The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Measuring Conditional Independence by Independent Residuals: Theoretical Results and Application in Causal Discovery

Hao Zhang,[†] **Shuigeng Zhou,**[†*] **Jihong Guan**[‡]
[†]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China.
[‡]Department of Computer Science & Technology, Tongji University, China
[†]{haoz15, sgzhou}@fudan.edu.cn; [‡]jhguan@tongji.edu.cn

Abstract

We investigate the relationship between conditional independence (CI) $x \perp y|Z$ and the independence of two residuals $x - E(x|Z) \perp y - E(y|Z)$, where x and y are two random variables, and Z is a set of random variables. We show that if x, y and Z are generated by following linear structural equation model and all external influences follow Gaussian distributions, then $x \perp y|Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$. That is, the test of $x \perp y|Z$ can be relaxed to a simpler unconditional independence test of $x - E(x|Z) \perp y - E(y|Z)$. Furthermore, if all these external influences follow non-Gaussian distributions and the model satisfies structural faithfulness condition, then we have $x \perp y|Z \Leftrightarrow x - E(x|Z) \perp y - E(y|Z)$.

We apply the results above to the causal discovery problem, where the causal directions are generally determined by a set of V-structures and their consistent propagations, so CI test-based methods can return a set of Markov equivalence classes. We show that in linear non-Gaussian context, $x - E(x|Z) \perp y - E(y|Z) \Rightarrow x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ $(\forall z \in Z)$ if Z is a minimal d-separator, which implies z causes x (or y) if z directly connects to x (or y). Therefore, we conclude that CIs have useful information for distinguishing Markov equivalence classes.

In summary, compared with the existing discretization-based and kernel-based CI testing methods, the proposed method provides a simpler way to measure CI, which needs only one unconditional independence test and two regression operations. When being applied to causal discovery, it can find more causal relationships, which is experimentally validated.

Introduction

Statistical independence and conditional independence (CI) are important concepts in statistics, artificial intelligence (AI) and other related fields. In causal discovery, causal relationships are usually revealed by checking CIs among variables. For example, for two sets of variables X and Y that are conditional independent given Z (denoted by $X \perp Y|Z$), it means that given Z, further knowing X (or Y) does not provide any additional information about Y (or X). Therefore, we know that X and Y have no directed causality under faithfulness assumption (Pearl 2009).

Generally speaking, independence and CI play a central role in causal discovery. The CI relationship $X \perp Y|Z$ allows us to separate $X - Y$ when constructing a probabilistic model based on $P(X, Y, Z)$, which results in a parsimonious representation (Zhang et al. 2011). By using CI tests, the PC algorithm (Spirtes, Glymour, and Scheines 2000), for example, can return a set of Markov equivalence classes (Pearl 2009). CI testing is much more difficult than marginal independence testing (Bergsma 2004). Most existing methods are based on explicit estimation of conditional densities or their variants, or discretize the conditional set Z to a set of bins, and transform CI to independence in each bin. Due to the curse of dimensionality, the conditional set becomes very large, inevitably the required sample size increases dramatically. For example, in (Su and White 2008) the authors used a characterization of CI, $P_{X|YZ} = P_{X|Z}$, to check CI by measuring the distance between estimates of conditional density. However, accurate estimation of conditional density or related quantity is not easy, which deteriorates the testing result, especially when the conditional set is very large.

Concretely, if Z takes a finite number of values $\{z_1, ..., z_k\}$, then $X \perp Y|Z$ if and only if $X \perp Y|Z = z_i$ for each value z_i. Given a sample of size n, even if the data points are distributed evenly on the values of Z, we must show the independence within each subset of the sample with the same Z value by using only approximately n/k points in each subset. When Z is real-valued and P_z is continuous, or Z contains several variables, the observed values of Z are almost surely unique. To extend the above procedure to the continuous cases, we must infer conditional independence using nonidentical but neighboring values of Z, where “neighboring” is quantified by some distance metric. Finding neighboring points becomes more difficult as the dimensionality of Z grows. To approximate CI to unconditional independence between X and Y in each subset, we need a large number of subsets of Z. However, with too many subsets, the subsets may have not enough data points to evaluate independence.

To alleviate these problems, researchers resort to kernel-based methods. With the ability to represent high order moments, mapping of variables into reproducing kernel Hilbert spaces (RKHSs) allows us to infer properties of distributions, such as independence and homogeneity (Gretton et al. 2006). In (Fukumizu et al. 2007), the authors proposed

[*]Correspondence author.

to use the Hilbert-Schmidt norm of the conditional cross covariance operator, which is a measure of conditional covariance of the images of X and Y under the corresponding functions from RKHSs. When the RKHSs are characteristic kernels, the operator norm is zero if and only if $X \perp Y|Z$. A later method (denoted by KCIT in short) proposed in (Zhang et al. 2011), uses partial association of regression functions to measure CI, $X \perp Y|Z$ iff for all $f \in L^2_{XZ}$ and $g \in L^2_Y$ (L^2_{XZ} and L^2_Y denote the spaces of square integrable functions of (X, Z) and Y, respectively) such that $E(\tilde{f}\tilde{g}) = 0$ where $\tilde{f}(X, Z) = f(X, Z) - r_f(Z)$ and $\tilde{g}(Y, Z) = g(Y) - r_g(Z)$ ($r_f, r_g \in L^2_Z$ are regression functions). This method is motivated by Daudin's work (Daudin 1980) and relaxes the spaces of functions f, g, r_f and r_g to RKHSs, corresponding to kernels defined on these variables. (Doran et al. 2014) introduced the PKCIT method that utilizes permutation to convert the CI test problem into an easier two-sample test problem. However, PKCIT takes too much time to compute the required permutation. (Strobl, Zhang, and Visweswaran 2017) utilized random Fourier features to approximate KCIT, and developed two algorithms RCIT and RCoT, which are much faster than KCIT.

