Measuring Conditional Independence by Independent Residuals for Causal Discovery

HAO ZHANG, Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
SHUIGENG ZHOU, Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China; Shanghai Institute of Intelligent Electronics & Systems, Fudan University, China, China
JIHONG GUAN, Department of Computer Science and Technology, Tongji University, China
JUN (LUKE) HUAN, Big Data Lab, Baidu Research, China

We investigate the relationship between conditional independence (CI) $x \perp\!\!\!\perp y|Z$ and the independence of two residuals $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, 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 models and all external influences follow joint Gaussian distribution, then $x \perp\!\!\!\perp y|Z$ if and only if $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. That is, the test of $x \perp\!\!\!\perp y|Z$ can be relaxed to a simpler unconditional independence test of $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. Furthermore, testing $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ can be simplified by testing $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$. On the other side, if all these external influences follow non-Gaussian distributions and the model satisfies structural faithfulness condition, then we have $x \perp\!\!\!\perp y|Z \Leftrightarrow x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

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 the linear non-Gaussian context, in many cases $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp z$ ($\forall z \in Z$ and Z is a minimal d-separator) is satisfied when $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, 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, comparing 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 extensively validated by experiments.

CCS Concepts: • **Mathematics of computing → Statistical graphics**;

Additional Key Words and Phrases: Causal inference, causal discovery, conditional independence test, independent residual

This work was supported by National Natural Science Foundation of China (NSFC) (U1636205 and 61772367); Jihong Guan was supported by Special Fund for Shanghai Industrial Transformation and Upgrading under grant No. 18XI-05, Shanghai Municipal Commission of Economy and Informatization.
Authors' addresses: H. Zhang and S. Zhou, Room 502, Yifu Building, Fudan University, 220 Handan Road, Yangpu District, Shanghai, 200433, China; emails: {haoz15, sgzhou}@fudan.edu.cn; J. Gaun, Department of Computer Science and Technology, Tongji University, 4800 Cao'an Road, Shanghai, 201804, China; email: jhguan@tongji.edu.cn; J. (Luke) Huan, Baidu Technology Park, No. 10 Xibeiwang East Road, Haidian District, Beijing, 100085, China; email: huanjun@baidu.com.

2157-6904/2019/09-ART50 $15.00
https://doi.org/10.1145/3325708

ACM Reference format:
Hao Zhang, Shuigeng Zhou, Jihong Guan, and Jun (Luke) Huan. 2019. Measuring Conditional Independence by Independent Residuals for Causal Discovery. *ACM Trans. Intell. Syst. Technol.* 10, 5, Article 50 (September 2019), 19 pages.
https://doi.org/10.1145/3325708

1 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 $X \perp\!\!\!\perp Y|Z$), 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 [15].

Generally speaking, independence and CI play a central role in causal discovery [13]. The CI relationship $X \perp\!\!\!\perp Y|Z$ allows us to separate $X-Y$ when constructing a probabilistic model based on $P(X, Y, Z)$, which results in a parsimonious representation [30]. By using CI tests, the PC algorithm [22], for example, can return a set of Markov equivalence classes [15]. CI testing is much more difficult than marginal independence testing [2]. 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, which inevitably leads to dramatically increasing of the required sample size. For example, in Reference [25] 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 $\{z_1, \ldots, z_k\}$, then $X \perp\!\!\!\perp Y|Z$ if and only if $X \perp\!\!\!\perp Y|Z = z_i$ 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 in 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 [11]. In Reference [10], the authors proposed 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 $X \perp\!\!\!\perp Y|Z$. A later method (denoted by KCIT in short) proposed in Reference [30], uses partial association of regression functions to measure CI, $X \perp\!\!\!\perp Y|Z$ 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 $\mathbb{E}(\tilde{f}\tilde{g}) = 0$, 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 [6] and relaxes

the spaces of functions f, g, r_f, and r_g to RKHSs, corresponding to kernels defined on these variables. Reference [7] 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. Reference [24] 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 information of the observed data and have less random error. It was showed that causal learning based on kernel methods can discover more accurate causalities.

In this work, we aim to investigate the relationship between the two terms $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ and $x \perp\!\!\!\perp y|Z$ in the scenario that x, y and Z are generated by following the linear structural equation model (SEM) [20], 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 stands for the external influence. We show that if all external influences follow joint Gaussian distribution, then $x \perp\!\!\!\perp y|Z$ if and only if $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. We further show that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ is equivalent to $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ as well as $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$. 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 $x \perp\!\!\!\perp y|Z$ to a simpler unconditional independence test of $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ without considering d-separation. Furthermore, if all these external influences follow non-Gaussian distributions and the model satisfies structural faithfulness condition, we show that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Leftrightarrow x \perp\!\!\!\perp y|Z$.

It is well known that the existing causal discovery methods based on CI tests usually return a set of Markov equivalence classes [23] by detecting a set of V-structures and their consistent propagations [4, 14]. With the theoretical results above, we show that CI testing based on independent residuals contains useful information for causal direction inference. When CI testing-based methods like PC algorithm [22] return a causal skeleton, if two variables x and z are directly connected and $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$ holds, we can easily deduce that z is a cause of x in the non-Gaussian case.

In summary, compared to 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, when being applied to causal discovery, this method can infer more causal directions than the existing methods.

The rest of this article is organized as follows: Section 2 reviews the major related work. Section 3 introduces our method of measuring conditional independence by independent residuals. Section 4 provides the method of learning causal structure based on our CI testing method. Section 5 evaluates the proposed method. Section 6 concludes this article.

2 RELATED WORK

In this section, we briefly review the existing regression-based CI testing methods.

In Reference [12], the authors proved that if there exists a function f such that $x - f(Z) \perp\!\!\!\perp (y, Z)$ then $x \perp\!\!\!\perp y|Z$. Reference [27] further showed that if there exists two functions f and g such that $x - f(Z) \perp\!\!\!\perp (y - g(Z), Z)$ then $x \perp\!\!\!\perp y|Z$. We note that these two conclusions do not require additive noise assumption. However, in non-additive noise model (e.g., post-nonlinear model [28]), it is hard to find an appropriate function f or g from the corresponding spaces of square integrable functions. With the assumption of additive noise, 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, $x - f(Z) \perp\!\!\!\perp Z$ is a strong condition, as we can deduce Z causes x from $x - \mathbb{E}(x|Z) \perp\!\!\!\perp Z$ in many cases [28]. 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, even in linear non-Gaussian cases, we often need to conduct $|y| + |Z|$ marginal independence tests to check whether $x - f(Z) \perp\!\!\!\perp (y, Z)$ holds.

In Reference [9], the authors showed that given structural faithfulness and Markov condition [15] with additive noise assumption, whenever Z causes x or y, it follows that $x \perp\!\!\!\perp y|Z$ if and only if $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. Similarly, here a strong condition that Z causes x or y is assumed. We can see that if these conditions are given, then it is easy to derive the corresponding causalities. Moreover, faithfulness condition means $x \perp\!\!\!\perp y|Z \Rightarrow x$ and y are d-separated by Z, and Markov condition implies y are d-separated by $Z \Rightarrow x \perp\!\!\!\perp y|Z$, so CI is relaxed to d-separation given the faithfulness and Markov assumptions [15]. 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 Reference [17], 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 Reference [27], the authors further conjectured that $x - f(Z) \perp\!\!\!\perp y - g(Z)$ can lead to $x \perp\!\!\!\perp y|Z$ under nonlinear and faithfulness conditions, where f and g are arbitrary nonlinear functions, and x, y, and Z are generated by following nonlinear additive noise model [16, 28]. There are also some other works that study the covariance between residuals [18] or the independence between residuals [8]. However, these works focus on how to approximate CIs, rather than to detect the equivalence between CI and independence (or uncorrelatedness) of residuals.

