# Feature-to-Feature Regression for a Two-Step Conditional Independence Test

**Qinyi Zhang**
Department of Statistics
University of Oxford
qinyi.zhang@stats.ox.ac.uk

**Sarah Filippi**
School of Public Health
Department of Mathematics
Imperial College London
s.filippi@imperial.ac.uk

**Seth Flaxman**
Department of Statistics
University of Oxford
flaxman@stats.ox.ac.uk

**Dino Sejdinovic**
Department of Statistics
University of Oxford
dino.sejdinovic@stats.ox.ac.uk

## Abstract

The algorithms for causal discovery and more broadly for learning the structure of graphical models require well calibrated and consistent conditional independence (CI) tests. We revisit the CI tests which are based on two-step procedures and involve regression with subsequent (unconditional) independence test (RESIT) on regression residuals and investigate the assumptions under which these tests operate. In particular, we demonstrate that when going beyond simple functional relationships with additive noise, such tests can lead to an inflated number of false discoveries. We study the relationship of these tests with those based on dependence measures using reproducing kernel Hilbert spaces (RKHS) and propose an extension of RESIT which uses RKHS-valued regression. The resulting test inherits the simple two-step testing procedure of RESIT, while giving correct Type I control and competitive power. When used as a component of the PC algorithm, the proposed test is more robust to the case where hidden variables induce a switching behaviour in the associations present in the data.

## 1 INTRODUCTION

Conditional independence tests are an important component of causal discovery (cf. [Shalizi, 2016, Chapter 28]). For example, the popular PC algorithm [Spirtes et al., 2000] for recovering dependence structure among a set of variables starts from a complete undirected graph and recursively removes the edges between variables based on conditional independence testing. With normally distributed variables and linear relationships, conditional independence testing can be performed using estimates of partial correlation. Namely, when testing the hypothesis that $X \perp\!\!\!\perp Y | Z$, partial correlation is the correlation of the residuals of $X$ and $Y$ after linearly regressing each of them on $Z$ separately. However, in the presence of nonlinearities and when one needs to condition on a random vector $Z$ of larger dimensions, conditional independence testing is a challenging problem [Bergsma, 2004]. A popular approach for conditional independence testing within PC algorithm is RESIT (REgression with Subsequent Independence Test) [Hoyer et al., 2009, Peters et al., 2014], which can incorporate nonlinearities by extending the partial correlation approach in the following two ways: (1) it uses a flexible nonparametric regression of $X$ and $Y$ on $Z$, and (2) the subsequent test is based on a nonlinear dependence measure between the resulting residuals. While RESIT greatly broadens the class of models in which causal discovery with conditional independence tests is possible, as we will see, it is sensitive to departures from its modelling assumption, which itself is not straightforward to verify. In particular, it is likely to give an inflated number of false positives in the presence of associations which do not directly conform to functional relationships (as illustrated in Figure 1).

In the last decade, kernel embeddings of probability measures [Smola et al., 2007, Sriperumbudur et al., 2010] have been widely used to construct nonparametric hypothesis tests, including tests for the two-sample problem [Gretton et al., 2007, 2012], independence [Gretton et al., 2008, Chwialkowski et al., 2015], conditional independence [Fukumizu et al., 2008, Zhang et al., 2011, Doran et al., 2014], three-variable interaction [Sejdinovic et al., 2013], joint independence [Pfister et al., 2016], and goodness of fit [Chwialkowski et al., 2016]. The various kernel-based tests for conditional independence [Fukumizu et al., 2008, Zhang et al., 2011, Doran et al., 2014] allow to measure more general forms of conditional dependence than RESIT. However, these approaches typically involve complex statistics and computationally expensive procedures to estimate their distributions under the null hypothesis. In this paper we propose a kernel-based conditional independence approach which aims to strike a balance between the simplicity of the RESIT approach and the robustness to the functional association and additive noise assumptions made by the RESIT. The procedure we introduce is a test for *weak conditional independence* as defined by [Daudin, 1980] since, like RESIT, it focuses on individual effects[1] of the conditioning variable $Z$ on $X$ and $Y$. As such, it does not fully characterise conditional independence, but does benefit from a simpler testing procedure and an improved power in comparison to the "strong" tests which do characterise conditional independence. When these individual effects are not present and as such weak CI tests are insufficient to detect a complex joint effect on $(X, Y)$, they can be combined with tests for multivariate interaction [Sejdinovic et al., 2013].

An independent recent work, Strobl et al. [2017], considers a very similar approach to ours – indeed, the method we propose is essentially equivalent to the RCoT approach of Strobl et al. [2017], which in addition operates on explicit primal representations of feature maps leading to a decreased computational cost. However, Strobl et al. [2017] do not comment on the connections with the two-step procedures for CI testing and RESIT. Hence, our contribution is to provide the unifying framework which places the proposed method as a generalisation of RESIT, pointing the deficiencies of RESIT and the interplay between its structural assumptions and nonlinear dependencies. In addition, while RCoT uses approximations to the null distributions with parametric families, we simply employ a permutation-based approach using the two-step interpretation of the proposed test, and do not observe any deviations from the desired significance level in the experiments. The test maintains correct Type I control even under the challenging switching associations induced by hidden variables, where RESIT is inappropriate.

In Section 2, we overview the key notion of weak conditional independence and the existing approach for testing it as well as three related tests for general conditional independence using kernel methods. In Section 3 we develop our novel procedure for weak conditional independence testing. Section 4 evaluates the performance of our approach and compares it to RESIT and other conditional independence tests on both synthetic and real-world data.

# 2 BACKGROUND

We start by introducing notation and reviewing the existing kernel-based conditional independence tests. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be non-empty measurable spaces, with Borel $\sigma$-algebras $\mathcal{B}_\mathcal{X}$, $\mathcal{B}_\mathcal{Y}$ and $\mathcal{B}_\mathcal{Z}$ respectively. Let $k$, $l$ and $m$ be measurable positive definite kernels on these respective domains with the corresponding reproducing kernel Hilbert spaces (RKHSs) $\mathcal{H}_\mathcal{X}$, $\mathcal{H}_\mathcal{Y}$, $\mathcal{H}_\mathcal{Z}$. Let $(X, Y, Z)$ be a triple of random variables with the joint probability law $P_{XYZ}$ on $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{B}_\mathcal{X} \times \mathcal{B}_\mathcal{Y} \times \mathcal{B}_\mathcal{Z})$. We will assume $\mathbb{E}_{X\sim P_X}[k(X, X)] < \infty$ and similarly for $l$ and $m$, which will ensure that $\mathcal{H}_\mathcal{X} \subset L^2_\mathcal{X}(P_X) = \left\{f : \mathbb{E}\left[f^2(X)\right] < \infty\right\}$, and similarly $\mathcal{H}_\mathcal{Y} \subset L^2_\mathcal{Y}(P_Y)$ and $\mathcal{H}_\mathcal{Z} \subset L^2_\mathcal{Z}(P_Z)$ [Steinwart and Christmann, 2008, Section 4.3]. We will also assume that the kernels $k$, $l$ and $m$ are characteristic, such that their RKHSs are dense in the corresponding $L^2$-spaces [Sriperumbudur et al., 2010].