Compared to discretization-based CI testing methods, kernel methods exploit more complete information of the data and involve less random error. It was showed that causal learning methods based on kernel methods can discover more accurate causalities.

Recently, regression-based tests were proposed for CI testing. (Grosse-Wentrup et al. 2016) proved that if there exists a function f such that $x - f(Z) \perp (y, Z)$ then $x \perp y|Z$. (Zhang et al. 2017) showed that if there exists two functions f and g such that $x - f(Z) \perp (y - g(Z), Z)$ then $x \perp y|Z$. These methods find the function f (or g) by regressing x (or y) on Z, which are able to relax a CI test to a set of marginal independence tests. However, they both showed that their methods are just sufficient but not necessary to determine CI. In practice, $x - f(Z) \perp Z$ is a strong condition, as $x - E(x|Z) \perp Z \Rightarrow Z$ causes x in many cases (Zhang and Hyvärinen 2009). Moreover, when the dimension of Z is large, to check whether a variable $x - f(Z)$ is independent from a set of variables (y, Z) or $(y - g(Z), Z)$ (joint distribution) is still prohibitively expensive. For example, in linear non-Gaussian cases, we often conduct $|y| + |Z|$ marginal independence tests to check whether $x - f(Z) \perp (y, Z)$ holds. In (Flaxman, Neill, and Smola 2016), the authors showed that given structural faithfulness and Markov assumptions (Pearl 2009), whenever Z causes x or y, it follows that $x \perp y|Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$. Similarly, here a strong condition that Z causes x or y is assumed. We can see that if these conditions are given, then it is easy to derive the corresponding causalities. Moreover, faithfulness condition means $x \perp y|Z \Rightarrow x$ and y are d-separated by Z, and Markov condition implies y are d-separated by $Z \Rightarrow x \perp y|Z$, so CI is relaxed to d-separation given the faithfulness and Markov assumptions (Pearl 2009). However, CI is neither sufficient nor necessary to d-separation. In practice, given the faithfulness assumption, $x - E(x|Z) \perp y - E(y|Z)$ and $x \perp y|Z$ have significant correlations. For example, in (Ramsey 2014), the authors suggested to use $x - E(x|Z) \perp y - E(y|Z)$ to test $x \perp y|Z$ under the faithfulness assumption. In (Zhang et al. 2017), the authors further conjectured that $x - f(Z) \perp y - g(Z)$ can lead to $x \perp y|Z$ under nonlinear and faithfulness conditions, where f and g are arbitrary nonlinear functions, x, y and Z are generated by following nonlinear additive noise model (Zhang and Hyvärinen 2009; Peters, Janzing, and Schölkopf 2011).

In this work, we aim to investigate the relationship between the two terms $x - E(x|Z) \perp y - E(y|Z)$ and $x \perp y|Z$ in the scenario that x, y and Z are generated by following linear structural equation model (Shimizu et al. 2011), i.e., $x = \sum_{i=1}^{q} a_i s_i$, $y = \sum_{i=1}^{p} b_i s_i$ and $z_j = \sum_{i=1}^{r_j} c_i s_i$ ($\forall z_j \in Z$) where s_i is the external influence. We prove that if all external influences follow Gaussian distributions, then $x \perp y|Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$. Note that, here we do not assume the faithfulness and Markov conditions, but only require that x, y and $\forall z_j \in Z$ are linear combinations of those external influences. Therefore, we can relax the test of $x \perp y|Z$ to a simpler unconditional independence test of $x - E(x|Z) \perp y - E(y|Z)$ without considering d-separation. Furthermore, if all these external influences follow non-Gaussian distributions and the model satisfies structural faithfulness condition, then we show that $x - E(x|Z) \perp y - E(y|Z) \Leftrightarrow x \perp y|Z$.

It is well known that existing causal discovery methods based on CI tests usually return a set of Markov equivalence classes (Spirtes and Zhang 2016) by detecting a set of V-structures and their consistent propagations (Meek 1995; Chickering 2002). With the theoretical results above, we show that CI testing based on independent residuals contains information for causal direction inference. We prove that in the linear non-Gaussian context, $x - E(x|Z) \perp y - E(y|Z) \Rightarrow x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ ($\forall z \in Z$) if Z is a minimal d-separator (Tian, Pearl, and Paz 1998). When CI testing-based methods like PC algorithm (Spirtes, Glymour, and Scheines 2000) return a causal skeleton, if two variables x and z are directly connected and $x - E(x|Z) \perp z$ holds, we can easily deduce that z is a cause of x in the non-Gaussian case.

In summary, compared with the existing discretization-based and kernel-based CI testing methods, testing independence between two residuals needs only one marginal independence test and two regression operations. Moreover, this method can infer more causal directions than these methods when being applied to causal discovery.

Measuring conditional independence by independent residuals

Here we first quote Daudin's work (Daudin 1980) that gives the characterization of conditional independence by explicitly enforcing the uncorrelatedness of functions in suitable spaces, because it is used to prove our theorems.

Characterization of conditional independence (CCI) (Daudin 1980) Let X, Y and Z be three real random variables or sets of random variables, $E_1 = \{g \in L^2_{XZ}, E(g|Z) = 0\}$, $E_2 = \{h \in$

$L^2_{YZ}, E(h|Z) = 0\}$, $E_3 = \{g' \in L^2_X, E(g') = 0\}$ and $E_4 = \{h' \in L^2_Y, E(h') = 0\}$ where L^2_X, L^2_Y, L^2_{XZ} and L^2_{YZ} denote the spaces of square integrable functions of X, Z, (X, Z) and (Y, Z), respectively, then the following conditions are equivalent to each other:

1) $X \perp Y|Z$;
2) $\forall g \in E_1$ and $\forall h \in E_1$, $E(gh) = 0$;
3) $\forall g \in E_1$ and $\forall h' \in E_4$, $E(gh') = 0$;
4) $\forall h \in E_2$ and $\forall g' \in E_3$, $E(hg') = 0$.