We can see that all the above methods aim to measure CI by regression and they require respective preconditions or assumptions. Recently, we studied the relationship between conditional independence (CI) $x \perp\!\!\!\perp y|Z$ and the independence of two residuals $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, where x and y are two random variables, and Z is a set of random variables, and proposed a new method for causality discovery by checking CI based on independent residuals [26].

This article is an extension to Reference [26]. Our extensions lie in two aspects:

(1) Methodology: On the one hand, we extend the original conclusion in Reference [26]: $x \perp\!\!\!\perp y|Z \Leftrightarrow x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ in Gaussian case to a more general conclusion: $x \perp\!\!\!\perp y|Z, x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z), x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ and $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$ are equivalent to each other. However, we propose to measure CI in non-Gaussian case without assuming the faithfulness condition: $x \perp\!\!\!\perp y|Z$ if $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ or $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ (for any z in Z) and $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

(2) Experiments: In Reference [26], we compared our method to only one causal learning method the PC algorithm that cannot distinguish Markov equivalence classes. In this manuscript, we further compare our method with other four mainstream causal structure learning methods LiNGAM [19], DLiNGAM [20], Sparse-ICA LiNGAM [29], and SADA [3], which all are able to distinguish Markov equivalence classes. And the experimental results show that our method performs much better than these four existing causal learning methods.

3 MEASURING CONDITIONAL INDEPENDENCE BY INDEPENDENT RESIDUALS

Generally, the linear structural equation model (SEM) is defined as a tuple $(S, P(X))$, where $S = \{S_1, S_2, \ldots, S_n\}$ is a collection of n equations, $S_i : x_i = f_i(pa_{x_i}) + \varepsilon_i$, $i = 1, 2, \ldots, n$, where each f_i is a linear function, 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 i.i.d., and $\varepsilon_i \perp\!\!\!\perp pa_{x_i}$. SEM reflects the data-generating processes of X in the DAG G. We say a SEM is identifiable if it is asymmetrical in cause and effect and is capable of distinguishing them. In fact, SEM is generally identifiable in non-Gaussian cases, all the identifiable and non-identifiable cases

are summarized in Reference [28] (let the invertible mapping in Post-Nonlinear causal model [28] be identity mapping).

We consider such a scenario: Given a DAG G where the data-generating procedure follows SEM, there are two randomly selected nodes x_i and x_j, we want to test whether x_i and x_j are conditionally independent given a set of variables Z. By default, throughout this article we assume that all variables follow SEM. In what follows, we present the theoretical results for characterizing CIs (i.e., $x_i \perp\!\!\!\perp x_j|Z$) from the perspective of SEM, which underlies the proposed new method.

Here, we first present Daudin's work [6] that gives the characterization of conditional independence by explicitly enforcing the uncorrelatedness of functions in suitable spaces, because it is used to derive our theorems.

Characterization of conditional independence (CCI) [6]. Let X, Y, and Z be three real random variables or sets of random variables, $E_1 = \{g \in L^2_{XZ}, \mathbb{E}(g|Z) = 0\}$, $E_2 = \{h \in L^2_{YZ}, \mathbb{E}(h|Z) = 0\}$, $E_3 = \{g' \in L^2_X, \mathbb{E}(g') = 0\}$, and $E_4 = \{h' \in L^2_Y, \mathbb{E}(h') = 0\}$, 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) $X \perp\!\!\!\perp Y|Z$;
(2) $\forall g \in E_1$ and $\forall h \in E_2$, $\mathbb{E}(gh) = 0$;
(3) $\forall g \in E_1$ and $\forall h' \in E_4$, $\mathbb{E}(gh') = 0$;
(4) $\forall h \in E_2$ and $\forall g' \in E_3$, $\mathbb{E}(hg') = 0$.

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

In what follows, we review the theoretical results [1] on the relationship between CI and independent residuals in Gaussian and non-Gaussian cases, respectively.

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

It can be seen that if $x \perp\!\!\!\perp y|Z$, then $\forall g \in E_1$ and $\forall h \in E_2$, $\mathbb{E}(gh) = 0$ according to the condition (2) in CCI. As $\mathbb{E}(x - \mathbb{E}(x|Z)|Z) = 0$ and $\mathbb{E}(y - \mathbb{E}(y|Z)|Z) = 0$, then $x - \mathbb{E}(x|Z) \in E_1$ and $y - \mathbb{E}(y|Z) \in E_2$, we have $cov\{(x - \mathbb{E}(x|Z))(y - \mathbb{E}(y|Z))\} = E\{(x - \mathbb{E}(x|Z))(y - \mathbb{E}(y|Z))\} - \mathbb{E}(x - \mathbb{E}(x|Z))\mathbb{E}(y - \mathbb{E}(y|Z)) = 0$. Thus in the Gaussian case, we have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. 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 $\sigma_{xy.Z} = cov(x - \mathbb{E}(x|Z), y - \mathbb{E}(y|Z)) = 0$. In the linear Gaussian case, zero partial correlation is equivalent to the conditional independence [1], we therefore obtain $x \perp\!\!\!\perp y|Z$.

Remark 1 implies that CI and the independence between two residuals are equivalent in the Gaussian case, i.e., CI can be simply checked by two times' regression and one time's unconditional independence test, $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. In what follows, we will show that this condition can be further simplified as $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$, which is also can be easily derived from Reference [1]. That is, we can reduce one time's regression.

Remark 2. Define $m + 2$ random variables x, y and $Z = \{z_1, \ldots, z_m\}$ as linear combinations of independent random variables s_i $(i = 1, \ldots, l)$, if all s_i follow joint Gaussian distribution, then the following conditions are equivalent to each other:

(1) $X \perp\!\!\!\perp Y|Z$;
(2) $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$;

(3) $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$;
(4) $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

Without loss of generality, assume x, y and $\forall z \in Z$ have zero mean. It can be seen that if $x \perp\!\!\!\perp y|Z$, then $\forall g \in E_1$ and $\forall h' \in E_4$, $\mathbb{E}(gh') = 0$ according to the condition (3) in CCI. As $\mathbb{E}(x - \mathbb{E}(x|Z)|Z) = 0$ and $\mathbb{E}(y - \alpha) = 0$ where $\alpha = mean(y)$, then $x - \mathbb{E}(x|Z) \in E_1$ and $y - \alpha \in E_4$, we have $cov\{(x - \mathbb{E}(x|Z))(y - \alpha)\} = E\{(x - \mathbb{E}(x|Z))(y - \alpha)\} - \mathbb{E}(x - \mathbb{E}(x|Z))\mathbb{E}(y - \alpha) = 0$. Thus in the Gaussian case, we have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \alpha$. As α is a constant term, $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \alpha \Rightarrow x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$. However, 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

$$\begin{aligned}
\sigma_{xy.Z} &= cov(x - \mathbb{E}(x|Z), y - \mathbb{E}(y|Z)) \\
&= E\{(x - \mathbb{E}(x|Z))(y - \mathbb{E}(y|Z))\} - E\{(x - \mathbb{E}(x|Z))\}E\{(y - \mathbb{E}(y|Z))\} \\
&= E\{(x - \mathbb{E}(x|Z))(y - \mathbb{E}(y|Z))\} \\
&= E\{(x - \mathbb{E}(x|Z))y\} - E\{(x - \mathbb{E}(x|Z))\mathbb{E}(y|Z)\} \\
&= 0 - E\left\{(x - \mathbb{E}(x|Z))\Sigma_{k=1}^{m}\alpha_k z_k\right\} \ (given\ x - \mathbb{E}(x|Z) \perp\!\!\!\perp y) \\
&= -\Sigma_{k=1}^{m}\alpha_k E\{(x - \mathbb{E}(x|Z))z_k\} \\
&= 0 \ (as\ z_k\ is\ Z - measurable)
\end{aligned} \tag{1}$$

In the linear Gaussian case, zero partial correlation is equivalent to the conditional independence [1], we therefore have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y \Rightarrow x \perp\!\!\!\perp y|Z$. Similarly, we have $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x \Leftrightarrow x \perp\!\!\!\perp y|Z$.