## 2.1 WEAK CONDITIONAL INDEPENDENCE

Conditional independence $X \perp\!\!\!\perp Y | Z$ is equivalent to

$$\mathbb{P}(X \in \mathsf{A} \,|\, Z)\,\mathbb{P}(Y \in \mathsf{B} \,|\, Z) = \mathbb{P}(X \in \mathsf{A}, Y \in \mathsf{B} \,|\, Z) \tag{1}$$

as random variables, for all measurable sets $\mathsf{A} \in \mathcal{B}_\mathcal{X}$ and $\mathsf{B} \in \mathcal{B}_\mathcal{Y}$, which in turn can be written as

$$\begin{aligned}&\mathbb{E}_Z\left[\mathbf{1}\{Z \in \mathsf{C}\}\mathbb{P}(X \in \mathsf{A} \,|\, Z)\,\mathbb{P}(Y \in \mathsf{B} \,|\, Z)\right] \\ &\quad = \mathbb{E}_Z\left[\mathbf{1}\{Z \in \mathsf{C}\}\mathbb{P}(X \in \mathsf{A}, Y \in \mathsf{B} \,|\, Z)\right] \\ &\quad = \mathbb{P}(X \in \mathsf{A}, Y \in \mathsf{B}, Z \in \mathsf{C})\end{aligned}$$

for all measurable sets $\mathsf{A} \in \mathcal{B}_\mathcal{X}$, $\mathsf{B} \in \mathcal{B}_\mathcal{Y}$ and $\mathsf{C} \in \mathcal{B}_\mathcal{Z}$. Relaxation of this property which only requires it to hold for $\mathsf{C} = \mathcal{Z}$ gives rise to the notion of *weak conditional independence*, i.e. that $\forall \mathsf{A} \in \mathcal{B}_\mathcal{X}, \mathsf{B} \in \mathcal{B}_\mathcal{Y}$

$$\mathbb{E}_Z\left[\mathbb{P}(X \in \mathsf{A} \,|\, Z)\,\mathbb{P}(Y \in \mathsf{B} \,|\, Z)\right] = \mathbb{P}(X \in \mathsf{A}, Y \in \mathsf{B}), \tag{2}$$

which we can write as $\mathbb{E}_Z[P_{X|Z} \otimes P_{Y|Z}] = P_{XY}$, understood as the equality of probability measures defined on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_\mathcal{X} \times \mathcal{B}_\mathcal{Y})$. Weak conditional independence is studied by Daudin [1980], who interprets it in terms of zero expected conditional covariances of square integrable functions, i.e.

$$\mathbb{E}_Z\left[\mathrm{Cov}\left[f(X), g(Y)|Z\right]\right] \;=\; 0 \tag{3}$$

for all $f \in L^2_\mathcal{X}(P_X)$ and $g \in L^2_\mathcal{Y}(P_Y)$. In other words, the residuals in all square integrable functions of $X$ and $Y$, given $Z$, are uncorrelated. Fukumizu et al. [2004] give another characterisation of weak conditional independence using conditional cross-covariance operators between RKHSs. Namely, they define the operator $\Sigma_{YX|Z} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, such that $\forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}$

$$\langle g, \Sigma_{YX|Z} f\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_Z\left[\mathrm{Cov}\left[f(X), g(Y)|Z\right]\right], \tag{4}$$

under the additional smoothness assumptions on conditioning variables, i.e. that $\forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}, \mathbb{E}[f(X)|Z = \cdot]$ and $\mathbb{E}[g(Y)|Z = \cdot]$ both belong to $\mathcal{H}_\mathcal{Z}$ [Fukumizu et al., 2004, Proposition 4], [Alpay, 2001]. Theorem 8 of Fukumizu et al. [2004] then shows that, for characteristic kernels, this operator vanishes if and only if weak conditional

[1] Since regressions of $X$ and $Y$ on $Z$ are done separately and hence cannot capture the joint dependence of $(X, Y)$.

independence holds. i.e.

$$\Sigma_{YX|Z} = 0 \iff \mathbb{E}[P_{X|Z} \otimes P_{Y|Z}] = P_{XY}. \quad (5)$$

The conditional independence clearly implies weak conditional independence but the converse is not true. A simple counterexample is where $X$, $Y$ and $Z$ are all pairwise independent, but are jointly dependent due to a three-variable interaction. For a concrete case, similar to the one described in Sejdinovic et al. [2013], let $X, Y, W \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and define $Z = \text{sign}(XY)|W|$. In this case, $\mathbb{E}(f(X)|Z) = \mathbb{E}(f(X))$ and $\mathbb{E}(g(Y)|Z) = \mathbb{E}(g(Y))$ implying that LHS in (3) vanishes. However, given $Z$, the pair $(X, Y)$ can only take values in two out of four quadrants, so that $X$ and $Y$ are clearly conditionally dependent given $Z$. Essentially, whenever the variables are pairwise independent but jointly dependent, any two variables are weakly conditionally independent given the third and need not be "strongly" conditionally independent. We note however that nonparametric tests exist which are directly aimed at such multivariate interaction: Sejdinovic et al. [2013] reports stronger power than the conditional independence tests in these cases. Thus, we challenge the paradigm by which all departures from conditional independence should be tested with the same procedure and instead choose to focus on weak conditional independence testing. When multivariate interaction is present and the individual effects of $Z$ on $X$ and on $Y$ are weak or non-existing and only the joint dependence of $(X, Y)$ on $Z$ is possible to detect, then arguably $(X, Y)$ should be considered a single random vector.

## 2.2 RESIT: THE TWO-STEP APPROACH

RESIT approach [Hoyer et al., 2009, Peters et al., 2014, Flaxman et al., 2015] converts a conditional independence testing problem into an unconditional one by removing the effect of a confounder $Z$ through a flexible nonparametric regression. More specifically, it assumes an additive noise model where $X$ and $Y$ are expressed as some deterministic functions of $Z$ plus an additive zero-mean noise term, i.e.

$$X = f(Z) + n_x \quad (6)$$

$$Y = g(Z) + n_y \quad (7)$$

where $n_x$ and $n_y$ are zero-mean random variables independent of $Z$. Under the assumptions (6) and (7), the conditional independence can be characterised as:

$$X \perp\!\!\!\perp Y | Z \iff n_x \perp\!\!\!\perp n_y. \quad (8)$$

Thus, the test proceeds by first regressing $X$ on $Z$ and $Y$ on $Z$, and then testing for the (unconditional) independence between the fitted residuals of these regressions $\hat{\epsilon}_x = X - \hat{\mathbb{E}}[X|Z]$ and $\hat{\epsilon}_y = Y - \hat{\mathbb{E}}[Y|Z]$. This approach obviously crucially depends on the regression procedure in order to remove the effect of $Z$ by using an appropriate model of regression functions $f$ and $g$. Gaussian Process regression and kernel ridge regression are often used for the first step, whereas Hilbert-Schmidt Independence Criterion [Gretton et al., 2008] can be used in the second step in order to capture potentially *nonlinear dependence* between residuals. Because of its focus on individual effects of the conditioning variable $Z$ on $X$ and on $Y$, it is clear that RESIT only tests for weak conditional independence.

