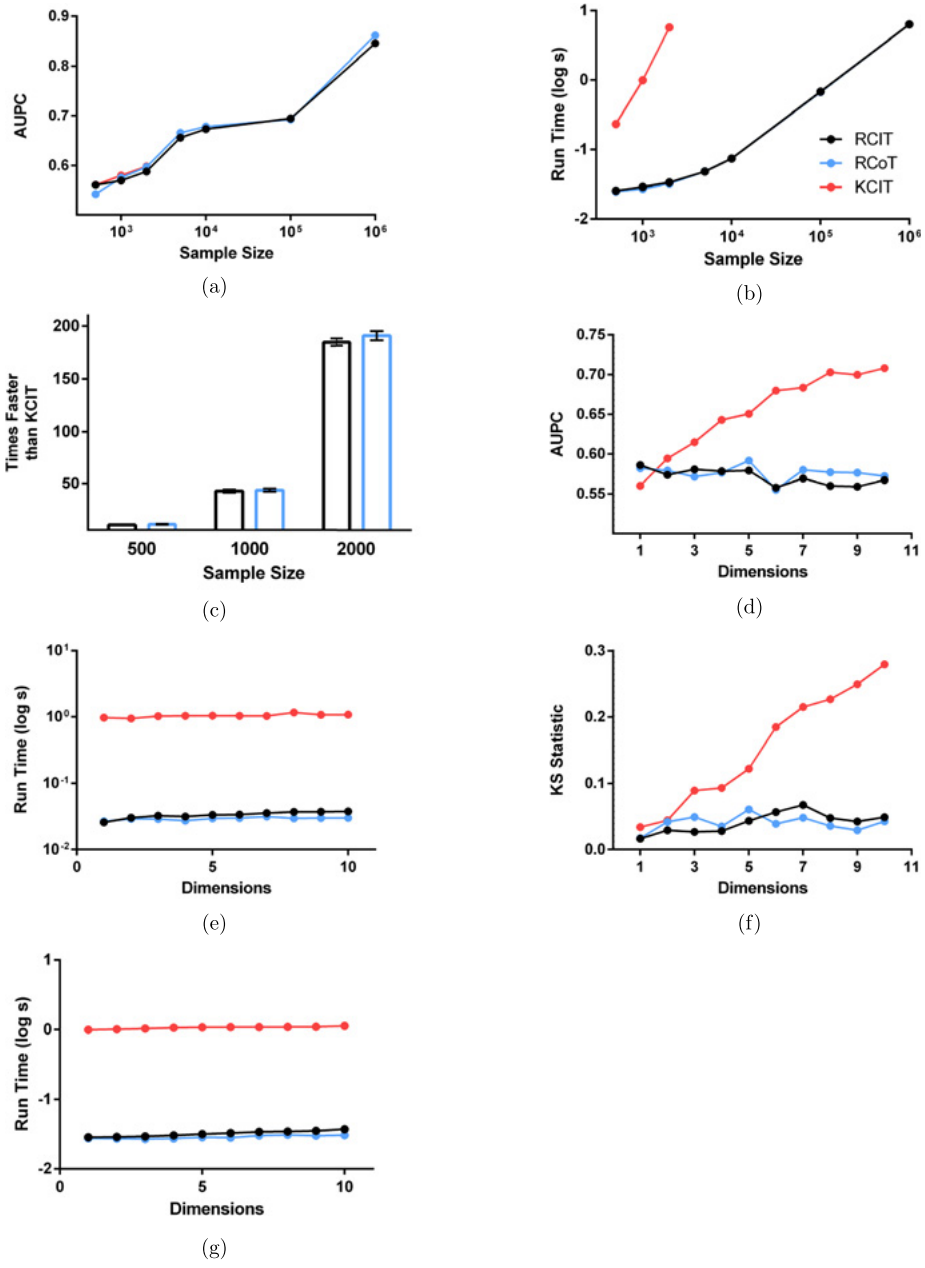

Figure 2: Experimental results with RCIT, RCoT and KCIT as a function of sample size and conditioning set size when conditional dependence holds. (a) All tests have comparable AUPC values as a function of sample size with a conditioning set size of one. (b–c) Both RCIT and RCoT again complete much faster than KCIT. (d) KCIT's AUPC value unexpectedly increases with the dimensionality of the conditioning set. Associated run times for (d) in (e). (f) The cause of KCIT's AUPC increase lies in a badly calibrated null distribution; here we see that only KCIT's KS statistic value increases under the null. Associated run times for (f) in (g).