This conclusion means that in linear Gaussian case, we can test the conditional independence between x, y and Z simply by $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$. For example, given a linear Gaussian structural equation model of $x \leftarrow z \leftarrow y$ where $z = \alpha y + \varepsilon_z$ and $x = \alpha\beta y + \beta\varepsilon_z + \varepsilon_x$. We have

$$\begin{aligned}
&E\{(y - \mathbb{E}(y|z))x\} \\
&= E\{(y - cov(y, \alpha y + \varepsilon_z)(\alpha y + \varepsilon_z)/var(\alpha y + \varepsilon_z)) \\
&\quad \times (\alpha\beta y + \beta\varepsilon_z + \varepsilon_x)\} \\
&= \alpha\beta\mathbb{E}(y^2) - \left(\alpha^3\beta\mathbb{E}(y^2)^2/\left(\alpha^2\mathbb{E}(y^2) + \mathbb{E}\left(\varepsilon_z^2\right)\right)\right) \\
&\quad - \left(\alpha\beta\mathbb{E}(y^2)\mathbb{E}\left(\varepsilon_z^2\right)/\left(\alpha^2\mathbb{E}(y^2) + \mathbb{E}\left(\varepsilon_z^2\right)\right)\right) \\
&= \alpha\beta\mathbb{E}(y^2) - \alpha\beta\mathbb{E}(y^2) \\
&= 0
\end{aligned} \tag{2}$$

We can see that $y - \mathbb{E}(y|z)$ contains the disturbance term ε_z, and so does x. However, $y - \mathbb{E}(y|z)$ is still independent of x according to Remark 2. Next, we consider the non-Gaussian case. Here, we first reinstate the Darmois-Skitovitch theorem [5, 21] as it is used to prove the subsequent theorems.

Darmois-Skitovitch theorem (DST). Define two random variables x and y as linear combinations of independent random variables s_i $(i = 1, \ldots, l)$, $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$. Then, if $x \perp\!\!\!\perp y$, 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 1. *Define $m + 2$ random variables x, y and $Z = \{z_1, \ldots, z_m\}$ generated by following a l-dimensional linear structural equation model, if all the external influences s_i $(i = 1, \ldots, l)$ are*

non-Gaussian, then $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ *or* $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ *(*$\forall z \in Z$*) and* $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Rightarrow x \perp\!\!\!\perp y|Z$.

Proof. In linear regression, the conditional expectations $\mathbb{E}(x|Z)$ and $\mathbb{E}(y|Z)$ are linear combinations of the independent variables $z_1, \ldots, z_m$. We therefore have $x - \mathbb{E}(x|Z) = \sum_{i=1}^{l} a_i s_i$ and $y - \mathbb{E}(y|Z) = \sum_{i=1}^{l} b_i s_i$. Given $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ and $\forall j$, if $a_j \neq 0$, then there must be $b_j = 0$ according to DST. Without loss of generality, assume $Z = \{z_1, z_2\}$, then there are two cases: (1) $Z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ (or $Z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$) and (2) $z_1 \perp\!\!\!\perp x - \mathbb{E}(x|Z)$, $z_2 \not\perp\!\!\!\perp x - \mathbb{E}(x|Z)$, $z_1 \not\perp\!\!\!\perp y - \mathbb{E}(y|Z)$ and $z_2 \perp\!\!\!\perp y - \mathbb{E}(x|Z)$.

Case 1. If $Z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$, for the conditional mutual information $I(x; y|Z)$ of x and y given Z, then we have

$$\begin{aligned} & I(x; y|Z) \\ &= I(x - \mathbb{E}(x|Z); y - \mathbb{E}(y|Z)|Z) \\ &= I(x - \mathbb{E}(x|Z); y - \mathbb{E}(y|Z), Z) - I(x - \mathbb{E}(x|Z); Z). \end{aligned} \tag{3}$$

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

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

$$\begin{aligned} & I(x; y|Z) \\ &= I(x - \mathbb{E}(x|Z); y - \mathbb{E}(y|Z), z_1, z_2) - I(x - \mathbb{E}(x|Z); z_1, z_2) \\ &= I(x - \mathbb{E}(x|Z); z_2) - I(x - \mathbb{E}(x|Z); z_2) = 0 \\ &\Rightarrow x \perp\!\!\!\perp y|Z. \end{aligned} \tag{4}$$

□

In what follows, we further demonstrate that there must be at least one Z in a certain SEM to meet the two conditions $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ or $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ ($\forall z \in Z$) and $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

Theorem 2. *If* x_i *and* x_j *are neither directly connected nor unconditionally independent, then there must exist a set of variables* Z *and two functions* f *and* g *such that* $x_i - f(Z) \perp\!\!\!\perp x_j - g(Z)$*, and* $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ *or* $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ *(*$\forall z \in Z$*).*

Proof. Without loss of generality, assume that x_j is an ancestor of x_i, and let pa_{x_i} denote the set of direct parents of x_i. Following the data-generating process of ANM, we have $x_i = f(pa_{x_i}) + \varepsilon_i$ and $\varepsilon_i \perp\!\!\!\perp pa_{x_i}$, i.e., $x_i - f(pa_{x_i}) \perp\!\!\!\perp pa_{x_i}$. For the reason that ε_i is an exogenous additive noise independent of x_i and all its non-descendant nodes, we have $\varepsilon_i \perp\!\!\!\perp (x_j, pa_{x_i})$. Thus, given an arbitrary function g, we have $x_i - f(pa_{x_i}) \perp\!\!\!\perp x_j - g(pa_{x_i})$. Similarly, if x_i is an ancestor of x_j, or x_i and x_j share common ancestors, we can also obtain $x_i - f(pa_{x_i}) \perp\!\!\!\perp (pa_{x_i}, x_j - g(pa_{x_i}))$. Therefore, let pa_{x_i} (or pa_{x_j}) be Z, we complete the proof of this theorem. □

Actually, $Z = pa_{x_i}$ or $Z = pa_{x_j}$ is just a specific case of Z that satisfies the condition of Theorem 2. In many cases, we need not restrict $Z = pa_{x_i}$ or $Z = pa_{x_j}$ to meet $x_i - \mathbb{E}(x_i|Z) \perp\!\!\!\perp x_j - \mathbb{E}(x_j|Z)$ and $x_i - \mathbb{E}(x_i|Z) \perp\!\!\!\perp Z$ or $x_j - \mathbb{E}(x_j|Z) \perp\!\!\!\perp Z$. For example, given a DAG of $x_1 \to z_1 \to x_2$ and $z_2 \to x_2$, if x_2 can be expressed by $x_2 = a_1 * z_1 + a_2 * z_2 + \varepsilon$, let $Z = z_1$, we can also obtain $x_1 - \mathbb{E}(x_1|Z) \perp\!\!\!\perp x_2 - \mathbb{E}(x_2|Z)$ and $x_2 - \mathbb{E}(x_2|Z) \perp\!\!\!\perp Z$.

