Residual Similarity Based Conditional Independence Test and Its Application in Causal Discovery

Hao Zhang[1] **Kun Zhang**[2] **Shuigeng Zhou**[1]* **Jihong Guan**[3]

[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Department of Philosophy, Carnegie Mellon University, USA
[3]Department of Computer Science & Technology, Tongji University, China
{haoz15, sgzhou}@fudan.edu.cn; kunz1@cmu.edu; jhguan@tongji.edu.cn

Abstract

Recently, many regression based conditional independence (CI) test methods have been proposed to solve the problem of causal discovery. These methods provide alternatives to test CI of x, y given Z by first removing the information of the controlling set Z from x and y, and then testing the independence between the two residuals $R_{x,Z}$ and $R_{y,Z}$. When the residuals are linearly uncorrelated, the independence test between them is nontrivial. With the ability to calculate inner product in high-dimensional space, kernel-based methods are usually used to achieve this goal, but still consume considerable time. In this paper, we investigate the independence between two linear combinations under linear non-Gaussian structural equation model. We show that the dependence between the two residuals can be captured by the difference between the similarity of $(R_{x,Z}, R_{y,Z})$ and that of $(R_{x,Z}, R_r)$ (R_r is an independent copy of $R_{y,Z}$) in high-dimensional space. With this result, we design a new method called SCIT for CI test, where permutation test is performed to control Type I error rate. The proposed method is simpler yet more efficient and effective than the existing ones. When applied to causal discovery, the proposed method outperforms the counterparts in terms of both speed and Type II error rate, especially in the case of small sample size, which is validated by our extensive experiments on various datasets.

Introduction

Independence and conditional independence (CI) are central notions in statistical model building, as well as being a foundational concept for much of statistical theory. In the problem of causal discovery, independence and CI tests are usually used for testing CIs among variables. In constraint-based methods (Pearl and Mackenzie 2018), the CI relationship $x \perp\!\!\!\perp y|Z$ allows us to separate x–y when constructing a probabilistic model based on the joint distribution, which results in a parsimonious representation (Zhang et al. 2011). By using CI tests, constraint-based methods (Pearl 2009) can generally return a partial directed acyclic graph (DAG) (Pearl 2009). In the causal functional model (Velikova et al. 2014; Peters et al. 2012; Zhang et al. 2016), there is a solution to infer causal directions by testing the independence between the set of independent variables x and the corresponding residual $R_{x \to y}$ (or the causal process of $P(y|x)$).

Without given any assumption or precondition, CI testing is generally more difficult than independence testing. Many existing methods are based on explicit estimation of conditional densities or their variants such as ranks, kernel, copulas, nearest neighbours and discretizing-based methods (Diakonikolas and Kane 2016). For example, the characterization of CI of $P_{x|yZ}$=$P_{x|Z}$ can be used to test CI by measuring the distance between two conditional densities (Su and White 2008). Due to the curse of dimensionality, inevitably the required sample size increases dramatically with the size of controlling set Z, which makes accurate estimation of conditional density or related quantity hard to be accomplished. Assume that Z contains only one variable with a finite number of values $\{z_1, ..., z_k\}$, then $x \perp\!\!\!\perp y|Z$ iff $x \perp\!\!\!\perp y|Z$=z_i for each value z_i. Given a sample of size n, even if the data are distributed evenly on the values of Z, we must ensure that the independence within each subset of the sample with the same Z value by using only approximately n/k data points in each subset. When Z is continuously distributed or contains several variables, the observed values of Z are almost surely unique. To extend the above procedure to the continuous cases, we must consider the neighboring values of Z. However, it is also difficult for us to find appropriate neighboring points.

As kernel functions are able to represent high order moments by calculating similarity of high-dimensional implicit functions, a series of kernel-based CI tests were presented to solve the above problems. In practice, mapping variables into reproducing kernel Hilbert spaces (RKHSs) allows us to infer properties of distributions like independence (Gretton et al. 2006). Fukumizu et al. (2007) tested conditional covariance by using Hilbert-Schmidt norm of conditional cross-covariance operator, as the zero operator norm is equivalent to $x \perp\!\!\!\perp y|Z$ when the RKHSs are characteristic kernels. Daudin (1980) presented a characterization of CI that transforms CI to a set of zero correlations of regression functions. Concretely, $x \perp\!\!\!\perp y|Z$ if and only if for all $\psi \in L^2_{xZ}$ and $\phi \in L^2_y$ (L^2_{xZ} and L^2_y denote the spaces of square integrable functions of (x, Z) and y, respectively) such that $\mathbb{E}(\tilde{\psi}\tilde{\phi}) = 0$ where $\tilde{\psi}(x, Z) = \psi(x, Z) - r_\psi(Z)$ and $\tilde{\phi}(y, Z) = \phi(y) - r_\phi(Z)$, where $r_\psi, r_\phi \in L^2_Z$ are regression functions. To give an empirical estimate of this characterization of CI, Zhang et al. (2011) developed a method called KCIT, which relaxes the spaces of functions ψ, ϕ, r_ψ and r_ϕ to RKHSs. (Doran et al. 2014) intro-

*Corresponding author.

duced the PKCIT method that utilizes permutation to convert the CI test problem into an easier two-sample test problem. Strobl, Zhang, and Visweswaran (2017) used random Fourier features to approximate KCIT. Lee and Honavar (2017) employed a modified unbiased estimate of maximum mean discrepancy to measure CI. Compared to discretization-based CI testing methods, kernel methods exploit more complete information of the data and incur less random error.