Consider $Z = \emptyset$, we can derive $X \perp Y \Leftrightarrow \forall g' \in E_3$ and $\forall h' \in E_4$, $E(g'h') = 0$.

In what follows, we present the theoretical results on the relationship between CI and independent residuals in Gaussian and non-Gaussian cases respectively.

Theorem 1. *Define $m + 2$ random variables x, y and $Z = \{z_1, ..., z_m\}$ as linear combinations of independent random variables s_i ($i = 1, ..., l$), if all s_i follow Gaussian distributions, then $x \perp y|Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$.*

Proof. If $x \perp y|Z$, then $\forall g \in E_1$ and $\forall h \in E_2$, $E(gh) = 0$ according to the condition (2) in CCI. As $E(x - E(x|Z)|Z) = 0$ and $E(y - E(y|Z)|Z) = 0$, then $x - E(x|Z) \in E_1$ and $y - E(y|Z) \in E_2$, we have $cov\{(x - E(x|Z))(y - E(y|Z))\} = E\{(x - E(x|Z))(y - E(y|Z))\} - E(x - E(x|Z))E(y - E(y|Z)) = 0$. Thus in the Gaussian case, we have $x - E(x|Z) \perp y - E(y|Z)$.

On the other side, consider the partial correlation of x and y given Z, $\rho_{xy.Z} = \frac{\sigma_{xy.Z}}{\sqrt{\sigma_{xx.Z}\sigma_{yy.Z}}}$. The partial variance or covariance given Z, ($\sigma_{**.Z}$), can be considered as the variance or covariance between residuals of projections of x and y on the linear space spanned by Z, thus $\sigma_{xy.Z} = cov(x - E(x|Z), y - E(y|Z)) = 0$. In the linear Gaussian case, zero partial correlation is equivalent to the conditional independence (Baba, Shibata, and Sibuya 2004), we therefore obtain $x \perp y|Z$. □

Theorem 1 shows that CI and the independence between two residuals are equivalent in the Gaussian case. Next, we consider the non-Gaussian case. Here, we first quote Darmois-Skitovitch theorem (Darmois 1953; Skitovich 1953) as it is used to prove Theorem 2:

Darmois-Skitovitch theorem (DST) Define two random variables x and y as linear combinations of independent random variables s_i ($i = 1, ..., l$), $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$. Then, if $x \perp y$, all variables s_j for which $a_j b_j \neq 0$ are Gaussian.

This theorem means that if there exists a non-Gaussian s_j for which $a_j b_j \neq 0$, then x and y are dependent.

Theorem 2. *Define $m + 2$ random variables x, y and $Z = \{z_1, ..., z_m\}$ generated by following a l-dimensional linear structural equation model satisfying faithfulness condition, if all the external influences s_i ($i = 1, ..., l$) are non-Gaussian, then $x \perp y|Z \Leftrightarrow x - E(x|Z) \perp y - E(y|Z)$.*

Proof. In linear regression, the conditional expectations $E(x|Z)$ and $E(y|Z)$ are linear combinations of the independent variables $z_1, ..., z_m$, we therefore have $x - E(x|Z) = \sum_{i=1}^{l} a_i s_i$ and $y - E(y|Z) = \sum_{i=1}^{l} b_i s_i$. Given $x - E(x|Z) \perp y - E(y|Z)$ and $\forall j$, if $a_j \neq 0$, then there must be $b_j = 0$ according to DST. Consider $\forall z \in Z$, there are two cases: 1) $z \perp x - E(x|Z)$ or $z \perp y - E(y|Z)$; 2) $z \not\perp x - E(x|Z)$ and $z \not\perp y - E(y|Z)$.

Case 1: Without loss of generality, assume $Z = \{z_1, z_2\}$, then there are two subcases: (1) $Z \perp x - E(x|Z)$ (or $Z \perp y - E(y|Z)$) and (2) $z_1 \perp x - E(x|Z)$, $z_2 \not\perp x - E(x|Z)$, $z_1 \not\perp y - E(y|Z)$ and $z_2 \perp y - E(x|Z)$.

Subcase (1): If $Z \perp x - E(x|Z)$, for the conditional mutual information $I(x; y|Z)$ of x and y given Z, we have

$$
\begin{aligned}
&I(x; y|Z) \\
&= I(x - E(x|Z); y - E(y|Z)|Z) \\
&= I(x - E(x|Z); y - E(y|Z), Z) - I(x - E(x|Z); Z).
\end{aligned}
$$

Given $x - E(x|Z) \perp y - E(y|Z)$ and $Z \perp x - E(x|Z)$, we can deduce that $x - E(x|Z) \perp (y - E(y|Z), Z)$ according to DST. Therefore, $I(x - E(x|Z); y - E(y|Z), Z) = 0$ and $I(x - E(x|Z); Z) = 0$, we have $I(x; y|Z) = 0$, i.e., $x \perp y|Z$. Similar result can be derived when we consider $Z \perp y - E(y|Z)$.