As we will demonstrate below, RESIT can result in a substantial inflation of false discoveries when assumptions (6) and (7) are violated. Indeed, RESIT is unable to handle non-functional dependence between the variables $X$ and $Y$ and the conditioning variable $Z$ – the regression will not remove the dependence on $Z$ and the fitted residuals will be dependent even in the cases where $X$ and $Y$ are independent given $Z$. We see this as an undesirable property for many real data applications, especially when conditioning on a multivariate $Z$, as these assumptions are difficult to verify and may be violated. In particular, there could be hidden categorical confounders which introduce a switching behaviour in the way $X$ and $Y$ depend on $Z$, i.e. $X$ and $Y$ are functionally dependent on $Z$ and those confounders, but not on $Z$ itself. An example of such a relationship is illustrated in Figure 2 which plots the total expenditure on health per capita vs the gross national income per capita for 178 countries: there is a superposition of two functional relationships corresponding to two subsets of countries depending on whether their economy relies on oil [WHO, 2009].

For a simple example of how brittle the RESIT approach is, consider the dataset presented in Figure 1. Here, the dependence of both $X$ and $Y$ on $Z$ is a mixture of two functional relationships (linear for $X$, quadratic for $Y$), arguably very simple – but also indicating the presence of some latent switching mechanism for both $X$ and $Y$. However, the conditional expectations in both cases are constant and thus independent of $Z$. Hence, the fitted regression functions do not capture any dependence on $Z$ and as a result, the residuals (rightmost plot) are clearly dependent, regardless of whether in fact $X \perp\!\!\!\perp Y|Z$ (indeed, in this case, as described in Section 4.2, the conditional independence does hold because switching variables are independent). RESIT with a nonlinear independence test on these residuals will therefore falsely reject the null hypothesis. This is a highly undesirable property as it leads to an inflated number of false positives, and when the same method is used inside PC algorithm, it can potentially result in reporting spurious causal links as illustrated in Section 4.3. This problem is exacerbated for $Z$ being a random vector, even of a small dimension, since such cases which go beyond simple functional relationships become difficult to notice. One way to avoid the inflation of false positives in these cases is to only test for *linear dependence* between the residuals. However, this clearly results in a reduction of power as it misses many alternative models where (6) and (7) are satisfied and $n_x$ and $n_y$ are nonlinearly dependent.

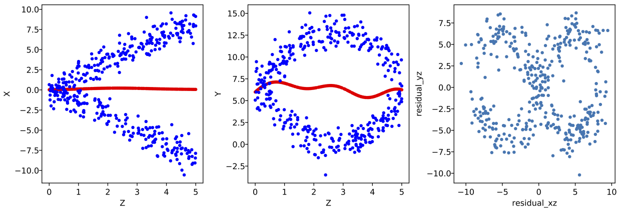

Figure 1: Data generated from the null model of Section 4.2 and kernel ridge regressions are used to remove dependency on $Z$. Left: $X$ against $Z$. Middle: $Y$ against $Z$. Simulated data (blue) and fitted values (red). Right: residuals of the regression for $Y$ against residuals of the regression for $X$ which clearly exhibits non-linear dependency.

## 2.3 KERNEL TESTS FOR "STRONG" CONDITIONAL INDEPENDENCE

In contrast to the RESIT test, there have been several approaches in the literature that aim to measure all forms of conditional dependence, typically at the expense of more complex statistics which have the distribution under the null hypothesis that is more difficult to estimate.

**Cross-Covariance Operators** Fukumizu et al. [2008] proposed a general nonparametric characterisation of conditional independence based on the conditional cross-covariance operators between RKHSs. The cross-covariance operator $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is defined through

$$\langle g, \Sigma_{YX} f \rangle = \text{Cov}(f(X), g(Y)), \quad f \in \mathcal{H}_\mathcal{X},\ g \in \mathcal{H}_\mathcal{Y},$$

i.e., it is a nonlinear extension of the cross-covariance matrix (cf. Baker [1973], Fukumizu et al. [2004]). The conditional cross covariance operator $\Sigma_{YX|Z}$ is then defined as

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \tag{9}$$

in analogy to the conditional cross-covariance matrix $C_{YX|Z} = C_{YX} - C_{YZ}C_{ZZ}^{-1}C_{ZX}$ formula for jointly Gaussian random vectors.

The conditional dependence measure can then be based on estimating $\|\Sigma_{\ddot{Y}\ddot{X}|Z}\|_{HS}^2$, where $\ddot{Y} = (Y, Z)$ and $\ddot{X} = (X, Z)$, as it can be shown [Fukumizu et al., 2008] that $\Sigma_{\ddot{Y}\ddot{X}|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$ when the product of the three positive definite kernels, $klm$, results in a characteristic kernel on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. In addition, Fukumizu et al. [2008] considers the conditional cross-correlation operators, which can be expressed as $V_{YX|Z} = \Sigma_{YY}^{-1}\Sigma_{YX|Z}\Sigma_{XX}^{-1}$.

The test statistics with these approaches however do not have clear asymptotic null distributions. Consequently, one needs to adopt a *local permutation-based approach* which needs to take into account an artificial discretisation of condition $Z$ (e.g. through clustering) so that the marginal structures are preserved. As remarked by Zhang et al. [2011], this usually requires large sample size and the results become unreliable when the dimension of $Z$ increases.

**Kernel Conditional Independence Test (KCI-Test[2])** Following similar reasoning, Zhang et al. [2011] characterise conditional independence based on partial association [Daudin, 1980] where it was shown that

$$X \perp\!\!\!\perp Y|Z \iff \mathbb{E}(\tilde{f}\tilde{g}) = 0 \tag{10}$$

for all $\tilde{f} \in \mathcal{E}_{XZ} := \{\tilde{f} \in L^2_{\mathcal{X}\times\mathcal{Z}}(P_{XZ})|\mathbb{E}(\tilde{f}|Z) = 0\}$ and $\tilde{f}(\ddot{X}) = f(\ddot{X}) - \mathbb{E}(f|Z)$ for $f \in L^2_{\mathcal{X}\times\mathcal{Z}}(P_{XZ})$. The notations for $\tilde{g}$, $g$ and $\mathcal{E}_{YZ}$ are defined similarly. As noted, if the functions $f$ and $g$ are restricted to the spaces $\mathcal{H}_{\mathcal{X}\times\mathcal{Z}}$ and $\mathcal{H}_{\mathcal{Y}\times\mathcal{Z}}$, then such characterisation is exactly the same as Fukumizu et al. [2008]. Zhang et al. [2011] derive an explicit formulae for the asymptotic null distribution and propose an approach for approximating it based on eigendecompositions of kernel matrices or fitting a parametric family. However, this null distribution becomes harder to accurately approximate in practice as the dimension of the conditioning variable increases.