Recently, regression-based tests were proposed for CI testing. Generally, these methods can be divided into two steps, regression and independence test. An indispensable assumption used by these methods is that any information of the controlling set Z can be removed from x and y by regression. As we know that this assumption is not always true but it works well in general continuous cases. Especially, when the information of Z can be totally removed from x and y by regression, regression-based CI tests generally works better than kernel-based methods. Grosse-Wentrup et al. (2016) transformed the CI of $x \perp\!\!\!\perp y|Z$ to independence between $x - \psi(Z)$ and (y, Z). (Zhang et al. 2017) used $x - \psi(Z) \perp\!\!\!\perp (y - \phi(Z), Z)$ to test $x \perp\!\!\!\perp y|Z$. In the two methods, ψ (or ϕ) is obtained by regressing x (or y) on Z, then CI test can be reduced to a set of regression and independence tests. In practice, $x - \psi(Z) \perp\!\!\!\perp Z$ is a strong condition, as $x - \mathbb{E}(x|Z) \perp\!\!\!\perp Z \Rightarrow Z$ causes x in many cases (Zhang and Hyvärinen 2009). On the other side, when Z contains several variables, checking whether or not $P(x - \psi(Z))$ is independent from the joint distribution $P(y, Z)$ or $P(y - \phi(Z), Z)$ tends to be prohibitively expensive. Note that in the two methods, independent-residuals is just sufficient but not necessary to meet CI. Flaxman, Neill, and Smola (2016) showed that given structural faithfulness and Markov assumptions (Pearl 2009), if Z causes x or y, $x \perp\!\!\!\perp y|Z$ is equivalent to $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. Similarly, here a strong condition that Z causes x or y is assumed, hence it is easy to derive the corresponding causal relations. Moreover, faithfulness condition means that $x \perp\!\!\!\perp y|Z \Rightarrow x$ and y are d-separated by Z, and Markov condition implies that y are d-separated by $Z \Rightarrow x \perp\!\!\!\perp y|Z$, so CI is relaxed to d-separation given the faithfulness and Markov assumptions. However, CI is neither sufficient nor necessary to d-separation. In practice, given the faithfulness assumption, $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ and $x \perp\!\!\!\perp y|Z$ have significant correlations. For example, in (Ramsey 2014), the authors suggested to use $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ to test $x \perp\!\!\!\perp y|Z$ under the faithfulness assumption. In (Zhang et al. 2017), the authors further conjectured that $x - \psi(Z) \perp\!\!\!\perp y - \phi(Z)$ can lead to $x \perp\!\!\!\perp y|Z$ under nonlinear and faithfulness conditions, where ψ and ϕ are nonlinear functions, x, y and Z are generated by nonlinear additive noise model. Zhang, Zhou, and Guan (2018) showed that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ is sufficient to support $x \perp\!\!\!\perp y|Z$ if the data is generated by following the linear non-Gaussian structural equation model (SEM) under the faithfulness assumption. As the residuals can be easily calculated by linear regression, the performance mainly depends on the independence test. Note that in this case, $cov(x - \mathbb{E}(x|Z), y - \mathbb{E}(y|Z)) = 0$ often holds. Therefore, it is difficult to detect the common component shared by $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$. To get the best performance, this method (denoted by ReCIT) uses KCIT to achieve this goal, but it is computationally rather demanding. In (Zhang et al. 2021), the authors used kurtosis to test independence, they proved that in linear case with $x \not\perp\!\!\!\perp y$, $x - h * y$ and $x - h * r$ have different kurtosis, where h is the number of interpolation points, k is the times of permutations. This methods works very efficient in simple case. However, when the scenario becomes complicated with two residuals being very Gaussian, it is easy to cause Type II error where the CI hypothesis is not rejected although it is false.

In this work, we aim to test the independence between the two residuals $R_{x,Z} = x - \mathbb{E}(x|Z)$ and $R_{y,Z} = y - \mathbb{E}(y|Z)$ returned by linear regression, where $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$, $z_j = \sum_{i=1}^{l} c_i s_i$ $(\forall z_j \in Z)$ and $s_{1,\ldots,l}$ are noise mutually independent. We show that the dependence between residuals can be captured by the difference between the similarity of $(R_{x,Z}, R_{y,Z})$ and that of $(R_{x,Z}, R_r)$ where R_r is an independent copy of $R_{y,Z}$ in high-dimensional space, denoted by $S[\psi(R_{x,Z}), \psi(R_{y,Z})]$ and $S[\psi(R_{x,Z}), \psi(R_r)]$. We design an elaborate test criterion for measuring the difference between the two $S[*]$, by kernel and permutation based methods. The proposed method needs to calculate $n \times 1$ similarity matrix instead of the trace of product between two $n \times n$ matrices, therefore it works more efficient. Extensive experiments show that our method performs better on regression based CI test than the counterparts, which can work faster and get a better performance in causal discovery.

Similarity Based CI Test

In this work, we assume that the given variables are generated by the linear non-Gaussian structural equation model (SEM), which is defined as a tuple $(S, P(X))$ where $S = \{S_1, ..., S_n\}$ is a collection of n equations, $S_i : x_i = \sum pa_{x_i} + \varepsilon_i$, $i = \{1, ..., n\}$ and pa_{x_i} corresponds to the set of direct parents of x_i in a DAG G. The noise variables ε_i have a strictly positive density with respect to the Lebesgue measure and are independent, all of them have the same non-Gaussian distribution. SEM reflects the data-generating processes of X in G. We say a SEM is identifiable if it is asymmetrical in cause and effect and is able to distinguish between them. In fact, linear SEM is generally identifiable in non-Gaussian cases (Zhang and Hyvärinen 2009).

Regression Based CI Test

Consider the task as follows: given two randomly selected nodes x' and y', we want to test whether x' and y' are conditionally independent given a set of variables Z. According to the mechanism of regression based CI test, the CI test of $x' \perp\!\!\!\perp y'|Z$ can be relaxed to an independence test between two residuals $x = x' - \mathbb{E}(x'|Z)$ and $y = y' - \mathbb{E}(y'|Z)$ in the linear non-Gaussian case. As the residuals x and y can be easily calculated by linear regression, the task turns to testing the independence between x and y. Concretely, the two variables (residuals) x and y are linear combinations of independent noise s_i $(i = 1, ..., l)$ such that $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$. When x and y are correlated, we know $x \not\perp\!\!\!\perp y$ holds. However, if x and y are uncorrelated, then it is difficult to check whether x and y are independent or not. In what follows, we try to develop a low complexity method (compared to kernel-based methods) to test independence between two residuals.