Subcase (2): If $z_1 \perp x - E(x|Z)$, $z_2 \not\perp x - E(x|Z)$, $z_1 \not\perp y - E(y|Z)$ and $z_2 \perp y - E(x|Z)$, then there must be $z_1 \perp z_2$, otherwise $x - E(x|Z)$ cannot be independent of $y - E(y|Z)$ according to DST. Similar to subcase (1), considering the conditional mutual information $I(x; y|Z)$ of x and y given Z, we have

$$
\begin{aligned}
&I(x; y|Z) \\
&= I(x - E(x|Z); y - E(y|Z), z_1, z_2) - I(x - E(x|Z); z_1, z_2) \\
&= I(x - E(x|Z); z_2) - I(x - E(x|Z); z_2) = 0. \\
&\Rightarrow x \perp y|Z
\end{aligned}
$$

Case 2: Consider the external influence of x, denote by s_x, we have $x - E(x|Z) \not\perp s_x$ or faithfulness must be violated. As $x - E(x|Z) \perp y - E(y|Z)$, we have $s_x \perp y - E(y|Z)$. Note that, there is only one edge between s_x and x, $s_x \rightarrow x$, then we can further deduce that $x \perp y - E(y|Z)$, which means x cannot be directly connected to y or faithfulness is violated. $\forall z \in Z$, there are three scenarios: 1) z is on the path between x and y; 2) z is a collider or a descendant of a collider w.r.t. x and y; 3) z is not included in any path between x and y.

Scenario (1): If z is on the path between x and y, let s_z denote the external influence of z, we have $s_z \perp x - E(x|Z)$ or $s_z \perp y - E(y|Z)$ according to $x - E(x|Z) \perp y - E(y|Z)$ and DST. Similarly, there is only one edge between s_z and z, $s_z \rightarrow z$, then we can further deduce that $z \perp x - E(x|Z)$ or $z \perp y - E(y|Z)$. As aforementioned in Case 1, we have $x \perp y|Z$.

Scenario (2): Given z is a collider or a descendant of a collider w.r.t. x and y, let s_z denote the external influence of z. To ensure that $x - E(x|Z) \perp y - E(y|Z)$ holds, s_z must be removed from $x - E(x|Z)$ and $y - E(y|Z)$, thus there must be at least a descendant of z contained in Z. However, such a descendant will lead to $x - E(x|Z) \not\perp y - E(y|Z)$, this is contradictory.

Scenario (3): As $x \perp y - E(y|Z)$ or $y \perp x - E(x|Z)$, if z is not contained in any path between x and y, we can easily deduce that $z \perp x - E(x|Z)$ or $z \perp y - E(y|Z)$, i.e., $x \perp y|Z$.

On the other side, $x \perp y|Z$ can also lead to $x - E(x|Z) \perp y - E(y|Z)$. Given x and y are d-separated by Z, without loss of generality, we assume that Z is the minimal d-separator. If $\forall z \in Z$ is either a descendant of x (or y) and an ancestor of y (or x) (i.e., $x \rightarrow \ldots \rightarrow z \rightarrow \ldots \rightarrow y$) or a common ancestor of x and y (i.e., $x \leftarrow \ldots \leftarrow z \rightarrow \ldots \rightarrow y$), then we can deduce that $x - E(x|Z) \perp Z$ (or $y - E(y|Z) \perp Z$) according to the mechanism of additive noise model (Hoyer et al. 2009; Peters, Janzing, and Schölkopf 2011), and further conclude that $x - E(x|Z) \perp y - E(y|Z)$. Similarly, other cases can also be analyzed by following this way. Finally, we have $x \perp y|Z \Rightarrow x - E(x|Z) \perp y - E(y|Z)$. □

Theorem 1 and 2 indicate that we can do a CI test by just testing the independence of two residuals in linear Gaussian and non-Gaussian cases. We denote this CI test method as **ReCIT** (the abbreviation of **Re**sidual-based **C**onditional **I**ndependence **T**est). In next section, we apply ReCIT to causal discovery. We show that CI contains information about causal direction, which can distinguish Markov equivalent classes.

Causal discovery based on ReCIT

We have the following theorem:

Theorem 3. *Given $m+2$ random variables x, y and $Z = \{z_1, ..., z_m\}$ that are generated by following a linear non-Gaussian structural equation model that satisfies the faithfulness and Markov conditions, for any $z \in Z$ directly connecting to x (or y), if $x - E(x|Z) \perp y - E(y|Z)$, then we have $x - E(x|Z) \perp z$ (or $y - E(y|Z) \perp z$) $\Rightarrow$ z causes x (or z causes y).*

Proof. If x and y are d-separable, we can find a Z such that $x - E(x|Z) \perp y - E(y|Z)$ under the faithfulness and Markov assumptions. From the proof of Theorem 2, we know that there must be $x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ if Z is a minimal d-separator. Without loss of generality, we assume $x - E(x|Z) \perp z$. Let ε denote the exogenous disturbance of z, then $x - E(x|Z) \perp z$ means $x - E(x|Z) \perp \varepsilon$ in the linear non-Gaussian case according to DST. If z is a child of x, then $x \rightarrow z \leftarrow \varepsilon$ forms a V-structure, we can deduce that $x - E(x|Z) \not\perp \varepsilon$ as z is a collider, which is a contradiction. Therefore, z can only be the parent of x. Similarly, we can prove the case w.r.t. y. □

Note that, the conclusion $x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ $(\forall z \in Z)$ of Theorem 3 is different from the assumptions or preconditions in the existing regression-based CI testing methods (Flaxman, Neill, and Smola 2016; Grosse-Wentrup et al. 2016; Zhang et al. 2017), they all require x-$E(x|Z) \perp Z$ or y-$E(y|Z) \perp Z$, or Z causes x or y.