In contrast to Remark 1, Theorem 1 and 2 imply that if we want to check CI by two independent residuals in non-Gaussian case, we need one more additional condition $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ or $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ ($\forall z \in Z$). However, there is a problem that such a Z does not always exist for all CIs, i.e., we may encounter a situation that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ holds but $\exists z \in Z$ for $z \not\perp\!\!\!\perp x - \mathbb{E}(x|Z)$ and $z \not\perp\!\!\!\perp y - \mathbb{E}(y|Z)$. In what follows, we show that the condition $z \perp\!\!\!\perp x - \mathbb{E}(x|Z)$ or $z \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ in Theorem 1 can be removed when the faithfulness condition is given.

Theorem 3. *Define $m+2$ random variables x, y, and $Z = \{z_1, \ldots, z_m\}$ generated by following a l-dimensional linear structural equation model that satisfies the faithfulness condition, if all the external influences s_i ($i = 1, \ldots, l$) are non-Gaussian, then $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Leftrightarrow x \perp\!\!\!\perp y|Z$.*

Proof. We first consider $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Rightarrow x \perp\!\!\!\perp y|Z$. Assume that x is directly connected to y, without loss of generality, $x \rightarrow y$. Let ε_x be the exogenous disturbance of x, then $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp \varepsilon_x$ or faithfulness is violated. We further have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Rightarrow y - \mathbb{E}(y|Z) \perp\!\!\!\perp \varepsilon_x$ according to DST. This means that y and ε_x are d-separated by Z under the faithfulness condition. This is contradictory, as ε_x directly causes x.

However, if x is not adjacent to y and $x \not\perp\!\!\!\perp y|Z$, then there must be one of the following two scenarios: (1) at least one active path P between x and y cannot be blocked by Z or any subset of Z, or (2) Z contains a collider (or a descendant of a collider) w.r.t. x and y.

For the first scenario, let $Z \cup c$ be a minimal d-separator where c blocks P. We have $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp c$ and $y - \mathbb{E}(y|Z) \not\perp\!\!\!\perp c$ or x (or y) and c can be d-separated by Z that violates the faithfulness condition. We therefore have $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp y - \mathbb{E}(y|Z)$ according to DST, which is also contradictory.

For the second scenario, if x and y have the same collider, then $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$ must contain a common exogenous disturbance of a certain descendant. Therefore, we have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z) \Rightarrow x \perp\!\!\!\perp y|Z$.

The conclusion above indicates that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ can lead to $x \perp\!\!\!\perp y|Z$ in linear non-Gaussian case given the faithfulness condition. However, there is still a problem, that is, whether $x \perp\!\!\!\perp y|Z$ can cause $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. In what follows, we prove that $x \perp\!\!\!\perp y|Z \Rightarrow x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

As x and y cannot be adjacent given $x \perp\!\!\!\perp y|Z$ under the faithfulness assumption, we first consider all of the possible active paths between a pair of variables x and y that go through a third variable z. There are three cases: (1) x is an indirect cause of y (or y is an indirect cause of x), formally, $x \dashrightarrow z \dashrightarrow y$ (or $x \dashleftarrow z \dashleftarrow y$); (2) z is a common cause of x and y, i.e., $x \dashleftarrow z \dashrightarrow y$; (3) z is a collider regarding x and a subset of the d-separator, that is, $x \dashrightarrow z \dashleftarrow c \dashleftarrow d \dashrightarrow y \;\&\; z \dashrightarrow y$.

For both Case 1 and Case 2, without loss of generality, assume that x indirectly causes y and let z be a minimal d-separator. Then, z must be a set of descendants of x (or y) and a set of ancestors of y (or x). We can remove all the disturbances involved in Z from y by $y - \mathbb{E}(y|Z)$, obtain $y - \mathbb{E}(y|Z) \perp\!\!\!\perp Z$, and further deduce $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

For Case 3, $Z = \{z, c\}$ forms a minimal d-separator of x and y. Because of the transitivity in linear ANM, $y - \mathbb{E}(y|z, c)$ contains no information of x and ε_z (ε_z denotes the disturbances of z). Here, x can also be treated as an additive disturbance term of z. On the other side, the information that c can provide about x is $z - \mathbb{E}(z|c))$, that is, $x - \mathbb{E}(x|z, c) = x - \mathbb{E}(x|z, z - \mathbb{E}(z|c))$ contains only the information of x and ε_z. Therefore, $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. □

In what follows, we give several examples to describe the above theorem. As shown in Figure 1(a), $Z = \{z_1, z_2\}$, we can remove all the disturbances involved in Z from y by $y - \mathbb{E}(y|Z)$, this is, $y - \mathbb{E}(y|Z) \perp\!\!\!\perp Z$, and can further deduce that $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. As shown in Figure 1(b), where z_1 is a common cause between z_3 and y, and z_3 is a collider w.r.t. x and z_2 as well

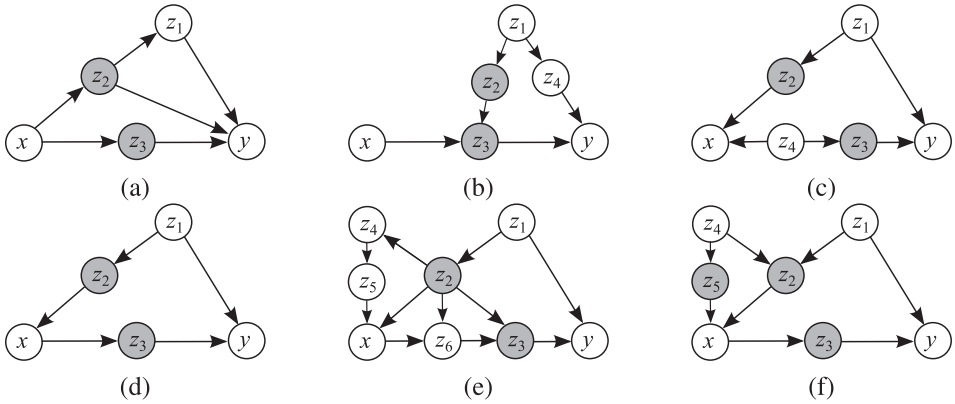


Fig. 1. Six local causal structures.

as z_1 along $x \dashrightarrow z_3 \dashleftarrow z_1 \dashrightarrow y$. We can see that

$$\begin{aligned} & x - \mathbb{E}(x|z_2, z_3) \\ = & x - \mathbb{E}(x|z_2, z_3, z_2 - \mathbb{E}(z_2|z_3)) \\ = & x - \mathbb{E}(x|z_2, z_3, *x + \varepsilon_{z_3}) \\ = & x - \mathbb{E}(x| * x + \varepsilon_{z_3}) \ (as\ x \perp\!\!\!\perp z_2\ and\ x \perp\!\!\!\perp z_3|\varepsilon_{z_3}) \\ = & f(x, \varepsilon_{z_3}). \end{aligned} \tag{5}$$

where ε_{z_3} is the disturbance term corresponding to z_3, f is a linear function, and for simplicity we denote a coefficient times x as $*x$.

However, we have

$$\begin{aligned} & y - \mathbb{E}(y|z_2, z_3) \\ = & y - \mathbb{E}(*z_3 + *z_4 + \varepsilon_y|z_2, z_3) \\ = & y - \mathbb{E}(*z_3|z_2, z_3) - \mathbb{E}(*z_4 + \varepsilon_y|z_2, z_3) \\ = & *z_4 + \varepsilon_y - \mathbb{E}(*z_4 + \varepsilon_y|z_2, z_3) \\ = & *z_4 + \varepsilon_y - \mathbb{E}(*z_4 + \varepsilon_y|z_2) \ (as\ *z_4 + \varepsilon_y \perp\!\!\!\perp z_3|\varepsilon_{z_2}) \\ = & g(z_2, z_4, \varepsilon_y). \end{aligned} \tag{6}$$

That is $f(x, \varepsilon_{z_3}) \perp\!\!\!\perp g(z_2, z_4, \varepsilon_y)$, i.e., $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. Similarly, if we change the d-separator Z by replacing z_2 with z_1 or z_4, we can still obtain the same result.