We know the mutual information of x and y is

$$I(x,y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1}$$

Then, $I(x,y) = 0$ implies $p(x,y) = p(x)p(y)$, i.e., x and y are independent. In continuous case, we need to use some methods to measure probability density. We first review how the existing methods based on Hilbert-Schmidt independence criterion (HSIC) work. Lets consider $p(x,y) = p(x)p(y)$, then for any ψ and ϕ, the square integrable function of x and y, respectively, we have

$$\begin{aligned}
&C[\psi,\phi] \\
=&\iint p(x,y)\psi(x)\phi(y)dxdy - \iint p(x)p(y)\psi(x)\phi(y)dxdy \\
=&\mathbb{E}_{x,y\sim p(x,y)}[\psi(x)\phi(y)] - \mathbb{E}_{x\sim p(x)}[\psi(x)]\mathbb{E}_{y\sim p(y)}[\phi(y)] \\
=&0
\end{aligned} \tag{2}$$

Therefore, to solve this problem, we need to select enough ψ and ϕ to see how close $C[\psi,\phi]$ is to zero. Consider

$$L = \sum_{\psi,\phi} (C[\psi,\phi])^2 \tag{3}$$

if $p(x,y) = p(x)p(y)$ holds, $L = 0$. We calculate $(C[\psi,\phi])^2$,

$$\begin{aligned}
&(C[\psi,\phi])^2 \\
=&\mathbb{E}_{x_1,y_1\sim p(x,y),x_2,y_2\sim p(x,y)}[\psi(x_1)\psi(x_2)\phi(y_1)\phi(y_2)] \\
&+ \mathbb{E}_{x_1\sim p(x),x_2\sim p(x),y_1\sim p(y),y_2\sim p(y)}[\psi(x_1)\psi(x_2)\phi(y_1)\phi(y_2)] \\
&- 2\mathbb{E}_{x_1,y_1\sim p(x,y),x_2\sim p(x),y_2\sim p(y)}[\psi(x_1)\psi(x_2)\phi(y_1)\phi(y_2)]
\end{aligned} \tag{4}$$

At this time, we can use kernel function to calculate the inner product between any ψ and ϕ, then L can be changed to

$$\begin{aligned}
L =&\mathbb{E}_{x_1,y_1\sim p(x,y),x_2,y_2\sim p(x,y)}[K(x_1,x_2)K(y_1,y_2)] \\
&+ \mathbb{E}_{x_1\sim p(x),x_2\sim p(x),y_1\sim p(y),y_2\sim p(y)}[K(x_1,x_2)K(y_1,y_2)] \\
&- 2\mathbb{E}_{x_1,y_1\sim p(x,y),x_2\sim p(x),y_2\sim p(y)}[K(x_1,x_2)K(y_1,y_2)]
\end{aligned} \tag{5}$$

We need to multiply $n \times n$ kernel matrices, computing L is expensive with complexity $O(n^3)$, n being the sample size.

CI Test Criterion Based on Similarity

In this section, we present a method for test CI based on similarity. Consider three variables, x, y and r, where r is an independent copy of y, that is y and r are independent and identically distributed, $r{\sim}p(y)$. Intuitively, if x and y are independent, then the similarity between $\psi(x)$ and $\psi(y)$ equals to that between $\psi(x)$ and $\psi(r)$, denoted by $S[\psi(x),\psi(y)] = S[\psi(x),\psi(r)]$, where ψ is any square integrable function of x, y and r. On the contrary, there must be some ψ such that $S[\psi(x),\psi(y)] \neq S[\psi(x),\psi(r)]$ if x and y are not independent. Therefore, we can derive the following theoretical result,

Proposition 1. *Given three random variables x, y and r, where r is an independent copy of y, if $x \perp\!\!\!\perp y$, then $\forall\psi$, $C[\psi(x),\psi(y),\psi(r)] = 0$; if $x \not\!\perp\!\!\!\perp y$, then almost surely $\exists\psi$ such that $C[\psi(x),\psi(y),\psi(r)] \neq 0$, where*

$$\begin{aligned}
&C[\psi(x),\psi(y),\psi(r)] \\
&= \iint p(x,y)S[\psi(x),\psi(y)]dxdy \\
&\quad - \iint p(x)p(r)S[\psi(x),\psi(r)]dxdr \\
&= \iiint p(x,y)p(r)(S[\psi(x),\psi(y)] - S[\psi(x),\psi(r)])dxdydr \\
&=\mathbb{E}_{x,y\sim p(x,y),r\sim p(y)}(S[\psi(x),\psi(y)] - S[\psi(x),\psi(r)]).
\end{aligned} \tag{6}$$

We therefore can derive a test criterion by following Equ. (2)-(6) as

$$L_{xyr} = \sum_{\psi} (C[\psi(x),\psi(y),\psi(r)])^2 \tag{7}$$

in which

$$\begin{aligned}
&(C[\psi(x),\psi(y),\psi(r)])^2 \\
=&\mathbb{E}_{x,y\sim p(x,y)}(S[\psi(x),\psi(y)])^2 + \mathbb{E}_{x\sim p(x),r\sim p(y)}(S[\psi(x),\psi(r)])^2 \\
&- 2\mathbb{E}_{x,y\sim p(x,y)}(S[\psi(x),\psi(y)])\mathbb{E}_{x\sim p(x),r\sim p(y)}(S[\psi(x),\psi(r)])
\end{aligned} \tag{8}$$

We need to measure how close is L_{xyr} to zero. Next step, we use kernel function to calculate the inner product between $\psi(x)$ and $\psi(y)$ or $\psi(r)$ w.r.t. a set of $\psi(*)$, then

$$\begin{aligned}
L_{xyr} =& \sum_{\psi} (C[\psi(x),\psi(y),\psi(r)])^2 \\
&=\mathbb{E}_{x,y\sim p(x,y)}(K(x,y))^2 + \mathbb{E}_{x\sim p(x),r\sim p(y)}(K(x,r))^2 \\
&\quad - 2\mathbb{E}_{x,y\sim p(x,y)}(K(x,y))\mathbb{E}_{x\sim p(x),r\sim p(y)}(K(x,r))
\end{aligned} \tag{9}$$

Note that, the data we have is a finite sample $(x,y) = ((x_1,y_1),(x_2,y_2),...,(x_n,y_n))$ from a pair of variables x and y. Here we use permutation method to test the hypothesis of independence: H_0: x and y are independent, versus H_1: x and y are not independent. The idea behind it is that permuting y removes any dependency between x and y. Therefore we can compare L_{xyr} with L_{xy_pr}, where y_p is the permutation p applied to the sample y. We choose the number of permutations k, and create k permuted samples y_{p_i}, $i = 1,...,k$. Then if x and y are truly independent, permuting y will not change much L_{xyr}, therefore we will not be able to reject the null hypothesis (x and y are independent). On the other hand, if x and y were not independent, we can reject the H_0 with greatly changed L_{xyr}.