**Permutation-based Maximum Mean Discrepancy (MMD)** Following the difficulties faced by the approaches of Fukumizu et al. [2008] and Zhang et al. [2011] in estimating the null distributions, Doran et al. [2014] introduce a reduction of the conditional independence testing problem into a two-sample testing one, by considering a carefully selected permutation which mimics a sample from $P_Z P_{X|Z} P_{Y|Z}$ for which the null hypothesis of conditional independence holds. While casting the problem as a simple two-sample test is attractive, promising a better calibration and improved performance in the cases where dimensionality of the conditioning variable is high, this approach does require a costly optimization procedure over the space of doubly stochastic matrices in order to select the required permutation.

# 3 PROPOSED METHOD

We have seen that RESIT, although a simple and effective conditional independence test for additive noise models, can lead to undesirable properties when the additive noise modelling assumption is violated. On the other hand, strong kernel-based independence tests require no restrictive assumptions on the type of the underlying relationships

[2] Termed by Zhang et al. [2011]

between the variables, but need complex and difficult to tune testing procedures, often requiring to solve difficult side-problems in order to have an estimate of the distribution under the null hypothesis (clustering the conditioning variable or optimisation over the space of permutations). Therefore, our goal is to identify an approach which aims to strike a balance between these two approaches, and construct a test which is more robust to the departures from the modelling assumption in (6) and (7). At the same time, we seek an approach which can be cast as an unconditional independence test on some form of regression residuals – allowing us to employ the well established existing methodology for unconditional tests, including straightforward permutation-based approaches for estimation of the null distribution. Using heuristic arguments, we derive in this section an intuitive test based on well known empirical mean RKHS quantities. The rigorous statistical analysis of this approach is left to further work.

We first note that the assumptions (6) and (7) specify that the conditional distributions $P_{X|Z=z}$ and $P_{Y|Z=z}$ depend on $z$ only through the expectations $\mathbb{E}\left[X|Z=z\right]$ and $\mathbb{E}\left[Y|Z=z\right]$. Therefore, any additional dependence on $z$, e.g. in the conditional variance: $\text{Var}\left[X|Z=z\right] \neq \text{Var}\left[X\right]$ will not be captured by the regression approaches, implying that the residuals remain dependent by construction, leading to spurious rejections when nonlinear dependence tests on residuals are employed. Therefore, we are interested in also modelling how higher-order moments of these distributions depend on $z$, leading us to consider *conditional expectations of feature maps* $\mathbb{E}\left[\phi(X)|Z=z\right]$ and $\mathbb{E}\left[\psi(Y)|Z=z\right]$ for some $\phi\colon \mathcal{X} \to \mathcal{H}_\mathcal{X}$, $\psi\colon \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and Hilbert spaces $\mathcal{H}_\mathcal{X}$, $\mathcal{H}_\mathcal{Y}$, i.e. the mean embeddings of the corresponding conditional distributions. For characteristic kernels [Sriperumbudur et al., 2010], these embeddings fully characterise the corresponding distributions (and thus, their dependence on $z$).

We thus propose the following two-step approach: first construct feature representations of $X$ and $Y$ and then perform the *vector-valued regression* of these feature representations of each of them on $Z$ separately. Just like in RESIT, the second step is a test of independence between the resulting residuals, but note that the residuals are themselves elements of the corresponding (potentially infinite-dimensional) feature spaces. Fortunately, as we show in the remainder of this section, using the kernel trick allows to perform each of these steps without ever explicitly computing the feature maps. The final result is a simple test statistic, very similar to HSIC of Gretton et al. [2008], which lends itself to direct computation. Moreover, its null distribution can be straightforwardly estimated using a permutation-based approach. Thus, the key step of the procedure is the feature transformation of the responses in regression – we will see that this greatly relaxes the modelling assumptions and in particular allows to model complex dependencies between the responses and the conditioning variable which cannot be expressed in functional forms with additive noise.

Consider now $k$, $l$ and $m$ to be measurable positive definite kernels with the corresponding RKHSs $\mathcal{H}_\mathcal{X}$, $\mathcal{H}_\mathcal{Y}$, $\mathcal{H}_\mathcal{Z}$. We denote by $\phi(x) \in \mathcal{H}_\mathcal{X}$ and $\psi(y) \in \mathcal{H}_\mathcal{Y}$ the feature map representations of $k$ and $l$. Note that Daudin's condition (3) for weak conditional independence can be written as

$$\begin{aligned} 0 &= \mathbb{E}_Z\left[\text{Cov}\left[f(X), g(Y)|Z\right]\right] \\ &= \mathbb{E}_{XYZ}\left[\left(f(X) - \mathbb{E}f(X|Z)\right)\left(g(Y) - \mathbb{E}g(Y|Z)\right)\right] \\ &= \mathbb{E}_{XYZ}\left[n_f(X,Z)n_g(Y,Z)\right] \qquad (11) \end{aligned}$$

where we denoted $n_f(X,Z) = f(X) - \mathbb{E}f(X|Z)$, $n_g(Y,Z) = g(Y) - \mathbb{E}g(Y|Z)$ to be the regression residuals for $X$ and $Y$ transformed through square integrable functions $f \in L^2_\mathcal{X}(P_X)$ and $g \in L^2_\mathcal{Y}(P_Y)$. Note that these residuals would in general depend on $Z$ and that they have mean zero by construction. Thus, weak conditional independence is equivalent to *uncorrelatedness* of residuals $n_f(X,Z)$ and $n_g(Y,Z)$ for all $f \in L^2_\mathcal{X}(P_X)$ and $g \in L^2_\mathcal{Y}(P_Y)$. Moreover, when $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are dense in $L^2_\mathcal{X}(P_X)$ and $L^2_\mathcal{Y}(P_Y)$, it suffices to consider $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. But then

$$\begin{aligned} &\mathbb{E}_{XYZ}\left[\left(f(X) - \mathbb{E}f(X|Z)\right)\left(g(Y) - \mathbb{E}g(Y|Z)\right)\right] \\ &= \mathbb{E}_{XYZ}\langle f, \phi(X) - \mathbb{E}\phi(X|Z)\rangle_{\mathcal{H}_\mathcal{X}}\langle g, \psi(Y) - \mathbb{E}\psi(Y|Z)\rangle_{\mathcal{H}_\mathcal{Y}} \\ &= \langle f, \mathbb{E}_{XYZ}\left(\phi(X) - \mathbb{E}\phi(X|Z)\right) \otimes \left(\psi(Y) - \mathbb{E}\psi(Y|Z)\right) g\rangle_{\mathcal{H}_\mathcal{X}} \\ &= \langle f, \mathbb{E}_{XYZ}\left[n_\phi(X,Z) \otimes n_\psi(Y,Z)\right] g\rangle_{\mathcal{H}_\mathcal{X}}. \end{aligned}$$

This suggests the method which regresses $\phi(X)$ and $\psi(Y)$ on $Z$ using vector-valued kernel ridge regression and tests for unconditional linear independence of the residuals $n_\phi(X,Z)$ and $n_\psi(Y,Z)$ which are elements of RKHSs.