As shown in Figure 1(c), z_1 and z_4 are two common causes of x and y, and $Z = \{z_2, z_3\}$. In this case, we can remove the disturbances involved in z_2 from x by $x - \mathbb{E}(x|Z)$, and those involved in z_3 from y by $y - \mathbb{E}(y|Z)$, and obtain $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. If we combine Case 3 with Case 1 and Case 2, as shown in Figure 1(d), then we can also see that $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$ have no common disturbance, because

$$\begin{aligned} & x - \mathbb{E}(x|z_2, z_3) \\ = & *z_2 + \varepsilon_x - \mathbb{E}(*z_2|z_2, z_3) - \mathbb{E}(\varepsilon_x|z_2, z_3) \\ = & \varepsilon_x - \mathbb{E}(\varepsilon_x|z_2, z_3, z_3 - \mathbb{E}(z_3|z_2)) \\ = & \varepsilon_x - \mathbb{E}(\varepsilon_x| * \varepsilon_x + \varepsilon_{z_3}) \\ = & f(\varepsilon_{z_3}, \varepsilon_x). \end{aligned} \tag{7}$$

And, similarly, we have $y - \mathbb{E}(y|z_2, z_3) = g(z_1, z_2, \varepsilon_y)$.

In practice, these three cases above can be generalized to more complicated cases, such as in Figure 1(e) and (f), from which we can also derive similar result by doing $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$ to remove their common disturbances.

Remark 1 and 2 indicate that we can do a CI test by just testing the independence of two residuals in linear Gaussian cases, and Theorem 3 means that CI can still be measured by two independence residuals in linear non-Gaussian case under the faithfulness condition. We denote this CI test method as **ReCIT** (the abbreviation of **Re**sidual-based **C**onditional **I**ndependence **T**est). In the next section, we apply ReCIT to causal discovery. We show that CI contains useful information about causal direction, which can distinguish Markov equivalent classes.

4 CAUSAL DISCOVERY BASED ON RECIT

We have the following theorem:

Theorem 4. *Given $m + 2$ random variables x, y, and $Z = \{z_1, \ldots, z_m\}$ 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 $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, then we have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$ (or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp z$) $\Rightarrow$ z causes x (or z causes y).*

Proof. Without loss of generality, we assume $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$. Let ε denote the exogenous disturbance of z. If $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp \varepsilon$, then $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp z$ according to Darmois-Skitovitch theorem [5, 21]. We therefore have $x - \mathbb{E}(x|Z) \perp\!\!\!\perp \varepsilon$, which means x and ε can be d-separated by Z under the faithfulness condition. If z is a child of x, then $x \to z \leftarrow \varepsilon$ forms a V-structure where z is a collider, then there must be $x - \mathbb{E}(x|Z) \not\perp\!\!\!\perp \varepsilon$ or faithfulness is violated. This 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 $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp z$ ($\forall z \in Z$) of Theorem 4 is different from the assumptions or preconditions in the existing regression-based CI testing methods [9, 12, 27], which all require $x - \mathbb{E}(x|Z) \perp\!\!\!\perp Z$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp Z$, or Z causes x or y.

Comparing 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: $x_1 \leftarrow x_2 \to x_3$, it is easy to find $x_1 - \mathbb{E}(x_1|x_2) \perp\!\!\!\perp x_2$ and $x_3 - \mathbb{E}(x_3|x_2) \perp\!\!\!\perp x_2$. We therefore can infer $x_1 \leftarrow x_2$ and $x_2 \to x_3$. However, it is difficult for the existing CI testing methods to distinguish the three structures $x_1 \leftarrow x_2 \to x_3$, $x_1 \leftarrow x_2 \leftarrow x_3$ and $x_1 \to x_2 \to x_3$, because all of them fit the observed conditional and unconditional independence, though obviously have 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 [30]) to test unconditional independence. Concretely, we calculate $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$ simply by least square regression, i.e., $x - \mathbb{E}(x|Z) = x - Z(Z^TZ)^{-1}Z^Tx$ and $y - \mathbb{E}(y|Z) = y - Z(Z^TZ)^{-1}Z^Ty$. And any independence testing method can be used to test the independence between $x - \mathbb{E}(x|Z)$ and $y - \mathbb{E}(y|Z)$.

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 $x_i - \mathbb{E}(x_i|Z)$ and $x_j - \mathbb{E}(x_j|Z)$ 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 the local structure $x_i - x_k - x_j$ can form a V-structure, then orient it as $x_i \to x_k \leftarrow x_j$. For a structure $x_i - x_k$, as

ALGORITHM 1: The PC algorithm based on ReCIT (PC_{ReCIT})

Input: a set of variables $X = \{x_1, \ldots, x_n\}$, a threshold k.
Output: a partial DAG G.

1: Form the complete undirected graph G on the variables set X.
2: **for** $\forall x_i, x_j \in X$ and adjacent in G **do**
3: **if** $\exists Z \subseteq X \setminus \{x_i, x_j\}$ and ($|Z| < k$) such that $x_i - \mathbb{E}(x_i|Z) \perp\!\!\!\perp x_j - \mathbb{E}(x_j|Z)$ **then**
4: remove edge $x_i - x_j$ from G and record Z in $Sepset(x_i, x_j)$ and record $x_i - \mathbb{E}(x_i|Z)$ and $x_j - \mathbb{E}(x_j|Z)$ in temporary sets $Temp_{x_i,Z}$ and $Temp_{x_j,Z}$.
5: **end if**
6: **end for**
7: **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**
8: **if** $x_k \notin Sepset(x_i, x_j) \cup Sepset(x_j, x_i)$ **then**
9: orient $x_i - x_k - x_j$ as $x_i \rightarrow x_k \leftarrow x_j$.
10: **end if**
11: **end for**
12: **for** $\forall x_i, x_k \in X$ such that x_i and x_k are adjacent **do**
13: **if** $\exists Z$ such that $x_i - \mathbb{E}(x_i|Z) \in Temp_{x_i}$ and $x_k \in Z$ **then**
14: **if** $x_i - f(Z) \perp\!\!\!\perp x_k$ **then**
15: orient $x_i - x_k$ as $x_i \leftarrow x_k$.
16: **end if**
17: **end if**
18: **end for**
19: Do consistent propagation.

$x_i - \mathbb{E}(x_i|Z) \perp\!\!\!\perp x_k$ ($x_k \in Z$) implies $x_i \leftarrow x_k$, if $x_i - \mathbb{E}(x_i|Z) \in Temp_{x_i}$, we test the dependence between $x_i - \mathbb{E}(x_i|Z)$ and x_k. If the 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).

5 PERFORMANCE EVALUATION

We first conduct extensive experiments to evaluate ReCIT, and compare it with KCIT [30]. We also compare the causal inference performance of ReCIT and KCIT under the PC algorithm framework [22], 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 the other existing CI testing methods in the literature [7, 24, 27, 30]. In the implementation of ReCIT, we do regression using least square regression, and do unconditional independence tests using KCIT.

5.1 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 it is false) errors of ReCIT change with the size of the conditioning set Z ($D = 1, 2, \ldots, 5$) and the sample size ($n = 100$ and 200) 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 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 $a \sim U(0.2, 1)$ is different for X and Y, and $\varepsilon \sim U(-0.2, 0.2)$. Hence, $X \perp\!\!\!\perp Y|Z$ holds. In our simulation, Z_i is i.i.d. $U(0, 1)$.