We ultimately created 250 different 500 sample datasets for PC, FCI and RFCI. We then ran the sample versions of PC, FCI and RFCI using RCIT, RCoT, KCIT and Fisher's z-test (FZT) at $\alpha = 0.05$. We also obtained the oracle graphs by running the oracle versions of PC, FCI and RFCI using the ground truth.

We have summarized the results as structural Hamming distances (SHDs) from the oracle graphs in Figure 3a. PC run with RCIT and PC run with RCoT both outperformed PC run with KCIT by a large margin according to paired t-tests (PC RCIT vs. KCIT, $t = -14.76$, $p < 2.2E{-}16$; PC RCoT vs. KCIT, $t = -12.87$, $p < 2.2E{-}16$). We

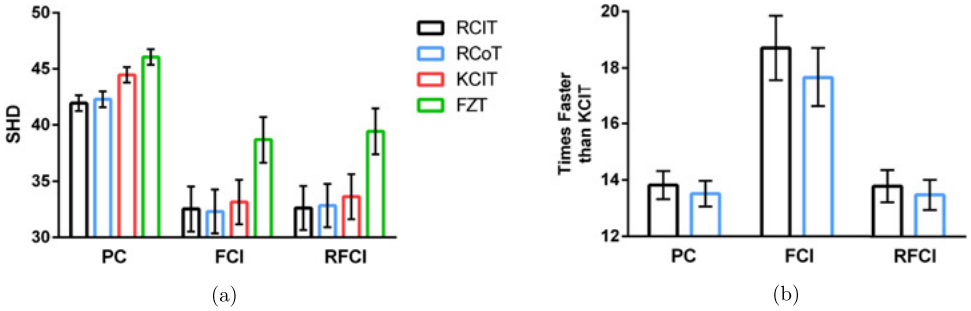

Figure 3: Results of CCD algorithms as evaluated by mean (a) SHD and (b) run times. The CCD algorithms run with KCIT perform comparably (or even slightly worse) to those run with RCIT and RCoT in (a). Run times in (b) show that the CCD algorithms run with RCIT and RCoT complete at least 13 times faster on average than those with KCIT. Error bars denote 95 % confidence intervals of the mean.

found similar results with FCI and RFCI, although by only a small margin; 3 of the 4 comparisons fell below the Bonferonni corrected threshold of 0.05/6 and the other comparison fell below the uncorrected threshold of 0.05 (FCI RCIT vs. KCIT, $t = -2.00$, $p = 0.047$; FCI RCoT vs. KCIT $t = -2.96$, $p = 0.0034$; RFCI RCIT vs. KCIT, $t = -3.56$, $p = 4.5\text{E}{-4}$; RFCI RCoT vs. KCIT, $t = -2.80$, $p = 0.0055$). All algorithms with any of the kernel-based tests outperformed the same algorithms with FZT by a large margin ($p < 7\text{E}{-14}$ in all cases). Finally, the run time results in Figure 3b show that the CCD algorithms run with RCIT and RCoT complete at least 13 times faster on average than those run with KCIT. We conclude that both RCIT and RCoT help CCD algorithms at least match the performance of the same algorithms run with KCIT, but RCIT and RCoT do so within a much shorter time frame than KCIT.

7.5 Real data

We finally ran PC, FCI and RFCI using RCIT, RCoT, KCIT and FZT at $\alpha = 0.05$ on a publicly available longitudinal dataset from the Cognition and Aging USA (CogUSA) study [44], where scientists measured the cognition of men and women above 50 years of age. The dataset contains 815 samples, 18 variables and two waves (thus $18/2 = 9$ variables in each wave) separated by two years after some data cleaning.[5] Note that we do not have access to a gold standard solution set in this case. However, we can utilize the time information in the dataset to detect false positive ancestral relations directed backwards in time.

We ran the CCD algorithms on 30 bootstrapped datasets. Results are summarized in Figure 4. Comparisons with PC did not reach the Bonferonni level among the kernel-based tests, although PC run with either RCIT or RCoT yielded fewer false positive ancestral relations on average than PC run with KCIT near an uncorrected level of 0.05 (PC RCIT vs. KCIT, $t = -2.76$, $p = 9.85\text{E}{-3}$; PC RCoT vs. KCIT, $t = -1.99$, $p = 0.056$). However, FCI and RFCI run with either RCIT or RCoT performed better than those run with KCIT at a Bonferroni corrected level of 0.05/6 (FCI RCIT vs. KCIT, $t = -29.57$, $p < 2.2\text{E}{-16}$; FCI RCoT vs. KCIT, $t = -17.41$, $p < 2.2\text{E}{-16}$; RFCI RCIT vs. KCIT, $t = -6.50$, $p = 4.13\text{E}{-7}$; RFCI RCoT vs. KCIT, $t = -7.39$, $p = 3.85\text{E}{-8}$). The CCD algorithms run with FZT also gave inconsistent results; PC run with FZT performed the best on average, but FCI and RFCI run with FZT also performed second from the worst. Here, we should trust the outputs of FCI and RFCI more strongly than those of PC, since both FCI and RFCI allow latent common causes and selection bias which often exist in real data. Next, CCD algorithms run with RCIT performed comparably to those run with RCoT (PC RCIT vs. RCoT, $t = -1.05$, $p = 0.301$; FCI RCIT vs. RCoT, $t = -1.54$, $p = 0.134$; RFCI RCIT vs. RCoT, $t = -0.89$, $p = 0.380$). We finally report that the CCD algorithms run with RCIT and RCoT complete at