In practice, permutation tests are particularly attractive because of their simplicity and their ability to control Type I error without any distributional assumptions (Berrett et al. 2020). Recall that our task is given two variables x and y, test whether x and y are conditionally independent given a set of variables Z. We use least square method to regress x on Z, and denote the obtained residual by $R_{x,Z} = x - \mathbb{E}(x|Z) = x - Z(Z^TZ)^{-1}Z^Tx$. Similarly, we can get the residual of regressing y on Z, $R_{y,Z} = y - \mathbb{E}(y|Z) = y - Z(Z^TZ)^{-1}Z^Ty$. By subtracting the two residuals, we obtain

$$\begin{aligned}
R_{x,Z} - R_{y,Z} =&R_{x-y,Z} \\
=&(x-y) - Z(Z^TZ)^{-1}Z^T(x-y) \\
=&(1-M)(x-y)
\end{aligned} \tag{10}$$

where we simply denote the matrix $Z(Z^TZ)^{-1}Z^T$ by M. Suppose the Gaussian Radial Basis Function (RBF) is used to calculate inner product, then

$$\begin{aligned}
L =&L_{R_{x,Z}R_{y,Z}R_r} = (C[\psi(R_{x,Z}), \psi(R_{y,Z}), \psi(R_r)])^2 \\
=&\mathbb{E}_{R_{x,Z},R_{y,Z}\sim p(R_{x,Z},R_{y,Z})}(K(x,y))^2 + \mathbb{E}_{R_{x,Z}\sim p(R_{x,Z}),R_r\sim p(R_{y,Z})}(K(x,r))^2 \\
&- 2\mathbb{E}_{R_{x,Z},R_{y,Z}\sim p(R_{x,Z},R_{y,Z})}(K(x,y))\mathbb{E}_{R_{x,Z}\sim p(R_{x,Z}),R_r\sim p(R_{y,Z})}(K(x,r)) \\
=&\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|(1-M)(x-y)\|^2))^2 \\
&+ \mathbb{E}_{x,y,Z\sim p(x,y,Z),r\sim p(y)}(exp(-\gamma\|x - Mx - r + M_{p_r}y\|^2))^2 \\
&- 2\mathbb{E}(*)\mathbb{E}(*)
\end{aligned} \tag{11}$$

where M_{p_r} is the permutation p_r applied to M on row, as

$$R_r =(R_{y,Z})_{p_r} = (y - My)_{p_r} = y_{p_r} - M_{p_r}y = r - M_{p_r}y \tag{12}$$

Consider the test criterion with permutation p

$$\begin{aligned}
L_p =&L_{R_{x,Z}(R_{y,Z})_pR_r} = (C[\psi(R_{x,Z}), \psi((R_{y,Z})_p), \psi(R_r)])^2 \\
=&\mathbb{E}_{x,Z\sim p(x,Z),y_p\sim p(y)}(exp(-\gamma\|x - Mx - y_p + M_py\|^2))^2 \\
&+ \mathbb{E}_{x,y,Z\sim p(x,y,Z),r\sim p(y)}(exp(-\gamma\|x - Mx - r + M_{p_r}y\|^2))^2 \\
&- 2\mathbb{E}(*)\mathbb{E}(*)
\end{aligned} \tag{13}$$

We can see $+\mathbb{E}_{x,y,Z\sim p(x,y,Z),r\sim p(y)}(exp(-\gamma\|x-Mx-r+M_{p_r}y\|^2))^2$ simultaneously exists in L and L_{p_i}, therefore this term can be removed, i.e.,

$$\begin{aligned}
L =&\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|(1-M)(x-y)\|^2))^2 \\
&- 2\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|(1-M)(x-y)\|^2)) \\
&\times \mathbb{E}_{x,y,Z\sim p(x,y,Z),r\sim p(y)}(exp(-\gamma\|x - Mx - r + M_{p_r}y\|^2))
\end{aligned} \tag{14}$$

and

$$\begin{aligned}
L_{p_i} =&\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|x - Mx - y_{p_i} + M_{p_i}y\|^2))^2 \\
&- 2\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|x - Mx - y_{p_i} + M_{p_i}y\|^2)) \\
&\times \mathbb{E}_{x,y,Z\sim p(x,y,Z),r\sim p(y)}(exp(-\gamma\|x - Mx - r + M_{p_r}y\|^2))
\end{aligned} \tag{15}$$

Recall that

$$\begin{aligned}
&C[\psi(R_{x,Z}), \psi(R_{y,Z}), \psi(R_r)] \\
=&\mathbb{E}_{R_{x,Z},R_{y,Z}\sim p(R_{x,Z},R_{y,Z}),R_r\sim p(R_{y,Z})}(S[\psi(R_{x,Z}), \psi(R_{y,Z})] \\
&- S[\psi(R_{x,Z}), \psi(R_r)]) = \mathbb{E}(S(A) - S(B))
\end{aligned} \tag{16}$$

$$C[\psi(R_{x,Z}), \psi((R_{y,Z})_p), \psi(R_r)] = \mathbb{E}(S(C) - S(B)) \tag{17}$$

As Equ. (16) and Equ. (17) share the term of $S(B)$, L and L_{p_i} can be simply reduced to

$$L =\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|(1-M)(x-y)\|^2)) \tag{18}$$

and

$$L_{p_i} =\mathbb{E}_{x,y,Z\sim p(x,y,Z)}(exp(-\gamma\|x - Mx - y_{p_i} + M_{p_i}y\|^2)) \tag{19}$$

Assume $i = 1, ..., k$, then P-value can be defined as

$$P - value = \frac{\sum_i \mathbf{1}\{L < L_{p_i}\}}{k} \tag{20}$$

where $\mathbf{1}$ is indicator function. Then, given a significant value α, if P-value$\geq \alpha$, we accept H_0: x and y are independent, otherwise accept H_1: x and y are not independent.