The resulting regression functions are now simply the empirical mean embeddings of the conditional distributions $P_{X|Z=z}$ and $P_{Y|Z=z}$ [Fukumizu et al., 2011, Theorem 1] and can be expressed as follows [Song et al., 2009, Theorem 5]:

$$\hat{\mathbb{E}}(\phi(X)|Z=z) = \sum_{i=1}^{n} \beta^{\mathbf{x}}_i(z)\phi(x_i) = \Phi_\mathbf{x}^\top \beta^\mathbf{x}(z) \qquad (12)$$

$$\hat{\mathbb{E}}(\psi(Y)|Z=z) = \sum_{i=1}^{n} \beta^{\mathbf{y}}_i(z)\psi(y_i) = \Psi_\mathbf{y}^\top \beta^\mathbf{y}(z) \qquad (13)$$

with $\beta^\mathbf{x}(z) = (M + \lambda^\mathbf{x} I_n)^{-1}(m(z,z_1), ..., m(z,z_n))^\top$, $M$ is the $n \times n$ matrix such that $M_{ij} = m(z_i, z_j)$, $\Phi_\mathbf{x} = (\phi(x_1), ..., \phi(x_n))^\top$ and $\Psi_\mathbf{y} = (\psi(y_1), ..., \psi(y_n))^\top$. The notation $\beta^\mathbf{x}$ emphasises the fact that the coefficient vector may differ for $\mathbf{x}$ and $\mathbf{y}$; this difference would only be due to different choice of the regularisation parameters $\lambda^\mathbf{x}$ and $\lambda^\mathbf{y}$. Note that the procedure can straightforwardly be amended to use different kernels $m^\mathbf{x}$ and $m^\mathbf{y}$ for the two responses, but we do not pursue this further for simplicity of exposition. We have enclosed two parameter optimisation schemes in the Appendix.

The residual of the kernel ridge regression of $X$ on $Z$ for the $i$-th data point is given by

$$\begin{aligned}\hat{\epsilon}_{x,i} &= \phi(x_i) - \hat{\mathbb{E}}(\phi(X)|Z = z_i) \\ &= \Phi_{\mathbf{x}}^\top \mathbf{e}_i - \Phi_{\mathbf{x}}^\top \beta^{\mathbf{x}}(z_i) \in \mathcal{H}_k \end{aligned} \tag{14}$$

so that the whole set of residuals $\hat{\epsilon}_x = (\epsilon_{x,1}, ..., \epsilon_{x,n})^\top$ can be written as:

$$\hat{\epsilon}_x = (I_n - B_{\mathbf{x}})\Phi_{\mathbf{x}} = (\frac{1}{\lambda^{\mathbf{x}}} M + I_n)^{-1} \Phi_{\mathbf{x}} \tag{15}$$

with $B_{\mathbf{x}} = M(M + \lambda^{\mathbf{x}} I_n)^{-1}$ and $\lambda^{\mathbf{x}} > 0$, so that $I_n - B_{\mathbf{x}} = (\frac{1}{\lambda^{\mathbf{x}}} M + I_n)^{-1}$. Similarly the residual of the kernel ridge regression of $Y$ on $Z$ is given by

$$\hat{\epsilon}_y = (\frac{1}{\lambda^{\mathbf{y}}} M + I_n)^{-1} \Psi_{\mathbf{y}}. \tag{16}$$

The problem of testing weak conditional independence between $X$ and $Y$ given $Z$ then translates into testing if $\hat{\epsilon}_x$ and $\hat{\epsilon}_y$ are correlated.

Since they are elements of RKHSs, we can use the Hilbert-Schmidt independence criterion (HSIC) of Gretton et al. [2008] with linear (inner product) kernels on these residuals, i.e. $\kappa_x(\epsilon_{x,i}, \epsilon_{x,j}) = \langle \epsilon_{x,i}, \epsilon_{x,j} \rangle_{\mathcal{H}_X}$ and $\kappa_y(\epsilon_{y,i}, \epsilon_{y,j}) = \langle \epsilon_{y,i}, \epsilon_{y,j} \rangle_{\mathcal{H}_Y}$. Let the empirically centred residuals be $\hat{\epsilon}^c_{x_i} := \hat{\epsilon}_{x_i} - \frac{1}{n}\sum_{j=1}^n \hat{\epsilon}_{x_j}$ and $\hat{\epsilon}^c_{y_i} := \hat{\epsilon}_{y_i} - \frac{1}{n}\sum_{j=1}^n \hat{\epsilon}_{y_j}$. Essentially, we build an empirical cross covariance operator between the residuals $\hat{\epsilon}_x$ and $\hat{\epsilon}_y$, and HSIC is then the squared Hilbert-Schmidt norm of such operator:

$$\begin{aligned}\Xi(\hat{\epsilon}_x, \hat{\epsilon}_y) &= \left\| \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}^c_{x_i} \otimes \hat{\epsilon}^c_{y_i} \right\|^2_{HS} &(17)\\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \langle \hat{\epsilon}^c_{x_i}, \hat{\epsilon}^c_{x_j} \rangle \langle \hat{\epsilon}^c_{y_i}, \hat{\epsilon}^c_{y_j} \rangle &(18)\\ &= \frac{1}{n^2} Trace(H \hat{\epsilon}_x \hat{\epsilon}_x^\top H H \hat{\epsilon}_y \hat{\epsilon}_y^\top H) &(19)\end{aligned}$$

where $H := I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. By definition of $\hat{\epsilon}_x$, we have

$$\hat{\epsilon}_x \hat{\epsilon}_x^\top = (\frac{1}{\lambda^{\mathbf{x}}} M + I_n)^{-1} \Phi_{\mathbf{x}} \Phi_{\mathbf{x}}^\top (\frac{1}{\lambda^{\mathbf{x}}} M + I_n)^{-1} \tag{20}$$

and similarly

$$\hat{\epsilon}_y \hat{\epsilon}_y^\top = (\frac{1}{\lambda^{\mathbf{y}}} M + I_n)^{-1} \Psi_{\mathbf{y}} \Psi_{\mathbf{y}}^\top (\frac{1}{\lambda^{\mathbf{y}}} M + I_n)^{-1} \tag{21}$$