5 We specifically removed redundant variables with deterministic relations, variables with more than 1000 missing values, and then samples with missing values in any of the remaining variables.

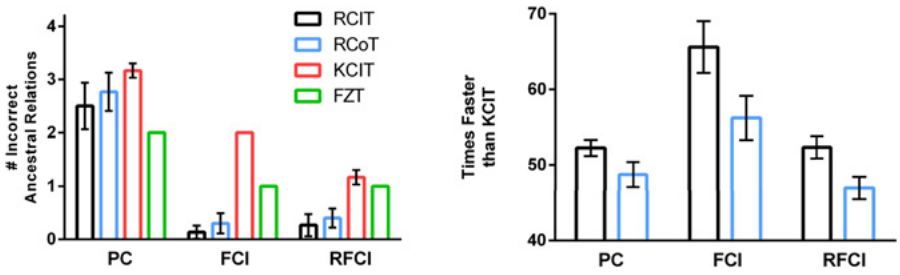

Figure 4: Results of CCD algorithms as evaluated on real longitudinal data. Part (a) displays mean counts of the number of ancestral relations directed backwards time. We do not display 95 % confidence intervals when we computed a standard error of zero. Part (b) summarizes the mean run times.

least 40 times faster on average than those run with KCIT (Figure 4b). We conclude that CCD algorithms run with either RCIT or RCoT perform at least as well as those run with KCIT on this real dataset but with large reductions run time.

8 Conclusion

We developed two statistical tests called RCIT and RCoT for fast non-parametric CI testing. Both RCIT and RCoT approximate KCIT by sampling Fourier features. Moreover, the proposed tests return p-values orders of magnitude faster than KCIT in the large sample size setting. RCoT in particular also has a better calibrated null distribution than KCIT especially with larger conditioning set sizes. In causal graph discovery, RCIT and RCoT help CCD algorithms recover graphical structures at least as accurately as KCIT but, most importantly, also allow the algorithms to complete in a much shorter time frame. We believe that the speedups provided by RCIT and RCoT will make non-parametric causal discovery more accessible to scientists who wish to apply CCD algorithms to their datasets.

Note that RCIT and RCoT may estimate the null distribution more accurately than KCIT for multiple reasons. First, RCIT and RCoT both assess for non-vanishing covariances across a smaller set of functions than KCIT. This in turn allows the two tests to more easily estimate the null distribution than KCIT because KCIT must deal with all of the functions in the associated RKHS. Second, both RCIT and RCoT utilize more advanced methods of estimating of the null distribution than KCIT. RCIT and RCoT more specifically utilize the Lindsay-Pilla-Basak method as explained in Section 6.2 as opposed to matching the first two moments of a gamma distribution. Third, RCoT in particular utilizes the variable set X rather than $\ddot{X}$ which allows for lower dimensional inferences as explained in Section 6.3. RCIT and RCoT thus take advantage of new technologies and additional structure inherent within real data in order to achieve better control of the Type I error rate.

Acknowledgment: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding: Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge initiative. The research was also supported by the National Library of Medicine of the National Institutes of Health under award numbers T15LM007059 and R01LM012095.

Appendix A

A.1 CLT for sample covariance

We will prove the central limit theorem (CLT) for the sample covariance matrix. We first have the following sample covariance matrices with known and unknown expectation vector, respectively:

$$\begin{aligned}\ddot{\mathcal{C}} &= \frac{1}{n}\sum_{i=1}^{n}[X_i - \mathbb{E}(X)][X_i - \mathbb{E}(X)]^T,\\ \widehat{\mathcal{C}} &= \frac{1}{n-1}\sum_{i=1}^{n}[X_i - \widehat{\mathbb{E}}(X)][X_i - \widehat{\mathbb{E}}(X)]^T.\end{aligned} \tag{42}$$