Implementation of Similarity Based CI Testing

As mentioned above, the difference between L and L_{p_i} can be used to test CI. With these theoretical results, we design a new method for CI testing called ***Similarity based Conditional Independence Test*** (SCIT in short). The details of SCIT are given in Alg. 1. To test the CI of $x \perp\!\!\!\perp y|Z$, we first apply $k + 1$ different permutations to y and obtain $k + 1$ permuted samples of y, the new variables are denoted by r and $y_{p_1}, ..., y_{p_k}$ (Line 1). Then, we calculate $k + 1$ statistics L and $L_{p_1}, ..., L_{p_k}$ according to Equ. (18) and Equ. (19). In this process, the time is mainly spent on calculate $M = Z(Z^TZ)^{-1}Z^T$, which contains an inverse operation of matrix Z^TZ. The matrix M_{p_i} can be easily obtained from M and permutation p_i (Line 2). In the final step, we calculate the P-value $= \sum_i(L < L_{p_i})/k$. If P-value$\geq \alpha$, we accept H_0: $x \perp\!\!\!\perp y|Z$, otherwise accept H_1: $x \not\perp\!\!\!\perp y|Z$ (Lines 3-8).

Algorithm 1 Similarity based conditional independence test (SCIT)

```
Input: variables: $x, y, Z$; the number of counterparts: $k$; significant value $\alpha$.
Output: accept $H_0$: $x \perp\!\!\!\perp y|Z$ or $H_1$: $x \not\perp\!\!\!\perp y|Z$.
1: create $k+1$ permuted samples of $y$, the new variables are denoted by $r$ and $y_{p_1}, ..., y_{p_k}$.
2: calculate $k + 1$ statistics $L$ and $L_{p_1}, ..., L_{p_k}$ according to Equ. (18) and Equ. (19).
3: calculate $P$-value = $\frac{\sum_i \mathbf{1}\{L<L_{p_i}\}}{k}$
4: if $P$-value$\geq \alpha$ then
5:     accept $H_0$: $x \perp\!\!\!\perp y|Z$.
6: else
7:     accept $H_1$: $x \not\perp\!\!\!\perp y|Z$.
8: end if
```

As SCIT is used for linear CI testing, therefore SCIT can be directly applied to the PC algorithm for linear causality discovery. For more details about using regression based CI test in the PC algorithm, the readers can refer to (Zhang, Zhou, and Guan 2018).

Discussion

Go back to Equ. (7), if the similarity $S(\cdot)$ is measured by using Pearson correlation coefficient $Corr(\cdot)$, then

$$\begin{aligned}
&\sum_\psi(C[\psi(x), \psi(y), \psi(r)])^2 \neq 0 \\
\Leftrightarrow &\sum_\psi(\mathbb{E}_{x,y\sim p(x,y)}(S[\psi(x), \psi(y)])^2 + \mathbb{E}_{x\sim p(x),r\sim p(y)}(S[\psi(x), \psi(r)])^2 \\
&- 2\mathbb{E}_{x,y\sim p(x,y)}(S[\psi(x), \psi(y)])\mathbb{E}_{x\sim p(x),r\sim p(y)}(S[\psi(x), \psi(r)])) \neq 0 \\
\Leftrightarrow &\sum_\psi \mathbb{E}_{x,y\sim p(x,y)}(Corr(\psi(x), \psi(y))^2 \neq 0 \Rightarrow x \not\perp\!\!\!\perp y.
\end{aligned} \tag{21}$$

Contrast to Daudin's work on characterization of CI (Daudin 1980), SCIT searches for only one function ψ, which means that Equ. (7) is sufficient but not necessary to support CI. But in practice, by assuming that any ψ can be covered by SCIT with a family of kernel functions, only in well-designed situations where counterexamples will be found.

Performance Evaluation

We first compare SCIT with ReCIT (Zhang et al. 2019) and KCIT (Zhang et al. 2011) by extensive simulated experiments, in which SCIT, ReCIT are residual-based CI test methods, ReCIT tests the independence between two residuals by using HSIC/KCIT. To the best of our knowledge, ReCIT is one of the best residual-based CI testing methods in linear cases, there are many comparisons between ReCIT and other CI testing methods like KCIT presented in the previous works (Zhang et al. 2017; Zhang, Zhou, and Guan 2018). We then illustrate the advantage of SCIT in causal skeleton learning. We compare our method (SCIT + PC algorithm) with the causal learning method PC_{ReCIT} over various causal graphs. The experimental platform adopts Matlab R2021b, Intel i7-11700K (3.60 GHz) CPU, Windows 10, and 32G memory. The source code of SCIT package is available at https://github.com/Causality-Inference/SCIT.

Effect of Controlling Set and Sample Sizes

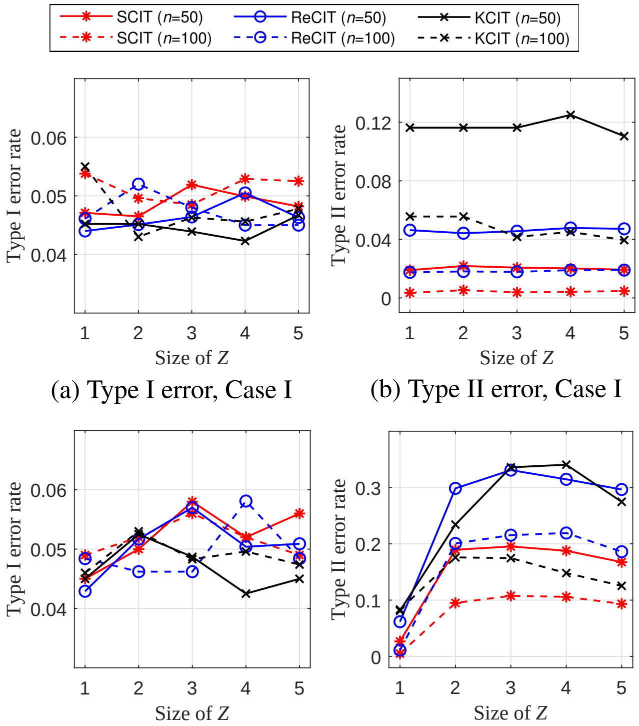
As we know, CI test methods are mainly affected by the size of the controlling set and the sample size, therefore we aim to examine how the probabilities of Type I error (where the CI hypothesis H_0 is incorrectly rejected) and Type II error (where the CI hypothesis is not rejected although it is false) errors of SCIT change with the size of the conditioning set Z ($|Z| = 1, 2, ..., 5$, respectively) by simulation. Here, we consider two cases as follows.