Compared with the existing CI testing methods, ReCIT can detect more causal directions even there is no V-structure contained in the corresponding DAG. To make it clearer, let us consider a simple example. Given a DAG: $x_1 \leftarrow x_2 \rightarrow x_3$, it is easy to find $x_1 - E(x_1|x_2) \perp x_2$ and $x_3 - E(x_3|x_2) \perp x_2$. We therefore can infer $x_1 \leftarrow x_2$ and $x_2 \rightarrow x_3$. However, it is difficult for the existing CI testing methods to distinguish the three structures $x_1 \leftarrow x_2 \rightarrow x_3$, $x_1 \leftarrow x_2 \leftarrow x_3$ and $x_1 \rightarrow x_2 \rightarrow x_3$, because all of them fit the observed conditional and unconditional independence, though obviously having completely different structures.

In what follows, we present a new causality discovery algorithm based on ReCIT under the PC algorithm framework. We denote the new algorithm as PC$_{ReCIT}$, where we use ReCIT to check CIs, and use existing methods (e.g. KCIT (Zhang et al. 2011)) to test unconditional independence. Concretely, we calculate $x - E(x|Z)$ and $y - E(y|Z)$ simply by least square regression, i.e., $x - E(x|Z) = x - Z(Z^T Z)^{-1} Z^T x$ and $y - E(y|Z) = y - Z(Z^T Z)^{-1} Z^T y$. And any independence testing method can be used to test the independence between $x - E(x|Z)$ and $y - E(y|Z)$.

PC$_{ReCIT}$ is outlined in Algorithm 1. The first step (lines 1 – 6) is to construct the causal skeleton by employing ReCIT. The procedure follows the PC algorithm. That is, we form the complete undirected graph G on the variables set X, then check whether every two variables x_i and x_j are conditional independent, given a set of variables Z. Here, we keep the corresponding regression results $x_i - E(x_i|Z)$ and $x_j - E(x_j|Z)$ in two sets $Temp_{x_i}$ and $Temp_{x_j}$, which are useful in the next step of inferring causal direction. We then detect V-structures as in the PC algorithm (lines 7 – 11). That is, to check whether a local structure $x_i - x_k - x_j$ can form a V-structure. If it is, orient it as $x_i \rightarrow x_k \leftarrow x_j$. For a structure $x_i - x_k$, as $x_i - E(x_i|Z) \perp x_k$ $(x_k \in Z)$ implies $x_i \leftarrow x_k$, if $x_i - E(x_i|Z) \in Temp_{x_i}$, we test the dependence between $x_i - E(x_i|Z)$ and x_k. If independence holds, then orient $x_i \leftarrow x_k$. These operations are shown in lines 12 – 18. Finally, we conduct consistent propagation to orient more directions and output the partial DAG (PDAG) w.r.t. the given data (line 19).

Performance evaluation

We conduct extensive experiments to evaluate ReCIT, and compare it with KCIT (Zhang et al. 2011). We also compare the causal inference performance of ReCIT and KCIT under the PC algorithm framework (Spirtes, Glymour, and Scheines 2000), i.e., PC$_{ReCIT}$ vs. PC$_{KCIT}$. To the best of our knowledge, KCIT is one of the best methods for CI testing in general cases. There are many comparisons between KCIT and other existing CI testing methods in the literature (Zhang et al. 2011; Doran et al. 2014; Zhang et al. 2017; Strobl, Zhang, and Visweswaran 2017). In our ReCIT implementation, we do regression using least square regression and the unconditional independence tests using KCIT.

Effect of Z's dimensionality and sample size

We first examine how the probabilities of Type I (where the CI hypothesis is incorrectly rejected) and Type II (where the CI hypothesis is not rejected although being false) errors of ReCIT change with the size of the conditioning set Z $(D = 1, 2, ..., 5)$ and the sample size ($n = 100$ and 200) by simulation. Here, we consider two cases as follows.

Algorithm 1 PC algorithm based on ReCIT (PC$_{ReCIT}$)

```
Input: a set of variables $X = \{x_1, ..., x_n\}$, a threshold $k$.
Output: a partial DAG $G$.
Form the complete undirected graph $G$ on the variables set $X$.
for $\forall x_i, x_j \in X$ and adjacent in $G$ do
    if $\exists Z \subseteq X \setminus \{x_i, x_j\}$ and ($|Z| < k$) such that $x_i - E(x_i|Z) \perp x_j - E(x_j|Z)$ then
        remove edge $x_i - x_j$ from $G$ and record $Z$ in $Sepset(x_i, x_j)$ and record $x_i - E(x_i|Z)$ and $x_j - E(x_j|Z)$ in temporary sets $Temp_{x_i,Z}$ and $Temp_{x_j,Z}$.
    end if
end for
for $\forall x_i, x_j, x_k \in X$ such that the pair $x_i$, $x_k$ and the pair $x_j$, $x_k$ are adjacent in $G$ but the pair $x_i$, $x_j$ are not adjacent in $G$ do
    if $x_k \notin Sepset(x_i, x_j) \cup Sepset(x_j, x_i)$ then
        orient $x_i - x_k - x_j$ as $x_i \rightarrow x_k \leftarrow x_j$.
    end if
end for
for $\forall x_i, x_k \in X$ such that $x_i$ and $x_k$ are adjacent do
    if $\exists Z$ such that $x_i - E(x_i|Z) \in Temp_{x_i}$ and $x_k \in Z$ then
        if $x_i - f(Z) \perp x_k$ then
            orient $x_i - x_k$ as $x_i \leftarrow x_k$.
        end if
    end if
end for
do consistent propagation.
```