Now observe that we may write:

$$\begin{aligned}&(n-1)\widehat{\mathcal{C}}\\ =&\sum_{i=1}^{n}[X_i - \mathbb{E}(X) - (\widehat{\mathbb{E}}(X) - \mathbb{E}(X))][X_i - \mathbb{E}(X) - (\widehat{\mathbb{E}}(X) - \mathbb{E}(X))]^T\\ =&\sum_{i=1}^{n}(X_i - \mathbb{E}(X))(X_i - \mathbb{E}(X))^T + n(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T\\ &- 2(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))\sum_{i=1}^{n}(X_i - \mathbb{E}(X))^T\\ =&\, n\ddot{\mathcal{C}} - n(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T\end{aligned} \tag{43}$$

It follows that:

$$\begin{aligned}&\sqrt{n}(\widehat{\mathcal{C}} - \mathcal{C})\\ =&\sqrt{n}(\frac{n-1}{n-1}\widehat{\mathcal{C}} - \mathcal{C})\\ =&\sqrt{n}(\frac{n}{n-1}\ddot{\mathcal{C}} - \frac{n}{n-1}(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T - \mathcal{C})\\ =&\frac{n\sqrt{n}}{n-1}\ddot{\mathcal{C}} - \frac{n\sqrt{n}}{n-1}(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T - \sqrt{n}\mathcal{C}\\ =&\frac{n\sqrt{n}}{n-1}\ddot{\mathcal{C}} - \frac{n\sqrt{n}}{n-1}(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T - \frac{n-1}{n-1}\sqrt{n}\mathcal{C}\\ =&\frac{n\sqrt{n}}{n-1}(\ddot{\mathcal{C}} - \mathcal{C}) - \frac{n\sqrt{n}}{n-1}(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))(\widehat{\mathbb{E}}(X) - \mathbb{E}(X))^T + \frac{\sqrt{n}}{n-1}\mathcal{C}\end{aligned} \tag{44}$$

We are now ready to state the result:

Lemma 1. *Let $X_1, \ldots, X_n$ refer to a sequence of i. i. d. random k-vectors. Denote the expectation vector and covariance matrix of X_1 as μ_1 and $\mathcal{C}_1$, respectively. Assume that $\check{\mathcal{C}}_1 = \mathrm{Cov}[v_u((X_1 - \mu_1)(X_1 - \mu_1)^T)]$ is positive definite, where $v_u(M)$ denotes the vectorization of the upper triangular portion of a real symmetric matrix M. Then, we have:*

$$\sqrt{n}(v_u(\widehat{\mathcal{C}}) - v_u(\mathcal{C}_1)) \xrightarrow{d} \mathcal{N}(0, \check{\mathcal{C}}_1). \tag{45}$$

Proof. Consider the quantity $a^T[\sqrt{n}(v_u(\widehat{\mathcal{C}}) - v_u(\mathcal{C}_1))] = \sqrt{n}(a^T v_u(\widehat{\mathcal{C}}) - a^T v_u(\mathcal{C}_1))$ where $a \in \mathbb{R}^{k(k+1)/2} \setminus \{0\}$. Note that $a^T v_u[(X_1 - \mu_1)(X_1 - \mu_1)^T], \ldots, a^T v_u[(X_n - \mu_1)(X_n - \mu_1)^T]$ is a sequence of i. i. d. random variables with expectation $a^T v_u(\mathcal{C}_1)$ and variance $a^T \check{\mathcal{C}}_1 a$. Moreover observe that $\check{\mathcal{C}}_1 < \infty$ because $\check{\mathcal{C}}_1$ is positive definite. We can therefore apply the univariate central limit theorem to conclude that:

$$\sqrt{n}(a^T v_u(\ddot{\mathcal{C}}_1) - a^T v_u(\mathcal{C}_1)) \xrightarrow{d} \mathcal{N}(0, a^T \check{\mathcal{C}}_1 a), \tag{46}$$

where $\ddot{C}_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)(X_i - \mu_1)^T$. We would however like to claim that:

$$\sqrt{n}(a^T v_u(\widehat{C}) - a^T v_u(C_1)) \xrightarrow{d} \mathcal{N}(0, a^T \check{C}_1 a). \tag{47}$$

In order to prove this, we use 44 and set:

$$\sqrt{n}(a^T v_u(\widehat{C}) - a^T v_u(C_1)) = a^T A_n + a^T B_n, \tag{48}$$

where we have:

$$\begin{aligned} A_n &= \frac{n}{n-1}\sqrt{n}(v_u(\ddot{C}_1) - v_u(C)), \\ B_n &= \frac{\sqrt{n}}{n-1}v_u(C_1) - \frac{n\sqrt{n}}{n-1}v_u[(\widehat{\mathbb{E}}(X) - \mu_1)(\widehat{\mathbb{E}}(X) - \mu_1)^T]. \end{aligned} \tag{49}$$