In Case I, only one variable in Z, denoted by z_1, is effective, i.e., the other variables are independent of x, y, and z_1. The causal link is $x \rightarrow z_1 \rightarrow y$, in which $z_1 = a * x + \varepsilon_x$, $y = b * z_1 + \varepsilon_y$. The other variables $x, z_2, ..., z_5$ are independently generated by following $U(-1, 1)$, $\varepsilon_x, \varepsilon_y \sim U(-0.2, 0.2)$ and $a, b \sim U(0.2, 1)$. The ground truth is $x \perp\!\!\!\perp y | z_1 \cup S$ and $x \not\perp\!\!\!\perp y | S$, where $\forall S \subseteq Z \backslash_{z_1}$.

In Case II, all variables in the conditioning set Z are effective in generating X and Y. The causal link is $x \rightarrow Z \rightarrow y$, in which $z_i = a_i * x + \varepsilon_i$ and $y = \sum_i b_i * z_i + \varepsilon_y$. The setting of coefficients a_i, b_i and noise terms $\varepsilon_i, \varepsilon_y$ are similar to those in Case 1. The ground truth is $x \perp\!\!\!\perp y | Z$ and $x \not\perp\!\!\!\perp y | S$ where $\forall S \subset Z$.

Recall that, the residuals can be easily recovered by linear regression in this simple setting. To evaluate the robustness of these methods, here we do not want the returned residuals being very accurate. Therefore we test CI with small sample size of 50 and 100. The significance levels are fixed at $\alpha = 0.05$. Note that for a good testing method, the probability of Type I error should be as close to the significance level as possible, and the probability of Type II error should be as small as possible. We check how the errors change when increasing the dimensionality of Z and the sample size n. For each parameter setting, we randomly repeat the testing 100 times and average their results.

Type I and II errors are calculated like this: take $|Z| = 3$ for example, in Case I, x is independent of y given (z_1), (z_1, z_2), (z_1, z_3) and (z_1, z_2, z_3), then Type I error rate =1-$the\ number\ of\ CIs/4$. On the other side, x is not independent of y given $\emptyset$, (z_2), (z_3) and (z_2, z_3), then Type II error rate = $the\ number\ of\ CIs/4$. Similarly, we can calculate Type I and II error rate in Case II.



(a) Type I error, Case I

(b) Type II error, Case I

(a) Type I error, Case II

(b) Type II error, Case II

Figure 1: (a) Type I and (b) Type II error rate in Case I; (c) Type I and (d) Type II error rate in Case II.

The results are presented in Fig. 1. We can see that

1. As shown in Fig. 1(a) and (c), the Type I error rate of SCIT is close to the significance level $\alpha = 0.05$ (between 0.04 - 0.06). This because we use permutation test to control Type I error rate;
2. As shown in Fig. 1(b), the Type II error rate of each method keeps stable when crossing $|Z| = \{1, 2, ..., 5\}$. The reason is that only one variable z_1 is effective in Case I, then the probability of rejecting CIs of $x \perp\!\!\!\perp y | z_1 \cup S$ for any $S \subseteq Z \backslash_{z_1}$ would be very close. We can see that SCIT outperforms the other two methods in terms of Type II error rate;
3. As shown in Fig. 1(d), the Type II error rate of each method increases/changes with different sizes of Z. This is because all the variables $z_1, ..., z_5$ are effective in Case II, then the probability of rejecting CIs of $x \perp\!\!\!\perp y | S$ for different $S \subseteq Z$ would be various. In this case, SCIT also achieves the best performance;
4. Increasing sample size can significantly reduce the Type II error rate, while Type I error rate is generally not impacted by sample size.

Efficiency Comparison

We compare the efficiency of SCIT, ReCIT and KCIT in terms of elapsed time with the sample size increasing from 50 to 1000. As presented in Table 1, SCIT takes significantly less time than the other methods. Recall that the same linear

Sample	Elapsed time (s)		
	SCIT	ReCIT	KCIT
50	**0.001** ± 0.0001	**0.001** ± 0.0001	0.16 ± 0.03
100	**0.001** ± 0.0001	0.002 ± 0.0003	0.25 ± 0.05
200	**0.002** ± 0.0001	0.005 ± 0.0012	0.45 ± 0.11
500	**0.021** ± 0.0055	0.042 ± 0.0053	3.67 ± 0.75
1000	**0.079** ± 0.0167	0.205 ± 0.0208	19.8 ± 5.06

Table 1: Efficiency comparison of SCIT, ReCIT and KCIT.

regression progress is performed in SCIT, ReCIT, therefore the time-consuming difference among them depends on the respective unconditional independence test methods. SCIT is evidently faster, as it only needs to calculate similarity vectors, while ReCIT needs to calculate the trace of product of two $n \times n$ matrices.

Performance on Small Graphs

In this section, we evaluate SCIT, ReCIT and KCIT in more complex scenarios. We generate data from a set of random DAGs. For each DAG G, we first create four nodes $v_1, ..., v_4$, and with probability 50% each possible edge is either present or absent, and orient arrow between them from v_i to v_j only for $i < j$. Then, each variable x_i corresponding to each root node in G is generated by following $U(-1, 1)$ and each variable x_i corresponding to leaf node is generated by $\sum_i a_i \cdot pa_{x_i} + \varepsilon$ where $a_i \sim U(0.2, 1)$ and $\varepsilon \sim U(-0.2, 0.2)$ independent across pa_{x_i}. For significance level 0.05 and sample sizes from 25 and 200, we simulate 100 DAGs and evaluate the performance of the three methods PC$_{SCIT}$, PC$_{ReCIT}$ and PC$_{KCIT}$ on discovering causal skeletons.

As shown in Fig. 2, we can see that when the sample size is small (e.g. less than 50), PC$_{SCIT}$ performs significantly better than other two methods. As the sample size increases, the performance of PC$_{SCIT}$ close to that of PC$_{ReCIT}$ and PC$_{KCIT}$. When the sample size up to 150, the *Recall*, *Precision* and $F1$ curves of the three methods tend to be overlapping. Therefore, PC$_{SCIT}$ performs significantly better in CI test in causal discovery when the sample size is small, which is the frequently-encountered case in reality.