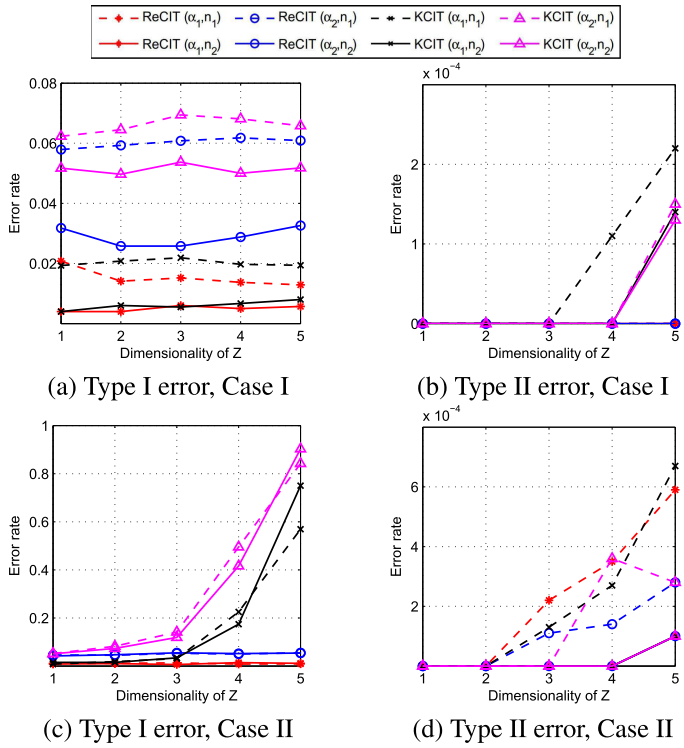


Fig. 2. The probabilities of Type I and Type II errors obtained by simulation in various situations. The significance level $\alpha_1 = 0.01$, $\alpha_2 = 0.05$ and the sample size $n_1 = 100$, $n_2 = 200$. Top: Case I (only one variable in Z is effective to X and Y). Bottom: Case II (all variables in Z are effective).

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 error of both type I and type II. 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 1,000 times and average their results.

Type I and II errors are calculated like this: take $D = 3$ for example, 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-$the\ number\ of\ CIs/4$. However, 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 Figure 2(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.

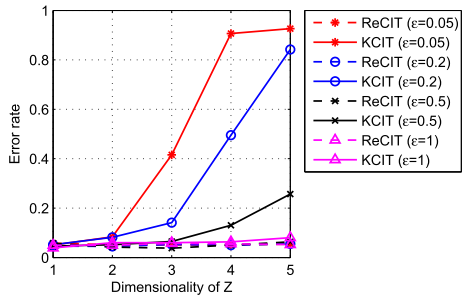


Fig. 3. The probability of Type II error in Case II (all variables in Z are effective) for different noise weights.

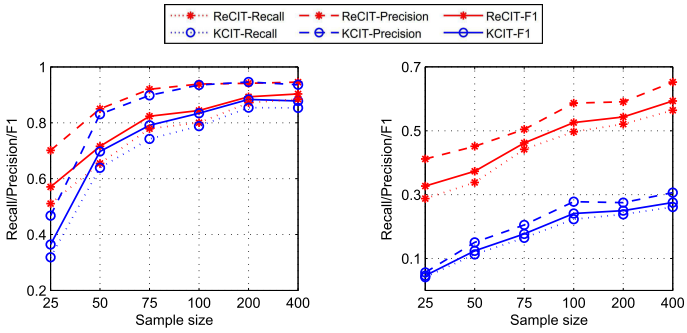
To further illustrate why ReCIT can perform much better than KCIT in terms of Type I error in Case II (see Figure 2(c)), we conduct another experiment to evaluate the two methods under different noise weights. We simulate four sets of noise, $\varepsilon_1 \sim U(-0.05, 0.05)$, $\varepsilon_2 \sim U(-0.2, 0.2)$, $\varepsilon_3 \sim U(-0.5, 0.5)$ 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 Figure 3. 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. However, 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, $x - \mathbb{E}(x|Z) = \varepsilon_x$ and $y - \mathbb{E}(y|Z) = \varepsilon_y$, then testing $x \perp\!\!\!\perp y|Z$ 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. Figure 2(a) and (c) and Figure 3 show that ReCIT performs better and is more robust than KCIT in different situations in terms of Type I error.

However, as shown in Figure 2(b) and Figure 2(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. However, as the sample size increases from 100 to 200, the probability of Type II error approaches zero. In particular, as shown in Figure 2(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 shown in the next subsection.

5.2 Performance in Causal Discovery

CI tests are frequently used in causal inference where we assume that the true causal structure of n random variables $x_1, \ldots, x_n$ 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



(a) Discovering skeleton (b) Discovering PDAG

Fig. 4. Performance comparison between PC_{ReCIT} and PC_{KCIT} with various sample sizes in discovering (a) causal skeleton and (b) PDAG.

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 [15]. 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 $x_1, \ldots, x_4$ 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 1,000 DAGs and evaluate the performance of the two methods PC_{ReCIT} and PC_{KCIT} on discovering causal skeletons and PDAGs (including identifiable causal directions).

As shown in Figure 4(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 $Z(Z^TZ)^{-1}Z^T$ 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 skeletons when the sample size is small, which is the frequently-encountered case in reality.

In practice, we have tested our method on many other simulation datasets generated by following a similar procedure with different parameters, and finally obtained the results that are very close to those shown in Figure 4(a).

We also evaluate the two methods in discovering PDAGs. The results are presented in Figure 4(b). We can see that PC_{ReCIT} achieves better result in all cases, though the performance of PC_{KCIT} in discovering causal skeletons 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 [15]. In other words, PC_{KCIT} returns only a set of Markov equivalence classes, while PC_{ReCIT} can uncover more causal directions according to Theorem 4.

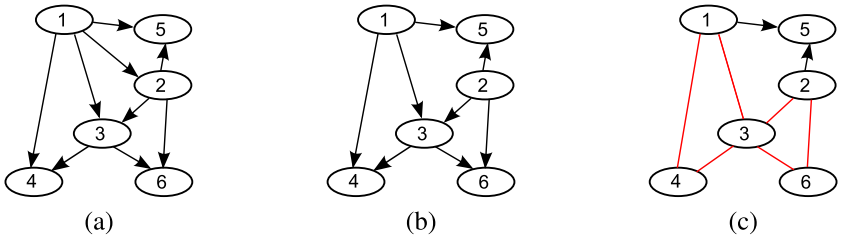


Fig. 5. Performance comparison in causal direction inference. (a) The ground-truth causal model; (b) the reconstructed DAG based on PC_{ReCIT}; (c) the reconstructed PDAG based on PC_{KCIT}. Here, the red lines are the edges whose directions are not determined.

Finally we apply PC_{ReCIT} to a causal graph presented in [19], which was generated by following a linear non-Gaussian structure equation model w.r.t. a DAG consisting of six variables as shown in Figure 5(a). We select this graph because it contains no V-structure, which is used to further show the advantage of ReCIT in inferring causal directions. The resulting skeletons reconstructed by PC_{ReCIT} and PC_{KCIT} are shown in Figure 5(b) and Figure 5(c), respectively. We can see that all the causal edges discovered by PC_{ReCIT} are correct. However, as shown in Figure 5(c), only two directions of edges $1\rightarrow5$ and $2\rightarrow5$ are correctly inferred by PC_{KCIT}, the others are not correctly 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 incorrectly removed by both ReCIT and KCIT, i.e., the CI hypotheses w.r.t. node 1 and node 2 are not rejected although they are false, therefore $1\rightarrow5\leftarrow2$ forms a false V-structure. It can be seen that even some local structures are incorrectly inferred by PC, but ReCIT can still distinguish the real causal directions. In one word, by exploiting the advantage of ReCIT, existing constraint-based methods (e.g., the PC algorithm) can greatly improve the performance of causal discovery, as ReCIT helps to distinguish the Markov equivalence classes.