We already know from 46 that:

$$\sqrt{n}(a^T v_u(\ddot{C}_1) - a^T v_u(C_1)) \xrightarrow{d} \mathcal{N}(0, a^T \check{C}_1 a). \tag{50}$$

Therefore, so does $a^T A_n$ by Slutsky's lemma, when we view the sequence of constants $\frac{n}{n-1}$ as a sequence of random variables. For $a^T B_n$, we know that:

$$\sqrt{n}(a^T \widehat{\mathbb{E}}(X) - a^T \mu_1) \xrightarrow{d} \mathcal{N}(0, a^T C_1 a), \tag{51}$$

by viewing $a^T X_1, \ldots, a^T X_n$ as a sequence of random variables, noting that $E(X_1 X_1^T) < \infty$ because $\check{C}_1$ is positive definite and then applying the univariate central limit theorem. We thus have $a^T B_n \xrightarrow{p} 0$. We may then invoke Slutsky's lemma again for $a^T A_n + a^T B_n$ and claim that:

$$\sqrt{n}(a^T v_u(\widehat{C}) - a^T v_u(C_1)) \xrightarrow{d} \mathcal{N}(0, a^T \check{C}_1 a). \tag{52}$$

We conclude the lemma by invoking the Cramer-Wold device. □

A.2 Kernel conditional correlation test (KCoT) results

We compared KCoT against RCIT, RCoT and KCIT. We report the results in Figure 5. All tests perform comparably as a function of sample size (Figures 5a and 5c). However, KCoT performs better than KCIT and underperforms RCoT and RCIT as a function of conditioning set size. In particular, RCIT and RCoT obtain smaller KS-statistic values as a function of the conditioning set size (Figure 5b). KCIT and KCoT also obtain larger AUPC values with an increasing conditioning set size because they fail to maintain a uniform distribution under the null (Figure 5d; we again permuted the values of X for Figure 5e). We therefore conclude that RCoT and RCIT control their Type I error rates better than KCIT and KCoT even with large conditioning set sizes while maintaining power.

A.3 Comparisons against permutation

We also compared RCIT and RCoT against permutation CI tests. Here, we estimated the null distribution of $\mathcal{S}$ and $\mathcal{S}'$ with permutations and call the resultant CI tests S-Perm and S′-Perm, respectively. The permutation tests specifically involve permuting the residuals of the random Fourier features of Y one thousand times in order to estimate the null distribution. We have summarized the Type I error rate results in Figure 6 and the AUPC results in Figure 7. We ran the sample size experiments up to only ten thousand samples due to the long run times of S-Perm and S′-Perm. We found that RCIT and RCoT perform similarly to the permutation tests but with significantly reduced average run time.