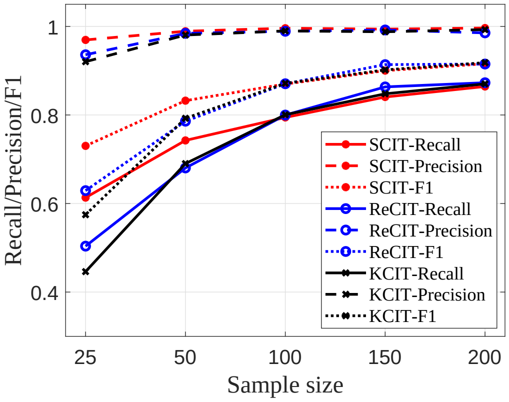


Figure 2: Performance comparison among PC$_{SCIT}$, PC$_{ReCIT}$ and PC$_{KCIT}$ with various sample sizes on causal skeleton learning.

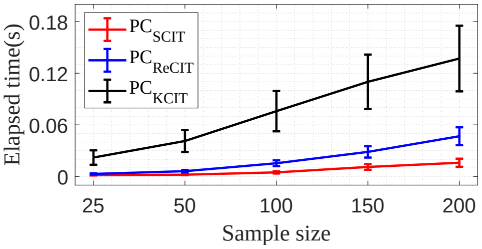


Figure 3: Efficiency comparison with various sample sizes. The elapsed time of PC$_{KCIT}$ is divided by 30, otherwise it is difficult to see the gap between PC$_{SCIT}$ and PC$_{ReCIT}$.

Fig. 3 shows the elapsed time of PC$_{SCIT}$, PC$_{ReCIT}$ and PC$_{FRCIT}$ with the sample size increasing from 25 to 200, it is consistent with the result present in Table 1. SCIT can be very efficient to test CI in causal discovery with small sample size ($n \leq 1000$).

Performance on Causal Discovery

In the experiments above, we compare SCIT and ReCIT in terms of learning causal skeletons of small DAGs, the result shows the two methods have almost the same accuracy when sample size is more than 100, though SCIT works much more efficient than the others. In this section, the two methods will be evaluated on six causal graphs [1] that cover a variety of applications, including biomedicine (*Cancer* and *Asia*), expert systems (*Child*), insurance evaluation (*Insurance*), medicine (*Alarm*) and agricultural industry (*Barley*). The structural statistics of these causal networks are summarized in Table 2.

To obtain the precise ground truth in every cases, the corresponding data-generating process follows the previous works (Cai, Zhang, and Hao 2013, 2017). As the residuals can be easily recovered by linear regression with enough samples, to evaluate the robustness of these methods, here we test CI with small sample size of 25 and 200. In causal discovery, partially correlation tests (Baba, Shibata, and Sibuya 2004) are often used to speed up CI tests based on the criterion: $pcorr(x, y|Z) \neq 0 \Rightarrow x \not\perp y|Z$. In order to evaluate these methods independently, here we do not perform any partially correlation test.

Table 2: Statistics of six causal graphs

Dataset	#Nodes	#Arcs	Max in-degree
Cancer	5	4	2
Asia	8	8	2
Child	20	25	2
Insurance	27	52	3
Alarm	37	46	4
Barley	48	84	4

[1] http://www.bnlearn.com/bnrepository/

Dataset	*Recall*		*Precision*		*F*1	
	PC$_{SCIT}$	PC$_{ReCIT}$	PC$_{SCIT}$	PC$_{ReCIT}$	PC$_{SCIT}$	PC$_{ReCIT}$
Cancer	**0.56** ± 0.17	0.40 ± 0.15	**0.94** ± 0.14	0.91 ± 0.21	**0.69** ± 0.15	0.54 ± 0.17
Asia	**0.62** ± 0.11	0.53 ± 0.12	0.92 ± 0.11	**0.95** ± 0.10	**0.74** ± 0.10	0.67 ± 0.11
Child	**0.51** ± 0.08	0.29 ± 0.08	0.82 ± 0.10	**0.86** ± 0.11	**0.62** ± 0.08	0.43 ± 0.09
Insurance	**0.33** ± 0.05	0.19 ± 0.04	0.68 ± 0.08	**0.76** ± 0.11	**0.44** ± 0.06	0.31 ± 0.06
Alarm	**0.39** ± 0.05	0.32 ± 0.04	0.78 ± 0.06	**0.82** ± 0.06	**0.52** ± 0.05	0.46 ± 0.05
Barley	**0.31** ± 0.03	0.23 ± 0.03	0.74 ± 0.05	**0.77** ± 0.07	**0.43** ± 0.04	0.36 ± 0.04

Table 3: Performance of PC$_{SCIT}$ and PC$_{ReCIT}$ with sample size = 25.

Dataset	*Recall*		*Precision*		*F*1	
	PC$_{SCIT}$	PC$_{ReCIT}$	PC$_{SCIT}$	PC$_{ReCIT}$	PC$_{SCIT}$	PC$_{ReCIT}$
Cancer	**0.93** ± 0.12	0.73 ± 0.18	0.94 ± 0.12	**0.95** ± 0.11	**0.93** ± 0.10	0.81 ± 0.13
Asia	**0.90** ± 0.05	0.81 ± 0.09	0.88 ± 0.10	**0.94** ± 0.07	**0.89** ± 0.06	0.87 ± 0.07
Child	**0.94** ± 0.03	0.88 ± 0.04	0.71 ± 0.05	**0.82** ± 0.06	0.81 ± 0.05	**0.85** ± 0.04
Insurance	**0.72** ± 0.05	0.63 ± 0.07	0.43 ± 0.05	**0.59** ± 0.03	0.54 ± 0.03	**0.61** ± 0.03
Alarm	**0.66** ± 0.05	0.62 ± 0.06	0.76 ± 0.03	**0.78** ± 0.04	**0.71** ± 0.04	0.69 ± 0.05
Barley	**0.51** ± 0.03	0.48 ± 0.03	**0.63** ± 0.02	**0.63** ± 0.04	**0.57** ± 0.02	0.55 ± 0.03

Table 4: Performance of PC$_{SCIT}$ and PC$_{ReCIT}$ with sample size = 200.