where $\Phi_{\mathbf{x}} \Phi_{\mathbf{x}}^\top = K$ and $\Psi_{\mathbf{y}} \Psi_{\mathbf{y}}^\top = L$ are the $n \times n$ matrices such as $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$. Combining equations (19), (20) and (21), we obtain

$$\Xi(\hat{\epsilon}_x, \hat{\epsilon}_y) = \frac{1}{n^2} Trace(\tilde{H}_{\mathbf{z},\mathbf{x}} K \tilde{H}_{\mathbf{z},\mathbf{x}}^\top \tilde{H}_{\mathbf{z},\mathbf{y}} L \tilde{H}_{\mathbf{z},\mathbf{y}}^\top)$$

where $\tilde{H}_{\mathbf{z},\mathbf{x}} = H(\frac{1}{\lambda^{\mathbf{x}}} M + I_n)^{-1}$ and $\tilde{H}_{\mathbf{z},\mathbf{y}} = H(\frac{1}{\lambda^{\mathbf{y}}} M + I_n)^{-1}$. Note that the HSIC statistic between the two residuals can be seen as the inner product between transformations of the kernel matrices $K$ and $L$ through the centering terms $\tilde{H}_{\mathbf{z},\mathbf{x}}$ and $\tilde{H}_{\mathbf{z},\mathbf{y}}$ which encompass information regarding the impact of $Z$ on respectively $X$ and $Y$. To obtain the test threshold, standard permutation approach can be used to estimate the null distribution. It is tempting to also consider the nonlinear tests on RKHS residuals, i.e. one may specify some kernel which only depends on RKHS distances (e.g. Gaussian RBF kernels), however, this results in inflated false positive rates for the same reasons as RESIT.

**Related Work** Our proposed conditional independence test which we term KRESIT[3] (Kernel RESIT) is a generalisation of the RESIT approach in which the regressions on $Z$ are done after feature transforming both responses $X$ and $Y$. KRESIT is closely related to the method of Zhang et al. [2011]. More precisely, Zhang et al. [2011] can be understood as performing ridge regressions of the feature transformations of pairs $(X, Z)$ and $(Y, Z)$ on $Z$ in order to achieve full characterisation of conditional independence. This makes the method more difficult to interpret in terms of two-step procedures. Recent extension by Strobl et al. [2017] of Zhang et al. [2011] presents the test called RCoT which essentially computes the same statistic as KRESIT, but uses a different estimation procedure of the asymptotic null distribution. RCoT can be viewed as a large scale approximation of KRESIT through the use of random Fourier features (RFF), similarly RCIT of Strobl et al. [2017] provides an approximation of the KCI-test via RFF. Strobl et al. [2017] has shown that RCIT and RCoT give similar performance as KCI-test but they are orders of magnitude faster in large-scale settings. While the large-scale approximations are not the focus of the present work, methods such as RFF or Nyström approximation can readily be employed in KRESIT (for comparisons of these approximation methods in unconditional independence testing, cf. Zhang et al. [2017]).

# 4 EXPERIMENTS

We apply the proposed method, KRESIT, to both synthetic and real data to evaluate its performance in terms of Type I error and statistical power. We start with a motivating real data example comparing KRESIT and RESIT on two socio-economic and health indicators from the World Health Organisation (WHO) data [WHO, 2009, Rosling, 2008]. Following this, we compare the performance of the proposed approach against RESIT and the three "strong" conditional independence tests described in Section 2.3 on a synthetic experiment with increasing dimensional-

[3] Code available at https://github.com/oxmlcs/kerpy.

ity of the conditioning variable $Z$ and sample sizes. We then apply the two "weak" conditional independence tests (RESIT and KRESIT) to unravel causal relationships in a synthetic dataset, the Boston Housing dataset [Pace and Gilley, 1997, Harrison and Rubinfeld, 1978] and the Ozone dataset [Breiman and Friedman, 1985] (presented in the Appendix).

Unless otherwise stated, the significance level is kept at $\alpha = 0.05$. For KRESIT, Gaussian kernels with median heuristic are used for all three random variables in regression, but the parameters $\lambda^{\mathbf{x}}$ and $\lambda^{\mathbf{y}}$ are optimised using grid search to find the minimum total 5-fold cross validation error over a grid of 30 evenly spaced values in the interval $(10^{-6}, 10^{1})$. For RESIT, Gaussian kernels with median heuristic are used for kernel ridge regression and on the residuals. The regularisation parameters are tuned in the same way as in KRESIT.

## 4.1 SIMPLE MOTIVATING EXAMPLE BASED ON WHO DATA

As a motivating example, we apply the proposed approach to two variables in the WHO data set used in Reshef et al. [2011]. We denote by $Z$ the log transformed gross national income per capita and by $Y$ the total expenditure on health per capita. Figure 2 (left) can be thought of as the superposition of two relationships where the small minority curve consists of countries whose economies rely largely on oil [WHO, 2009]. We have taken the log transform of the gross national income per capita so that the values are more evenly distributed. After removing all missing values, we have 178 data points. We then construct a synthetic $X$ with non-functional dependence on $Z$ as follows

$$x_i = \begin{cases} (z_i - 10)^2 + n_i^x & \text{if } c_i^x = 1 \\ -(z_i - 10)^2 + 35 + n_i^x & \text{if } c_i^x = 0 \end{cases} \quad (22)$$

where $c_i^x \overset{i.i.d.}{\sim} Bernoulli(0.5)$ and $n_i^x \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. $X$ and $Y$ are conditionally independent given $Z$ by construction.

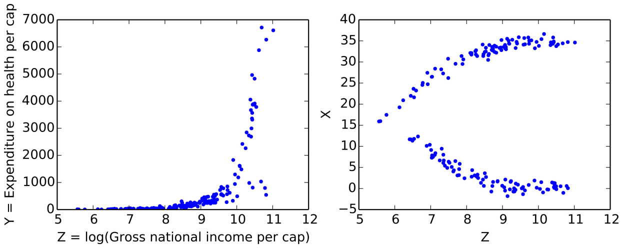

Figure 2: Left: Expenditure on health per cap against the logarithm of gross national income per cap ($Z$). Right: Synthetic data $X$ against $Z$.

RESIT incorrectly rejected the null hypothesis at $5\%$ significance level with a p-value of $0.0025$ owing to the fact that kernel ridge regression is unable to remove the dependency of $Z$ from $X$ and $Y$ and HSIC using Gaussian kernels is able to detect such nonlinear dependency between the residuals. KRESIT, on the other hand, gives a p-value of $0.91$ and hence does not reject the null hypothesis that $X$ and $Y$ are weakly conditionally independent. By relaxing the modelling assumption of additive noise and allowing more complex dependencies between responses and the conditioning variable, KRESIT is able to provide a better calibrated test.