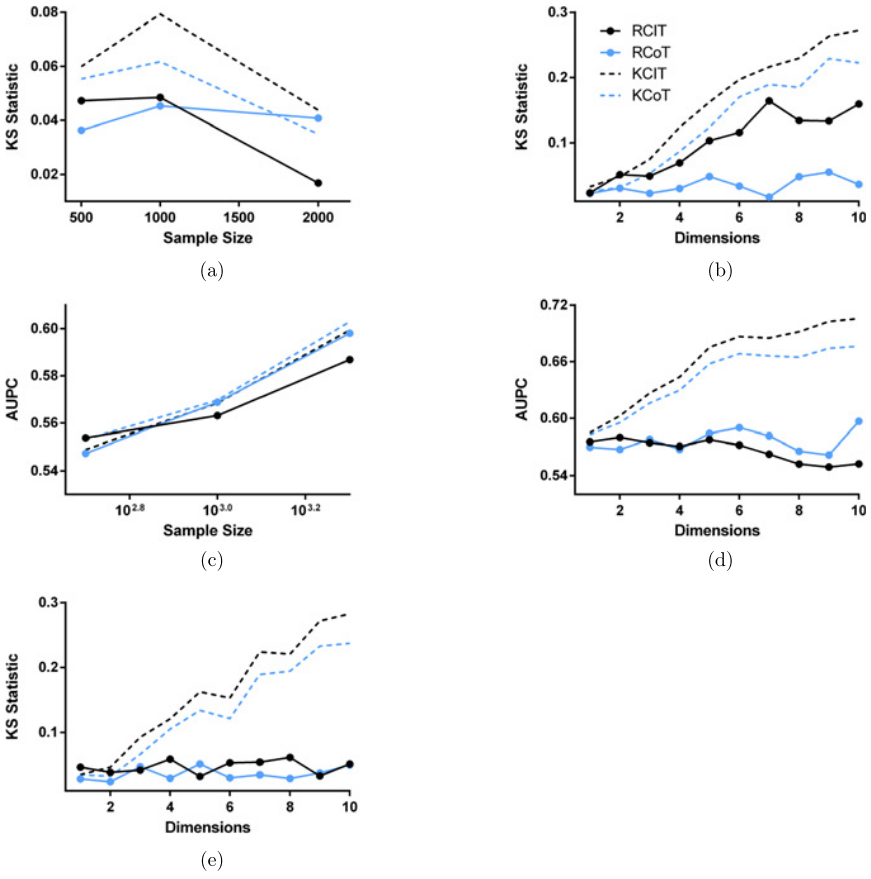

Figure 5: Results comparing against KCoT. Sub-figures (a) and (b) correspond to the KS-statistic values as a function of sample size and dimensions, respectively. Notice that RCIT and RCoT have progressively smaller KS-statistic values than both KCIT and KCoT with increasing dimensions. Next, sub-figures (c) and (d) correspond to the AUPC values also as a function of sample size and dimensions, respectively. KCIT and KCoT claim the largest AUPC values because both tests fail to control the Type I error rate well, as summarized by the large KS-statistic values in sub-figure (e) obtained after permuting the values of X.

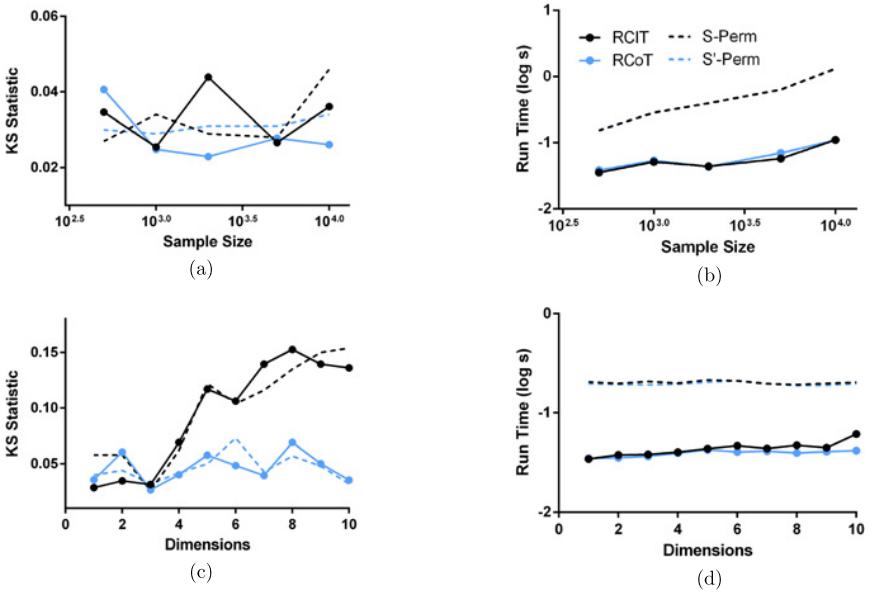

Figure 6: Results of the same KS statistic experiments as in Section 7.2 except comparing RCIT and RCoT against S-Perm and S′-Perm. Subfigures (a) and (b) again vary as a function of sample size, while subfigures (c) and (d) vary as a function of conditioning set size. RCIT and RCoT are faster than the permutation tests but yield comparable average KS statistic values.