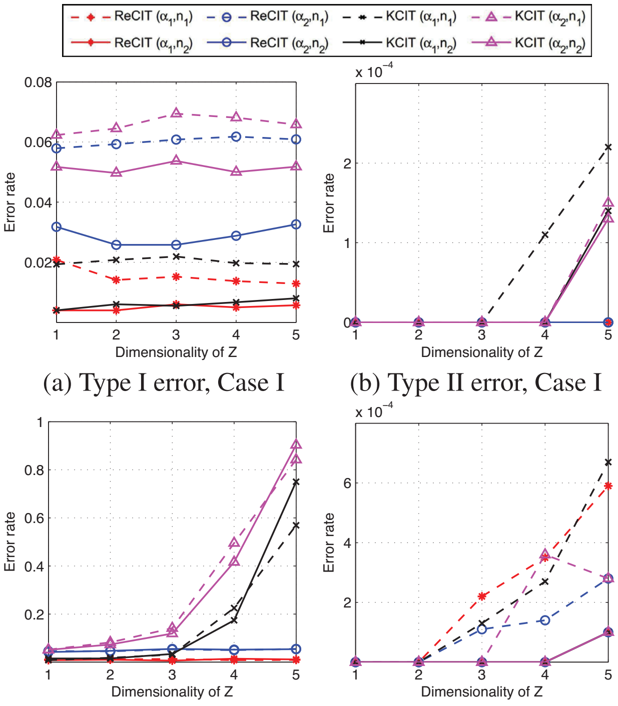
In Case I, only one variable in Z, denoted by Z_1, is effective, i.e., other conditioning variables are independent of X, Y, and Z_1. We generate X and Y from Z_1 according to the additive noise model (ANM) data generating procedure: they are generated as $a * Z_1 + \varepsilon$ where $a \sim U(0.2, 1)$ are different for X and Y, and $\varepsilon \sim U(-0.2, 0.2)$. Hence, $X \perp Y|Z$ holds. In our simulations, Z_i is i.i.d. $U(0, 1)$.

In Case II, all variables in the conditioning set Z are effective in generating X and Y. We first generate the independent variables Z_i, then X and Y are generated as $\sum_i b_i * Z_i + \varepsilon$ where b_i follows a.

We compare ReCIT with KCIT in terms of both types of error. The significance levels are fixed at $\alpha_1 = 0.01$ and $\alpha_2 = 0.05$ respectively. 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 1000 times and average their results.

Type I and II errors are calculated like this: for example $D = 3$, in Case I x should be 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 =1-$the\ number\ of\ CIs/4$. On the other side, x is independent of y given $\emptyset$, (Z_2), (Z_3) and (Z_2, Z_3), then Type II error = $the\ number\ of\ CIs/4$. Similarly, we can calculate Type I and II errors in Case II.

We first examine Type I error in Case I and Case II. As shown in Fig. 1(a) and (c), Type I error of ReCIT is close to the significance level. As D increases, the probability of Type I error increases slightly. In Case I, Type I error of ReCIT is lower than that of KCIT. However, in Case II, even when $D = 3$, the probability of Type I error of KCIT is obviously larger than the significance level. Furthermore, KCIT is very sensitive to D. We can see that increasing sample size (from 100 to 200) can obviously reduce Type I error in Case I, while in Case II the effect is not so obvious.



(a) Type I error, Case I (b) Type II error, Case I

(c) Type I error, Case II (d) Type II error, Case II

Figure 1: The probabilities of Type I and Type II errors obtained by simulation in various situations. The significance level $\alpha_1 = 0.01$, $\alpha_2 = 0.05$ and the sample size $n_1 = 100$, $n_2 = 200$. Top: Case I (only one variable in Z is effective to X and Y). Bottom: Case II (all variables in Z are effective).

To further illustrate why ReCIT can perform much better than KCIT in terms of Type I error in Case II (see Fig. 1(c)), we conduct another experiment to evaluate the two methods under different noise weights. We simulate four sets of noise, $\varepsilon_1 \sim U(-0.05, 0.05)$, $\varepsilon_2 \sim U(-0.2, 0.2)$, $\varepsilon_3 \sim U(-0.5, 0.5)$ and $\varepsilon_4 \sim U(-1, 1)$ and keep the sample size $n = 100$ and the significance level $\alpha = 0.05$. The results are showed in Fig. 2. We can see that, in the case of $\varepsilon_4 \sim U(-1, 1)$, the error rate of KCIT is extremely close to the significance level (0.05). However, as the noise weight grows, the error rate dramatically increases. Recall that the data-generating function is $\sum_i b_i * Z_i + \varepsilon$, which means if the noise ε is much less than the linear combination term $\sum_i b_i * Z_i$, KCIT tends to be unreliable in this case. On the other hand, we can see that the error rate of ReCIT keeps close to the significance level in all cases. This is because in the process of ReCIT, $x - E(x|Z) = \varepsilon_x$ and $y - E(y|Z) = \varepsilon_y$. Then, testing $x \perp y|Z$ is equivalent to testing the independence between two independent noise terms ε_x and ε_y. Therefore, the accuracy is not affected by noise weight. Fig. 1(a, c) and Fig. 2 show that ReCIT performs better and more robust than KCIT in different situations in terms of Type I error.

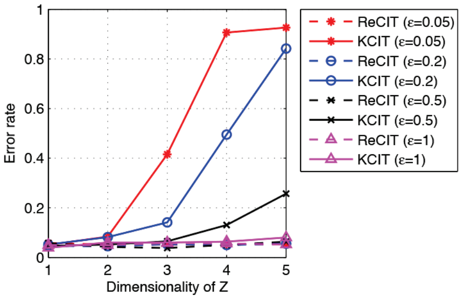


Figure 2: The probability of Type II error in Case II (all variables in Z are effective) for different noise weights.