5.3 Performance on Real-world Structures

In the experiments above, we compare our method with the kernel-based CI test method KCIT in terms of learning causal skeletons as well as causal directions (with PC algorithm), and show that ReCIT helps to break Markov equivalence classes, thus can discover much more causal directions. In this subsection, we compare our method with other four causal structure learning methods, including LiNGAM [19], DLiNGAM [20], Sparse-ICA LiNGAM [29], and SADA [3]. Note that all these methods can break Markov equivalence classes, we therefore can further evaluate the performance of our method in causal discovery. The implementations of LiNGAM, DLiNGAM, and SADA strictly follow the corresponding papers [3, 19, 20]. The implementation of Sparse-ICA LiNGAM is based on the sparse-ICA of [29] and the pruning algorithm of [19]. In these existing methods, SADA is the most effective approach for learning causal structures from high dimensional cases, and LiNGAM is selected as the base solver by default. All methods are evaluated on eight real-world causal network structures[1] that cover a variety of applications, including, insurance evaluation (*Insurance*), medicine (*Alarm and Pathfinder*), agricultural industry (*Barley*), weather forecasting (*Hailfinder*), system troubleshooting (*Win95pts and Andes*) and the pedigree of breeding pigs (*Pigs dataset*). The structural statistics of these causal networks are summarized in Table 1 and the corresponding data generating process follows the previous works [3]. Note that the performance of the four counterparts is highly influenced by the ratio of the sample size

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

Table 1. Statistics of Eight Causal Network Structures

Dataset	Nodes#	Avg. degree	Max degree
Insurance	27	3.95	9
Alarm	37	2.49	6
Barley	48	3.50	8
Hailfinder	56	2.36	17
Win95pts	76	1.84	9
Pathfinder	109	3.58	106
Andes	223	3.03	12
Pigs	441	2.68	41

Table 2. Performance of Five Causal Learning Methods on Eight Causal Networks

Dataset	*Recall*					*Precision*				
	PC_R	SADA	LG	DLG	SICA	PC_R	SADA	LG	DLG	SICA
Insurance	0.32	0.24	**0.47**	0.39	0.23	**0.65**	0.45	0.11	0.09	0.06
Alarm	**0.39**	0.38	0.38	0.27	**0.39**	**0.81**	0.44	0.24	0.16	0.22
Barley	**0.37**	0.26	0.33	0.23	0.36	**0.68**	0.43	0.21	0.15	0.22
Hailfinder	**0.55**	0.50	0.25	0.20	0.31	**0.74**	0.50	0.22	0.17	0.28
Win95pts	0.37	**0.49**	0.28	0.20	0.36	**0.49**	0.45	0.37	0.25	0.35
Pathfinder	**0.92**	0.81	0.35	0.35	0.32	**0.72**	0.09	0.22	0.22	0.20
Andes	**0.76**	0.55	0.21	0.12	0.26	**0.76**	0.18	0.46	0.26	0.51
Pigs	**0.78**	0.53	0.15	N.A.	N.A.	**0.72**	0.22	0.59	N.A.	N.A.
	F1 Score									
	PC_R	SADA	LG	DLG	SICA					
Insurance	**0.41**	0.30	0.18	0.14	0.10					
Alarm	**0.49**	0.41	0.29	0.20	0.28					
Barley	**0.47**	0.31	0.26	0.18	0.27					
Hailfinder	**0.63**	0.50	0.23	0.18	0.29					
Win95pts	0.40	**0.47**	0.32	0.22	0.35					
Pathfinder	**0.78**	0.16	0.27	0.27	0.25					
Andes	**0.76**	0.26	0.28	0.16	0.34					
Pigs	**0.75**	0.34	0.24	N.A.	N.A.					

to the number of nodes, and the baseline approach LiNGAM does not work when the number of samples is as small as $|V|$. Therefore, in the following experiments, we compare PC_{ReCIT} against the four existing methods by fixing the sample size at $2|V|$.

The results are reported in Table 2, where for compressing the space in the table, PC_{ReCIT}, LiNGAM, DLiNGAM, and Sparse-ICA LiNGAM are simply denoted as PC_R, LG, DLG, and SICA, respectively. It can be seen that PC_{ReCIT} achieves significantly better *F1* score on almost all datasets, except for the case of *Win95pts* where SADA works slightly better than PC_{ReCIT}. *Win95pts* has the simplest structure among the eight causal networks, its average degree is only 1.84 (see Table 1). The performance of the splitting phase of SADA is influenced by structural complexity [3]. In other cases, especially in larger causal networks (with $|V| > 100$), PC_{ReCIT} works much better than SADA. The other three methods LiNGAM, DLiNGAM, and Sparse-ICA LiNGAM are not

competitive in all these cases. To improve their learning accuracy, we have to increase the sample size. As DLiNGAM and Sparse-ICA LiNGAM are of high time-complexity, here we do not present their results on *Pigs* network.

Moreover, we have the following two observations:

(1) The performance (*Recall*, *Precision* and *F1* score) of PC_{ReCIT} grows with the sample size rather than the ratio of the sample size to the number of nodes ($2|V|$), while the performance of the four existing methods work more steady with a fixed ratio of the sample size to the number of nodes ($2|V|$). We can see that on larger networks, *Pathfinder*, *Andes*, and *Pigs*, the *F1* score of PC_{ReCIT} is from 2 to 3 times higher than that of the four existing methods. Therefore, PC_{ReCIT} is more applicable to causal discovery in high-dimensional cases in terms of inference accuracy when limited samples are given.

(2) As the dimensionality of causal networks increases, the ratio of *Recall* to *Precision* of PC_{ReCIT} remains relatively stable, therefore the *F1* score of PC_{ReCIT} maintains an acceptable level. On the contrary, we can see that on small causal networks, the *Precision* of SADA is slightly higher than the *Recall*, while in the cases of larger causal networks, like *Pathfinder* and *Andes*, the *Precision* of SADA is much lower than *Recall*. Note that the *Recall* is the fraction of actual causality found by the algorithm, the *Precision* is the actual fraction of inferred causality with respect to the true graph. This means that SADA cannot remove many incorrect causal relationships in these cases. However, it can be seen that PC_{ReCIT} does particularly well in all these cases in terms of precision. In summary, we can again conclude that PC_{ReCIT} is more applicable to causal discovery in high-dimensional cases than the existing methods.

5.4 Verifying the Equivalence between CI and Independent Residuals

The above experiments show that ReCIT not only works better than KCIT but also discovers more true causal relationships than the state-of-the-art causal structure learning methods. In some sense, we have verified that the independence between residuals $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ can lead to CI $x \perp\!\!\!\perp y|Z$. Now, we apply ReCIT to six datasets *Insurance*, *Alarm*, *Barley*, *Hailfinder*, $Win95pts$, and *Pathfinder* aforementioned to check whether independent residuals can cause CI. As the running time of KCIT is too long, here we do not use the larger networks Andes and Pigs, and fix the maximum size of d-separator at 3. As shown in Figure 4(a), when the sample size is beyond 400, the curves of ReCIT and KCIT are too close to nearly overlap. So in this group of experiments, we compare PC_{ReCIT} against PC_{KCIT} by fixing the sample size at 400. Due to the existence of Type I and Type II error, the CIs are calculated by this process. We first run PC_{ReCIT} to discover the causal structure, whenever a CI is detected by ReCIT, KCIT is further used to recheck this CI. Similarly, we run PC_{KCIT} to detect CI and run ReCIT to recheck the corresponding CI. Consequently, the total CIs of PC_{ReCIT} is the CIs detected by PC_{ReCIT} plus the CIs rechecked by ReCIT. Similarly, we can also calculate the total CIs of PC_{KCIT}.