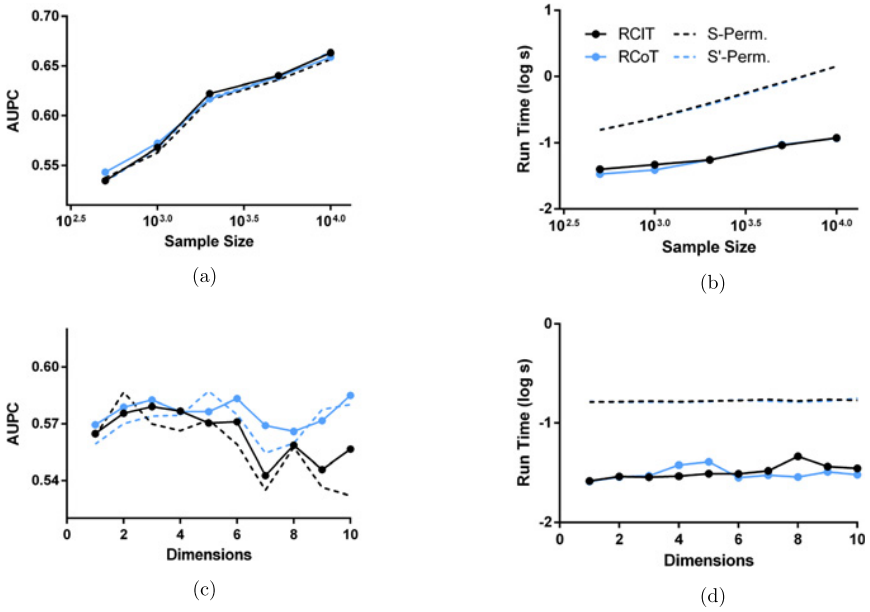

Figure 7: Results of the same AUPC experiments as in Section 7.3 except comparing RCIT and RCoT against S-Perm and S′-Perm. Subfigures (a) and (b) again vary as a function of sample size, while subfigures (c) and (d) vary as a function of conditioning set size. RCIT and RCoT are much faster than the permutation tests but yield comparable accuracy.

References

1. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd ed. MIT press; 2000.
2. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10(4):507–21. https://doi.org/10.2307/2331838. ISSN 00063444.
3. Fisher RA. On the probable error of a coefficient of correlation deduced from a small sample. Metron. 1921;1:3–32.
4. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag Ser 5. 1900;50:157–75.
5. Tsamardinos I, Borboudakis G. Permutation testing improves bayesian network learning. Berlin, Heidelberg: Springer; 2010. p. 322–37. http://dx.doi.org/10.1007/978-3-642-15939-8_21. ISBN 978-3-642-15939-8.
6. Margaritis D. Distribution-free learning of bayesian network structure in continuous domains. In: Proceedings, the twentieth national conference on artificial intelligence and the seventeenth innovative applications of artificial intelligence conference. July 9–13, 2005, Pittsburgh, Pennsylvania, USA. 2005. p. 825–30. http://www.aaai.org/Library/AAAI/2005/aaai05-130.php.
7. Huang T-M. Testing conditional independence using maximal nonlinear conditional correlation. Ann Stat. 2010;38(4):2047–91.
8. Su L, White H. A consistent characteristic function-based test for conditional independence. J Econom. 2007;141(2):807–34. https://doi.org/10.1016/j.jeconom.2006.11.006. http://www.sciencedirect.com/science/article/B6VC0-4MT59DD-4/2/267e7fc8dd979b6148fc4123998e94ee. ISSN 0304-4076.
9. Su L, White H. A nonparametric hellinger metric test for conditional independence. Econom Theory. 2008;24(4):829–64. http://www.jstor.org/stable/20142523. ISSN 02664666, 14694360.
10. Zhang K, Peters J, Janzing D, Schlkopf B. Kernel-based conditional independence test and application in causal discovery. In: Uncertainty in artificial intelligence. AUAI Press; 2011. p. 804–13. http://dblp.uni-trier.de/db/conf/uai/uai2011.html#ZhangPJS11. ISBN 978-0-9749039-7-2.
11. Doran G, Muandet K, Zhang K, Schölkopf B. A permutation-based kernel conditional independence test. In: Proceedings of the 30th conference on uncertainty in artificial intelligence (UAI2014). Oregon: AUAI Press Corvallis; 2014. p. 132–41.
12. Lopez-Paz D, Hennig P, Schölkopf B. The randomized dependence coefficient. In: Advances in neural information processing systems 26. 2013. p. 1–9.
13. Zhang Q, Filippi S, Gretton A, Sejdinovic D. Large-scale kernel methods for independence testing. Stat Comput. 2017; 1–18. https://doi.org/10.1007/s11222-016-9721-7. ISSN 1573-1375.
14. Rahimi A, Recht B. Random features for large-scale kernel machines. In: Neural information processing systems. 2007.