## 4.2 EFFECT OF DIMENSIONALITY OF $Z$ AND SAMPLE SIZE

We examine the probability of Type I and Type II errors of the "strong" conditional independence tests vs. the "weak" conditional independence tests when the dimensionality of the conditioning set $Z$ is increasing ($d = 1, 2, 3, 4, 5, 6, 7$) and the sample sizes take values in $\{40, 80, 120, 160, 200\}$.

Let us consider the following non-functional dependence as the null model:

$$Z_i^j \in \overset{i.i.d.}{\sim} Uniform(0,5),\ j = 1, \ldots, d$$

$$C_i^x, C_i^y \overset{i.i.d.}{\sim} Bernoulli(0.5)$$

$$n_i^x, n_i^y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$X_i \in \mathbb{R} = \begin{cases} 1.7Z_i^1 + n_i^x & \text{if } c_i^x = 1 \\ -1.7Z_i^1 + n_i^x & \text{if } c_i^x = 0 \end{cases}$$

$$Y_i \in \mathbb{R} = \begin{cases} (Z_i^1 - 2.7)^2 + n_i^y & \text{if } c_i^y = 1 \\ -(Z_i^1 - 2.7)^2 + 13 + n_i^y & \text{if } c_i^y = 0 \end{cases}$$

Note, $X$ and $Y$ only depend on the first dimension of the conditioning vector $Z$. To obtain the alternative model, we couple the latent variables $C^x$ and $C^y$:

$$u_i \overset{i.i.d.}{\sim} Uniform(0,1)$$

$$C_i^x \overset{i.i.d.}{\sim} Bernoulli(0.5)$$

$$C_i^y = \begin{cases} C_i^x & \text{if } u_i < 0.3 \\ \overset{i.i.d.}{\sim} Bernoulli(0.5) & \text{if } u_i \geq 0.3 \end{cases}$$

We compare the "weak" conditional independence tests: the proposed KRESIT, RESIT and LRESIT (RESIT with linear kernels on residuals) with the "strong" conditional independence tests: CIperm by Fukumizu et al. [2008], KCIPT by Doran et al. [2014] and KCI-test by Zhang et al. [2011]. The rejection rates are calculated out of 100 trials. Each rejection rate is associated with a 95% Wald confidence interval.

For a given dimension, we observe in Figure 3 that the Type I error of RESIT is increasing with the number of samples. As a simple kernel ridge regression cannot remove the dependency of $Z$ from $X$ and $Y$, the nonlinear dependency pattern in the residuals are more visible as sample size increases. As expected, the proposed approach, KRESIT, has correct Type I control together with LRESIT and the other "strong" conditional independence tests.

To establish sensible comparison in the alternative model (Figure 4), we only compare those methods with correct Type I control. When $Z$ is of dimension 1, KRESIT performs similarly to KCI test where the probability of Type II error (1-power) for both decreases to zero as the number of samples increases to 200, while the other tests give weaker power. In dimension 7, KRESIT again gives strong performance, outperforming other tests. Although LRESIT also tests for weak conditional independence, as expected, it has a reduced power comparing to KRESIT in both cases. While KCI test and KCIPT give similar power performance, CIperm struggles to detect the conditional dependence across all sample sizes.

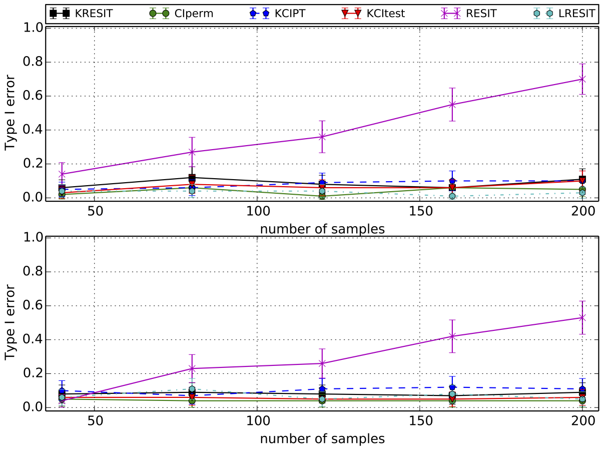

Figure 3: Null model experiments for $d = 1$ (Top) and $d = 7$ (Bottom).

## 4.3 APPLICATION IN CAUSAL DISCOVERY

Conditional independence tests are frequently used in causal discovery to recover the dependence relationships among a set of variables represented as a directed acyclic graph (DAG). Assuming causal Markov condition (i.e., any variable/node is conditionally independent of its nondescendents given its parents [Hausman and Woodward, 1999]) and faithfulness (i.e., the conditional independences of the distribution can be inferred from the d-separation in the graph and vice-versa [Kalisch and Buhlmann, 2007]), the popular constraint-based PC-algorithm [Spirtes et al., 2000] is often used to estimate a Markov equivalence class of DAGs (i.e. a set of graphs that impose the same independences and conditional independences). The estimation starts from a complete undirected graph and recursively deletes edges based on conditional independence decisions. The result is an undirected graph (skeleton) which is then partially directed and further extended to represent the underlying equivalence class of DAG [Kalisch and Buhlmann, 2007]. The quality of the conditional independence tests are hence crucial for the performance of the PC algorithm.

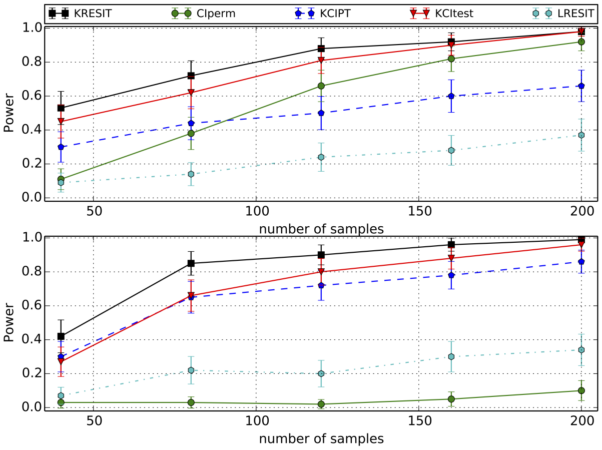

Figure 4: Alternative model experiments for $d = 1$ (Top) and $d = 7$ (Bottom).

The PC algorithm used in this section is a modified version of the pcalg implementation in python[4] with conditional independence test using KRESIT/RESIT and independence test using HSIC with Gaussian kernel using median heuristic. Completed partially acyclic graph (CPDAG) are used to visualise the equivalence classes of DAGs. Whenever an undirected edge $i - j$ is shown, there exists a DAG with $i \rightarrow j$ and a DAG with $i \leftarrow j$ in the equivalence class.