The results are shown in Table 3 and Table 4. One can see that the *Precision* is higher than the *Recall* in most cases. We know $Recall = \frac{Discovered\ edges \cap Actual\ edges}{Actual\ edges}$ and $Precision = \frac{Discovered\ edges \cap Actual\ edges}{Discovered\ edges}$, the Type I error occurred in SCIT and ReCIT would not affect PC($*$) much, that is because if Type I error occurs, the CI test will continue to test x and y given another controlling set Z. However, such a traversal search strategy will be greatly affected by Type II error. For example, assume that Type II error rate is r_i for each controlling set Z_i, then the rate of rejecting all CI hypothesis when they are really false is $\prod(1 - r_i)$, and we have

$$\lim_{k\to+\infty}\prod_{i=1,\ldots,k}(1 - r_i) = 0. \tag{22}$$

Therefore, the performance of constraint-based causal discovery is largely determined by the Type II error rate of CI tests. Compare Table 3 with Table 4, one can see that increasing samples can significantly reduce the Type II error rate, then improve the *Recall* of the two methods.

On the other side, we can see PC$_{SCIT}$ outperforms PC$_{ReCIT}$ in most of cases in terms of $F1$, although their *Precision* are very close to each other. As aforementioned, the Type I error occurred in SCIT and ReCIT would not affect PC($*$) much, therefore all of them obtain high *Precision*. Similarly, we can see that the *Recall* of PC$_{SCIT}$ is slightly better than that of PC$_{ReCIT}$. This result is consistent with the result presented in Fig. 1(b),(d) and Equ. (22), the lower the rate of Type II error, the higher the value of *Recall*. In addition, like the results presented in Table 1 and Fig.3, PC$_{SCIT}$ works much more efficient, it is very suitable for testing CI or discovering causalities in low-sample scenarios. Similar to the results presented in Fig.2, their accuracy will become very close given sufficient samples, and PC$_{SCIT}$ will lose its advantage on accuracy. Here we mainly consider the case of small sample size which is the most significant advantage of SCIT.

Conclusion

In this paper, we propose a new and fast residual similarity based conditional independence (CI) test method, called SCIT, to support effective and efficient causality discovery under the linear structural equation model (SEM) with non-Gaussian noise variables. Concretely, we provide a simple way to test the independence between two residuals $R_{x,Z}$=x-$\mathbb{E}(x|Z)$ and $R_{y,Z}$=y-$\mathbb{E}(y|Z)$ returned by linear regression. We show that the dependence between residuals can be captured by the difference between the similarity $S\,[\psi(R_{x,Z}), \psi(R_{y,Z})]$ and the similarity $S\,[\psi(R_{y,Z}), \psi(r)]$ given a set of square integrable functions ψ. Then kernel functions are used to calculate the inner product, i.e., similarity. As the value of similarity is not scale-free, we simply use permutation test to get the *P-value* to accept or reject CI hypothesis. Our theoretical analysis proves the correctness of the proposed method, and extensive experiments verify the advantage of SCIT.

As mentioned in Proposition 1, the dissimilarity of X and Y is measured by using only one function ψ, SCIT is sufficient but not necessary to support CI. But in practice, by assuming that any ψ can be covered by SCIT with a family of kernel functions, only in well-designed situations where counterexamples will be found.

Acknowledgement

This work was supported by National Natural Science Foundation (NSFC) (62006051 and U1936205). Hao Zhang was also supported by the Project of Young Innovative Talents in Colleges and Universities in Guangdong Province (2020KQNCX049). Shuigeng Zhou was also partially supported by 2019 Special Fund for Artificial Intelligence Innovation & Development, Shanghai Economy and Information Technology Commission (SHEITC).

References

Baba, K.; Shibata, R.; and Sibuya, M. 2004. Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, 46(4): 657–664.

Berrett, T. B.; Wang, Y.; Barber, R. F.; and Samworth, R. J. 2020. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B*, 82.

Cai, R.; Zhang, Z.; and Hao, Z. 2013. Sada: A general framework to support robust causation discovery. In *International Conference on Machine Learning*, 208–216.

Cai, R.; Zhang, Z.; and Hao, Z. 2017. SADA: A General Framework to Support Robust Causation Discovery with Theoretical Guarantee. *CoRR*, abs/1707.01283.

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

Diakonikolas, I.; and Kane, D. M. 2016. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 685–694. IEEE.

Doran, G.; Muandet, K.; Zhang, K.; and Schölkopf, B. 2014. A Permutation-Based Kernel Conditional Independence Test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, 132–141. Arlington, Virginia, USA: AUAI Press.

Flaxman, S. R.; Neill, D. B.; and Smola, A. J. 2016. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM TIST*, 7(2): 22–1.

Fukumizu, K.; Gretton, A.; Sun, X.; and Schölkopf, B. 2007. Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems*, 20(1): 167–204.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2006. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, 513–520.

Grosse-Wentrup, M.; Janzing, D.; Siegel, M.; and Schölkopf, B. 2016. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125: 825–833.

Lee, S.; and Honavar, V. G. 2017. Self-Discrepancy Conditional Independence Test. In *Uncertainty in artificial intelligence*, volume 33.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Peters, J.; Mooij, J.; Janzing, D.; and Schölkopf, B. 2012. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*.

Ramsey, J. D. 2014. A scalable conditional independence test for nonlinear, non-gaussian data. *arXiv preprint arXiv:1401.5031*.

Strobl, E. V.; Zhang, K.; and Visweswaran, S. 2017. Approximate Kernel-based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *arXiv preprint arXiv:1702.03877*.

Su, L.; and White, H. 2008. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(04): 829–864.

Velikova, M.; van Scheltinga, J. T.; Lucas, P. J.; and Spaanderman, M. 2014. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1): 59–73.

Zhang, H.; Zhou, S.; and Guan, J. 2018. Measuring Conditional Independence by Independent Residuals:Theoretical Results and Application in Causal Discovery. In *AAAI Conference on Artificial Intelligence*.

Zhang, H.; Zhou, S.; Guan, J.; and Huan, J. L. 2019. Measuring Conditional Independence by Independent Residuals for Causal Discovery. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5): 1–19.

Zhang, H.; Zhou, S.; Zhang, K.; and Guan, J. 2017. Causal Discovery Using Regression-Based Conditional Independence Tests. In *AAAI*, 1250–1256.

Zhang, H.; Zhou, S.; Zhang, K.; Guan, J.; and Zhang, J. 2021. Testing Independence Between Linear Combinations for Causal Discovery. In *AAAI*.

Zhang, K.; and Hyvärinen, A. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 647–655. AUAI Press.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-based Conditional Independence Test and Application in Causal Discovery. 804–813. Corvallis, OR, USA: AUAI Press.

Zhang, K.; Wang, Z.; Zhang, J.; and Schölkopf, B. 2016. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 7(2).