15. Sutherland DJ, Schneider JG. On the error of random fourier features. In: Meila M, Heskes T, editors. UAI. AUAI Press; 2015. p. 862–71. http://dblp.uni-trier.de/db/conf/uai/uai2015.html#SutherlandS15. ISBN 978-0-9966431-0-8.
16. Lopez-Paz D, Sra S, Smola A, Ghahramani Z, Schölkopf B. Randomized nonlinear component analysis. In: Proceedings of the 31st international conference on machine learning, W&CP 32 (1). JMLR; 2014. p. 1359–67.
17. Rahimi A, Recht B. Uniform approximation of functions with random bases. IEEE; 2008. https://doi.org/10.1109/ALLERTON.2008.4797607.
18. Gretton A, Bousquet O, Smola A, Schlkopf B. Measuring statistical dependence with hilbert-schmidt norms. In: Proceedings in algorithmic learning theory. Springer-Verlag; 2005. p. 63–77.
19. Berlinet A, Thomas-Agnan C. Reproducing kernel Hilbert spaces in probability and statistics. US: Springer; 2003. https://books.google.com/books?id=v79sBNG34coC. ISBN 9781402076794.
20. Hein M, Bousquet O. Kernels, associated structures and generalizations. Technical Report 127. Tübingen, Germany: Max Planck Institute for Biological Cybernetics; 2004.
21. Murphy G. C*-algebras and operator theory. Academic Press; 1990. https://books.google.com/books?id=emNvQgAACAAJ. ISBN 9780125113601.
22. Fukumizu K, Bach FR, Jordan MI. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. J Mach Learn Res. 2004;5:73–99. http://dl.acm.org/citation.cfm?id=1005332.1005335. ISSN 1532-4435.
23. Fukumizu K, Gretton A, Sun X, Schölkopf B. Kernel measures of conditional dependence. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran; 2008. p. 489–96. Max-Planck-Gesellschaft. http://papers.nips.cc/paper/3340-kernel-measures-of-conditional-dependence-supplemental.zip.
24. Daudin JJ. Partial association measures and an application to qualitative regression. 1980;67(3):581–590. https://doi.org/10.1093/biomet/67.3.581. https://doi.org/10.2307/2335127. http://www.jstor.org/stable/2335127. ISSN 0006-3444 (print), 1464-3510 (electronic).
25. Coppersmith D, Winograd S. Matrix multiplication via arithmetic progressions. J Symb Comput. 1990;9(3):251–80. https://doi.org/10.1016/S0747-7171(08)80013-2. ISSN 0747-7171.
26. Micchelli CA, Xu Y, Zhang H. Universal kernels. J Mach Learn Res. 2006;6:2651–67. http://dblp.uni-trier.de/db/journals/jmlr/jmlr7.html#MicchelliXZ06.
27. Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48(3/4):419–26. https://doi.org/10.2307/2332763.
28. Solomon H, Stephens MA. Distribution of a sum of weighted chi-square variables. J Am Stat Assoc. 1977;72:881–5.
29. Johnson NL, Kotz S, Balakrishnan N. Continuous multivariate distributions. 3rd ed. Wiley; 2002.
30. Welch BL. The significance of the difference between two means when the population variances are unequal. Biometrika. 1938;29(3–4):350–62. https://doi.org/10.1093/biomet/29.3-4.350.
31. Satterthwaite F. An approximate distribution of estimates of variance components. Biom Bull. 1946;2(6):110–4.
32. Fairfield-Smith H. The problem of comparing the results of two experiments with unequal errors. J Counc Sci Ind Res. 1936;9:211–2.
33. Hall P. Chi squared approximations to the distribution of a sum of independent random variables. Ann Probab. 1983;11(4):1028–36. https://doi.org/10.1214/aop/1176993451. 11.
34. Buckley MJ, Eagleson GK. An approximation to the distribution of quadratic forms in normal random variables. Aust N Z J Stat. 1988;30(1):150–9.
35. Wood ATA. An f approximation to the distribution of a linear combination of chi-squared variables. Commun Stat, Simul Comput. 1989;18:1439–56.
36. Lindsay B, Pilla R, Basak P. Moment-based approximations of distributions using mixtures: Theory and applications. Ann Inst Stat Math. 2000;52(2):215–30. http://EconPapers.repec.org/RePEc:spr:aistmt:v:52:y:2000:i:2:p:215-230.
37. Bodenham D, Adams N. A comparison of efficient approximations for a weighted sum of chi-squared random variables. Stat Comput. 2016;26:917–28. https://doi.org/10.1007/s11222-015-9583-4.
38. Bodenham D. momentchi2. 2015. http://cran.r-project.org/web/packages/momentchi2/.
39. Uspensky JVJV. Introduction to mathematical probability. 1st ed. New York, London: McGraw-Hill; 1937. “Problems for solution” with answers at end of each chapter.
40. Gretton A, Fukumizu K, Teo C, Song L, Schölkopf B, Smola A. A kernel statistical test of independence. In: Advances in neural information processing systems 20. Red Hook, NY, USA: Curran; 2008. p. 585–92. Max-Planck-Gesellschaft.
41. Ramsey JD. A scalable conditional independence test for nonlinear, non-gaussian data. CoRR. 2014;abs/1401.5031. http://arxiv.org/abs/1401.5031.
42. Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif Intell. 2008;172(16–17):1873–96. https://doi.org/10.1016/j.artint.2008.08.001. ISSN 0004-3702.
43. Colombo D, Maathius M, Kalisch M, Richardson T. Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann Stat. 2012;40(1):294–321. 10.1214/11-AOS940. http://projecteuclid.org/euclid.aos/1333567191.
44. McArdle J, Rodgers W, Willis R. Cognition and aging in the usa (cogusa), 2007–2009, 2015.