### 4.3.1 Synthetic Data

Consider variables of the null model from Section 4.2 and additionally let $A_i = (Y_i - 5)^2/3 + 5 + n_i^A$ and $B_i = 5.5\tanh(Y_i) + n_i^B$ where $n_i^A, n_i^B \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

We compare the results obtained from the PC algorithm with the variables $\{X, Y, Z, C^x, C^y, A, B\}$ using KRESIT and RESIT. Using a sample of size 800, we obtain Figure 5 (Left). Both KRESIT and RESIT recover the correct structure of the model.

We now turn to investigating the robustness of the method

[4] https://github.com/keiichishima/pcalg

upon removing the variables $C^x$ and $C^y$. PC algorithm with KRESIT is still able to discover the correct skeleton as shown in Figure 5 (Right). An incorrect graph is obtained with RESIT, reporting a spurious causal link $X - Y$. We note that when the switching variables $C^x, C^y$ are present, all associations are functional, but this is no longer the case when $C^x, C^y$ are unobserved.

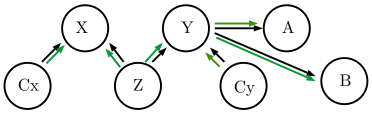

Figure 5: Left: The DAG obtained by using KRESIT (black) fully recovers the causal structure, but the one using RESIT (green) does not. Right: The DAG obtained by PC algorithm using KRESIT and RESIT with $C^x$ and $C^y$ being latent variables.

#### 4.3.2 Boston Housing Data

We consider the corrected Boston Housing dataset of Pace and Gilley [1997] and pre-whiten each variable using the spatial coordinates with a GP regression as in Flaxman et al. [2015] since there is significant spatial clustering in every single variable in the dataset. There are 14 variables in total and each with 506 observations. For a detailed explanation of each variable, we refer to the original paper [Harrison and Rubinfeld, 1978]. To correct for multiple testing, we set the significance level to $\alpha = 0.001$.

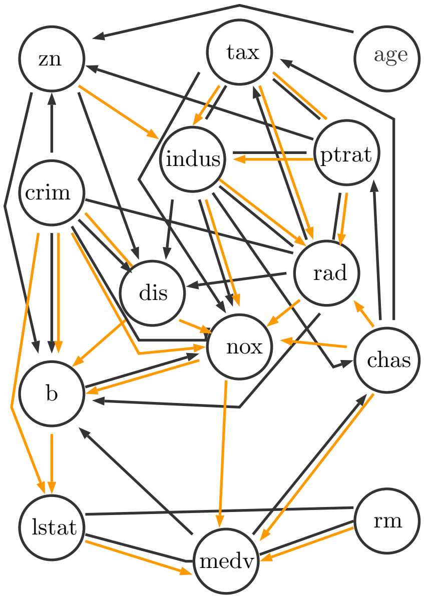

Figure 6: Results of PC algorithm with KRESIT (black) and RESIT [Flaxman et al., 2015] (orange) using pre-whitened Boston Housing dataset.

Overall, the CPDAG obtained by our proposed method is substantially different from previous results in the literature [Zhang et al., 2011, Flaxman et al., 2015]. We argue that many potentially unobserved variables could be driving such differences. The details and some possible interpretations are given in the Appendix: Boston Housing Data. Therefore, the links discovered using RESIT and KRESIT within PC algorithm may need further investigation and one should analyse the results with potential hidden variables in mind and be cautious about any conclusion drawn due to potential violation of structural assumptions.

# 5 CONCLUSION

We proposed a weak conditional independence test that extends the popular two-step approach RESIT, which combines regression with an unconditional independence test. We consider RKHS-valued ridge regressions and subsequently use a test for linear independence on RKHS-valued residuals. While maintaining simple and effective testing procedures of RESIT, the resulting test has a correct Type I control under more challenging scenarios where the modelling assumptions required by RESIT are violated. It also yields competitive or improved power performance to that of the other conditional independence tests. When used in the PC algorithm, the proposed method is more robust than RESIT to hidden variables inducing associations more complex than functional ones with additive noise.

**Acknowledgements**

QZ was supported by EPSRC (EP/M50659X/1), SF and DS were supported in part by ERC (FPT/617071).

# References

World health organization statistical information systems (whosis). World Health Organization Statistical Information Systems (WHOSIS), 2009. www.who.int/whosis/en/.

D. Alpay. *The Schur Algorithm, Reproducing Kernel Spaces and System Theory*. American Mathematics Society, 2001.

C.R. Baker. Joint Measures and Cross-Covariance Operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

W. Bergsma. Testing Conditional Independence for Continuous Random Variable. *EURANDOM-report 2004-049*, 2004.

L. Breiman and J.H. Friedman. Estimating Optimal Transformantions for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580 – 598, 1985.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

J. Dattorro. *Convex Optimisation & Euclidean Distance Geometry*. Meboo Publishing USA, second edition, 2016.

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

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A Permutation-Based Kernel Conditional Independence Test. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

S.R. Flaxman, D.B. Neill, and A.J. Smola. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM Transactions on Intelligent Systems and Technology*, 7 (2):22:1–22:23, 2015.

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

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

K. Fukumizu, Le Song, and A. Gretton. Kernel Bayes' Rule. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Method for the Two-Sample problem. *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Gretton, K. Fukumizu, B. Schölkopf, C.H. Teo, L. Song, and A.J. Smola. A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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

DJ.R. Harrison and D.L. Rubinfeld. Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

D.M. Hausman and J. Woodward. Independence, Invariance, and the Causal Markov Condition. *British Journal of Philosophy and Science*, 50(4):521–583, 1999.

P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear Causal Discovery with Additive Noise Models. In *Advances in Neural Information Processing Systems (NIPS)*. 2009.

M. Kalisch and P. Buhlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research (JMLR)*, 8:613–636, 2007.

R. Kelley Pace and Otis W. Gilley. Using the Spatial Configuration of the Data to Improve Estimation. *The Journal of Real Estate Finance and Economics*, 14(3):333–340, 1997.

J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research (JMLR)*, 15(1), 2014.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based Tests for Joint Independence. *Journal of the Royal Statistical Society*, page to appear, 2016.

D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524, 2011.

H. Rosling. Indicators in Gapminder World. Gapminder, 2008.

D. Sejdinovic, A. Gretton, and W. Bergsma. A Kernel Test for Three-Variable Interactions. In *Advances in Neural Information Processing Systems (NIPS)*. 2013.

C. Shalizi. *Advanced Data Analysis from an Elementary Point of View*. 2016.

A. Smola, A. Gretton, L. Song, and B. Schölkop. A Hilbert Space Embedding for Distributions. In *Algorithmic Learning Theory: 18th International Conference*, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, 2000.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research (JMLR)*, 11:1297–1322, 2010.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

E.V. Strobl, S. Visweswaran, and K. Zhang. Approximate Kernel-based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *ArXiv e-prints*, 2017.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Uncertainty in Artificial Intelligence (UAI)*, 2011.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, Jan 2017.