However, as shown in Fig. 1(b) and 1(d), we can see that the results of ReCIT and KCIT are close to zero in terms of Type II error in both Case I and Case II. As D increases, the probability of Type II error always increases. Intuitively, this is reasonable: due to the finite sample size effect, as the conditioning set becomes larger and larger, X and Y tend to be considered as conditionally independent. On the other hand, as the sample size increases from 100 to 200, the probability of Type II error approaches zero. In particular, as shown in Fig. 1(d), the curves of ReCIT with sample size=200 keep close to zero. That is, increasing sample size from 100 to 200 can dramatically reduce Type II error.

As far as causal discovery is concerned, the performance of CI test based methods is always heavily affected by Type II error instead of Type I error. This is due to two reasons: 1) two adjacent variables will not be affected by Type I error; 2) Assume that a CI of two non-adjacent variables is incorrectly rejected when Type I error is occurred, by increasing the size of d-separators, we can usually find another controlling set to d-separate the two variables. Despite this, KCIT and ReCIT have very similar performance when the dimensionality of Z is 1 and 2, which means that when the given DAG is very small (with small d-separators), these two methods perform similarly in discovering causal skeleton. However, ReCIT can learn more causal directions, which will be discussed in the next subsection.

Performance in causal discovery

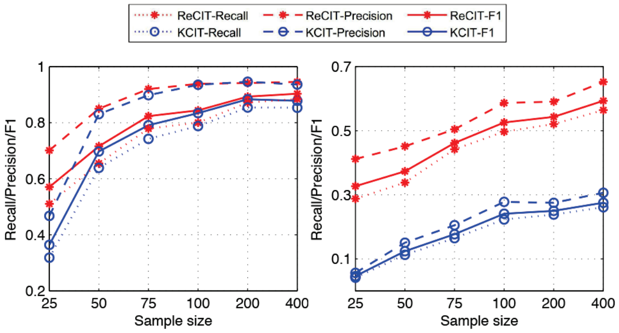
CI tests are frequently used in causal inference where we assume that the true causal structure of n random variables $x_1, ..., x_n$ can be represented by a directed acyclic graph (DAG) G. More specifically, the causal Markov condition assumes that the joint distribution satisfies all CIs that are imposed by the true causal graph. The constraint-based methods like the PC algorithm make additional assumption of faithfulness (i.e., the joint distribution does not allow any CI that is not entailed by the Markov condition) and recover the graph structure by exploiting the CIs and independence that can be found in the data. Obviously, this is only possible up to Markov equivalence classes, which are sets of graphs that impose exactly the same independence and CIs. Hence, the PC algorithm based on existing CI test methods orient causal directions by finding V-structures and consistent propagations (Pearl 2009). In our experiments, we show that PC$_{ReCIT}$ can reveal much more causal directions as mentioned above.

We generate data from a random DAG G. In particular, we sample four random variables $x_1, ..., x_4$ and allow arrows from x_i to x_j only for $i < j$. With probability 0.5 each possible arrow is either present or absent. The root variables are generated by $U(0, 1)$ and the leaf variables x_i are generated by $\sum_i a_i * pa_{x_i} + \varepsilon$ 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 between 25 and 400, we simulate 1000 DAGs and evaluate the performance of the two methods PC$_{ReCIT}$ and PC$_{KCIT}$on discovering causal skeleton and PDAG (including identifiable causal directions).

As shown in Fig. 3(a), we can see that when the sample size is small (e.g. less than 100), PC$_{ReCIT}$ performs significantly better than PC$_{KCIT}$. As the sample size increases, the performance of PC$_{KCIT}$ tends close to that of PC$_{ReCIT}$. When the sample size up to 400, the F1 curves of PC$_{ReCIT}$ and PC$_{KCIT}$ tend to overlap, but the former is still slightly (about 0.025) better than that of the latter. Considering that the regression coefficient $Z(Z^T Z)^{-1} Z^T$ in ReCIT can be easily calculated based on the least square method, and any possible error is generated by marginal independence test w.r.t. two residuals. Therefore, PC$_{ReCIT}$ performs significantly better than PC$_{KCIT}$ in discovering causal skeleton when the sample size is small, which is the frequently-encountered case in reality.

We also evaluate the two methods in discovering PDAG. The results are presented in Fig. 3(b). We can see that PC$_{ReCIT}$ achieves better result in all cases, though the performance of PC$_{KCIT}$ in discovering causal skeleton is very close to that of PC$_{ReCIT}$ when the sample size is large enough. The reason is that PC$_{KCIT}$ orients causal directions only based on V-structure and consistent propagation (Pearl 2009), in other words, returns only a set of Markov equivalence classes, while PC$_{ReCIT}$ can uncover more causal directions according to Theorem 3.

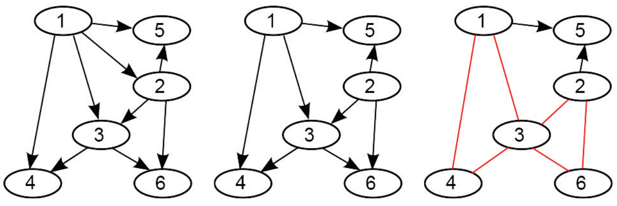
We apply PC$_{ReCIT}$ to a causal graph presented in (Shimizu et al. 2006), which was generated by following a linear non-Gaussian structure equation model w.r.t. a DAG consisting six variables as shown in Fig. 4(a). We select this graph because it contains no V-structure, which is used to further show the advantage of ReCIT in inferring causal direction. The resulting skeletons reconstructed by PC$_{ReCIT}$ and PC$_{KCIT}$ are shown in Fig. 4(b) and Fig. 4(c) respectively. We can see that all the causal edges discovering by PC$_{ReCIT}$ are correct. However, as shown in Fig. 4(c), only two directions of edges 1→5 and 2→5 are correctly inferred by PC$_{KCIT}$, others are failed to be inferred by any propagation. As there is no V-structure in this graph, theoretically none causal direction can be found by PC$_{KCIT}$. However, the edge between node 1 and node 2 are