The results are presented in Table 3. Here, the second column and the fourth column present the results of *Recall*, *Precision* and *F1* score (R/P/F1 in short) w.r.t. causal skeleton discovery of PC_{ReCIT} and PC_{KCIT} respectively. The third column and the fifth column indicate the numbers of CIs used by the two methods. We can see that the number of CIs used by PC_{ReCIT} is very close to that used by PC_{KCIT} on all these six causal networks. If ReCIT is only sufficient but not necessary to CI, then the number of CIs of KCIT should be substantially larger than that of ReCIT. In addition, as the performance of skeleton discovery corresponds to the accuracy of CI tests, we can see that PC_{ReCIT} and PC_{KIT} show similar performance on the six networks, which indicates that testing $x \perp\!\!\!\perp y|Z$ is probably equivalent to testing $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$.

Table 3. Experimental Verification of the Equivalence between CIs and Independent Residuals

Dataset	PC_{ReCIT}		PC_{KCIT}	
	R/P/F1	#CI	R/P/F1	#CI
Insurance	0.83 / 0.93 / 0.88	551	0.83 / 0.93 / 0.88	550
Alarm	0.74 / 0.99 / 0.85	437	0.74 / 0.99 / 0.85	443
Barley	0.56 / 0.98 / 0.71	1154	0.56 / 0.98 / 0.71	1143
Hailfinder	0.83 / 0.96 / 0.89	935	0.82 / 0.97 / 0.89	950
Win95pts	0.79 / 0.98 / 0.88	641	0.81 / 0.95 / 0.88	681
Pathfinder	0.91 / 0.81 / 0.86	11060	0.91 / 0.78 / 0.85	11118

6 CONCLUSION

This article studies the relationship between conditional independence $x \perp\!\!\!\perp y|Z$ and the independence of two residuals $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. In some previous works, the independence of two residuals is regarded as a weak condition for CI under the faithfulness and Markov assumptions. To make this weak condition be sufficient, some additional conditions such as $x - \mathbb{E}(x|Z) \perp\!\!\!\perp Z$ (or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp Z$), Z causes x (or y) are required. In this work, we show that if x, y, and Z are generated via a linear structural equation model and all external influences follow joint Gaussian distribution, then $x \perp\!\!\!\perp y|Z$ if and only if $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. In this case, $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$ is also equivalent to $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y$ as well as $y - \mathbb{E}(y|Z) \perp\!\!\!\perp x$. Furthermore, if all the external influences follow non-Gaussian distributions and the model satisfies the structural faithfulness condition, then we have $x \perp\!\!\!\perp y|Z \Leftrightarrow x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$. We therefore can relax the test of $x \perp\!\!\!\perp y|Z$ to a simpler unconditional independence test of $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, without assuming any other graph-related condition. Intuitively, our result means that if $x \perp\!\!\!\perp y|Z$ holds, then after 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. However, we show that CIs can distinguish Markov equivalence classes, as in many cases $x - \mathbb{E}(x|Z) \perp\!\!\!\perp z$ or $y - \mathbb{E}(y|Z) \perp\!\!\!\perp z$ ($\forall z \in Z$ and Z is a minimal d-separator) is satisfied when $x - \mathbb{E}(x|Z) \perp\!\!\!\perp y - \mathbb{E}(y|Z)$, 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 the proposed method outperforms the kernel-based method KCIT in discovering causality in linear non-Gaussian cases. Moreover, our experimental results also demonstrate that the PC algorithm using our method for CI testing performs much better than four existing causal learning methods SADA, LINGAM, DLINGAM and Spare-ICA, which all can distinguish Markov equivalence classes.

REFERENCES

[1] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. 2004. Partial correlation and conditional correlation as measures of conditional independence. *Austr. N. Z. J. Stat.* 46, 4 (2004), 657–664.

[2] Wicher Pieter Bergsma. 2004. *Testing Conditional Independence for Continuous Random Variables*. Eurandom.

[3] Ruichu Cai, Zhenjie Zhang, and Zhifeng Hao. 2013. Sada: A general framework to support robust causation discovery. In *Proceedings of the International Conference on Machine Learning*. 208–216.

[4] David Maxwell Chickering. 2002. Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* 2, 3 (2002), 150–157.

[5] George Darmois. 2002. Analyse générale des liaisons stochastiques: Etude particulière de l'analyse factorielle linéaire. *Rev. Inst. Int. Stat.* 21, 1/2 (2002), 2–8.

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

[7] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. 2014. A permutation-based kernel conditional independence test. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*. 132–141.
[8] Jianqing Fan, Yang Feng, and Lucy Xia. 2015. A projection based conditional dependence measure with applications to high-dimensional undirected graphical models. *Arxiv Preprint Arxiv:1501.01617* (2015).
[9] Seth R. Flaxman, Daniel B. Neill, and Alexander J. Smola. 2016. Gaussian processes for independence tests with non-iid data in causal inference.*Trans. Intell.Syst. Technol.* 7, 2 (2016), 22–1.
[10] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2007. Kernel measures of conditional dependence. *Adv. Neur. Inf. Process. Syst.* 20, 1 (2007), 167–204.
[11] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*. 513–520.
[12] Moritz Grosse-Wentrup, Dominik Janzing, Markus Siegel, and Bernhard Schölkopf. 2016. Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage* 125 (2016), 825–833. https://www.sciencedirect.com/science/article/abs/pii/S1053811915009751.
[13] Hui Liu, Shuigeng Zhou, Wai Lam, and Jihong Guan. 2017. A new hybrid method for learning bayesian networks: Separation and reunion. *Knowl.-Based Syst.* 121 (2017), 185–197. https://www.sciencedirect.com/science/article/abs/pii/S0950705117300412.
[14] Christopher Meek. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 403–410.
[15] Judea Pearl. 2009. *Causality*. Cambridge University Press.
[16] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Causal inference on discrete data using additive noise models. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 12 (2011), 2436–2450.
[17] Joseph D. Ramsey. 2014. A scalable conditional independence test for nonlinear, non-gaussian data. *Arxiv Preprint Arxiv:1401.5031* (2014).
[18] Rajen D. Shah and Jonas Peters. 2018. The hardness of conditional independence testing and the generalised covariance measure. *Arxiv Preprint Arxiv:1804.07203* (2018).
[19] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7, Oct. (2006), 2003–2030.
[20] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.* 12, Apr. (2011), 1225–1248.
[21] V. P. Skitovich. 1953. On a property of the normal distribution. *DAN SSSR* 89 (1953), 217–219.
[22] Peter Spirtes, Clark N. Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Vol. 81. MIT Press.
[23] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. In *Applied Informatics*, Vol. 3. Springer, 3.
[24] Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference* 7, 1 (2019), 1–24.
[25] Liangjun Su and Halbert White. 2008. A nonparametric hellinger metric test for conditional independence. *Econ. Theory* 24, 04 (2008), 829–864.
[26] Hao Zhang, Shuigeng Zhou, and Jihong Guan. 2018. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 2029–2036.
[27] Hao Zhang, Shuigeng Zhou, and Jihong Guan. 2018. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 1250–1256.
[28] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 647–655.
[29] Kun Zhang, Heng Peng, Laiwan Chan, and Aapo Hyvärinen. 2009. ICA with sparse connections: Revisited. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*. Springer, 195–202.
[30] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2011. *Kernel-based Conditional Independence Test and Application in Causal Discovery*. AUAI Press, Corvallis, OR, 804–813.

Received August 2018; revised February 2019; accepted April 2019