(a) Discovering skeleton (b) Discovering PDAG

Figure 3: Performance comparison between PC$_{ReCIT}$ and PC$_{KCIT}$ with various sample sizes in discovering (a) causal skeleton and (b) PDAG.

incorrectly removed by both ReCIT and KCIT, i.e., the CI hypothesis w.r.t. node 1 and node 2 is not rejected although being false, therefore 1→5←2 forms a false V-structure. It can be seen that even some local structures are incorrectly inferred by PC, ReCIT can still distinguish the real causal directions. In one word, by taking the advantage of ReCIT, existing constraint-based methods (say the PC algorithm) can greatly improve the performance in causal discovery, as ReCIT helps to distinguish the Markov equivalence classes.



(a) (b) (c)

Figure 4: Performance comparison in causal direction inference. (a) The ground truth causal model; (b) The reconstructed DAG based on PC$_{ReCIT}$; (c) The reconstructed PDAG based on PC$_{KCIT}$. Here, the red lines are the edges whose directions are not determined.

.

Verifying the equivalence between CI and independent residuals

The above experiments show that ReCIT can work better than KCIT. In some sense, we have verified that the independence between residuals $x - E(x|Z) \perp y - E(y|Z)$ can lead to CI $x \perp y|Z$. Now, we apply ReCIT to two datasets $Hailfinder$ and $Win95pts$ used in a previous work (Cai, Zhang, and Hao 2013) to check whether independent residuals can cause CI. $Hailfinder$ is a weather forecasting network including 56 variables and 66 edges, $Win95pts$ is a printer troubleshooting network containing 76 variables and 70 edges. The results are presented in Table 1. We can see that the number of CIs used in PC$_{ReCIT}$ is extremely close to that used by PC$_{KCIT}$. If ReCIT is only sufficient but not necessary to CI, then the number of ReCIT should be obviously larger than that of KCIT. In addition, we also present the *recall*, *precision* and *F1* (R/P/F1) w.r.t. skeleton discovery, as the performance of skeleton discovery corresponds to the accuracy of CI tests. We can see that they have very similar score. These results indicate that $x \perp y|Z$ is equivalent to $x - E(x|Z) \perp y - E(y|Z)$ almost surely.

Table 1: Performance on $Hailfinder$ and $Win95pts$ networks.

Dataset	PC$_{ReCIT}$		PC$_{KCIT}$	
	R/P/F1	CIs	R/P/F1	CIs
$Hail.$	0.5/0.7/0.6	75859	0.5/0.7/0.6	75854
$Win.$	0.5/0.6/0.5	141628	0.5/0.6/0.5	141628

Conclusion

This paper studies the relationship between conditional independence $x \perp y|Z$ and the independence of two residuals $x - E(x|Z) \perp y - E(y|Z)$. In some previous works, the independence of two residuals is regarded as a week condition for CI under faithfulness and Markov assumptions. To make the week condition be sufficient, some additional condition such as $x - E(x|Z) \perp Z$ (or $y - E(y|Z) \perp Z$), Z causes x (or y) are required. In this work, we prove that if x, y and Z are generated by following a linear structural equation model and all external influences follow Gaussian distributions, then $x \perp y|Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$. Furthermore, if all these external influences follow non-Gaussian distributions and the model satisfies structural faithfulness condition, then we have $x \perp y|Z \Leftrightarrow x - E(x|Z) \perp y - E(y|Z)$. We therefore can relax the test of $x \perp y|Z$ to a simpler unconditional independence test of $x - E(x|Z) \perp y - E(y|Z)$ without assuming any other graph-related condition. Intuitively, our result means that if $x \perp y|Z$ holds, then when removing the effect of Z from x and y by regression, the remaining effect of Z on x is independent from that of y, and vice versa. On the other hand, we show that CIs can distinguish Markov equivalence classes, as we deduce that $x - E(x|Z) \perp y - E(y|Z) \Rightarrow x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ where Z is a minimal d-separator ($\forall z \in Z$), which implies z causes x (or y) if z directly connects to x (or y). We conduct extensive experiments to evaluate the proposed method, and our experimental results show that our method outperforms the kernel-based method KCIT in discovering causality in linear non-Gaussian cases.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) (U1636205, 61772367, 61373036). Jihong Guan was also supported by the Program of Shanghai Subject Chief Scientist (15XD1503600).

References

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

Bergsma, W. P. 2004. *Testing conditional independence for continuous random variables*. Eurandom.

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.

Chickering, D. M. 2002. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research* 2:445–498.

Darmois, G. 1953. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique* 2–8.

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

Doran, G.; Muandet, K.; Zhang, K.; and Schölkopf, B. 2014. A permutation-based kernel conditional independence test. In *UAI*, 132–141.

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.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, 689–696.

Meek, C. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 403–410. Morgan Kaufmann Publishers Inc.

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

Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Causal inference on discrete data using additive noise models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33(12):2436–2450.

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

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research* 7:2003–2030.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; and Bollen, K. 2011. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research* 12:1225–1248.

Skitovich, V. 1953. On a property of the normal distribution. *DAN SSSR* 89:217–219.

Spirtes, P., and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, 3. Springer.

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

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.

Tian, J.; Pearl, J.; and Paz, A. 1998. Finding minimal d-separators.

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, H.; Zhou, S.; Zhang, K.; and Guan, J. 2017. Causal discovery using regression-based conditional independence tests. In *AAAI*, 1250–1256.