IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 11, 2019, accepted December 23, 2019, date of publication January 9, 2020, date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2965093

Distinguish Markov Equivalence Classes from Large-Scale Linear Non-Gaussian Data

GUIZHEN MAI[1], YINGHAN HONG[2], PINGHUA CHEN[1], (Member, IEEE), KEXI CHEN[3], HAN HUANG[4], AND GENGZHONG ZHENG[5]

[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China
[2]School of Physics and Electronic Engineering, Hanshan Normal University, Chaozhou 521041, China
[3]School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[4]School of Software Engineering, South China University of Technology, Guangzhou 510006, China
[5]School of Computer and Information Engineering, Hanshan Normal University, Chaozhou 521041, China

Corresponding author: Yinghan Hong (honyinghan@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572144, Grant 61976052, Grant 61876207, and Grant 61973092, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2014A030306050, in part by the Science and Technology Planning Project of Guangdong Province, China, under Grant 2017A040405063, Grant 2017B030307002, Grant 2019B101001021, Grant 2015B090922014, and Grant 2015A010103015, in part by the Hanshan Normal University Doctor Starting Fund under Grant QD20190628, in part by the Hanshan Normal University Scientific Research Talents Fund under Grant Z19113, in part by the Science and Technology Planning Project of Tianhe District, Guangzhou City, Guangdong Province under Grant 201703YG029, in part by the International Cooperation Project of Guangzhou under Grant 201807010047, in part by the Guangzhou Science and Technology Project under Grant 201802010007 Grant 201804010276, and in part by the Guangdong Province Key Area Research and Development Program under Grant 2018B010109003.

ABSTRACT In the problem of causal discovery, conditional independence (CI) tests are generally used to detect the causal relationships among observed data. Due to the curse of dimensionality and the limitation of causal direction learning based on V-structure learning, it is difficult for constraint-based methods to distinguish the actual graph from a set of Markov equivalence classes. To alleviate this problem, in this work, a novel regression-based method to test CIs over linear Non-Gaussian data is proposed. The main purpose of this proposal is to relax the CI test of $x \perp y | Z$ to two unconditional independence tests $x - f(Z) \perp y - g(Z) + \Sigma H(Z)$ and $x - f(Z) + \Sigma H(Z) \perp y - g(Z)$, where f and g can be estimated by linear regression independently. In addition, we further show that $x - f(Z) \perp y - g(Z) + \Sigma H(Z)$ (or $x - f(Z) + \Sigma H(Z) \perp y - g(Z)$) can lead to $x \leftarrow Z$ (or $y \leftarrow Z$). According to this regression-based method, we design a causal structure learning algorithm to learn the actual graph instead of a set of Markov equivalence classes over the observed data. Experiments indicate that our method can detect much more causal relationships than other existing methods, especially in large-scale cases.

INDEX TERMS Causal inference, linear non-Gaussian additive noise model, Markova equivalence classes.

I. INTRODUCTION

In the problem of causal discovery, the causal relationships among given variables are usually detected through statistical independence [1], [2] (or unconditional independence, marginal independence) and conditional independence (CI). Concretely, causality between two variables X and Y can be checked by testing $x \perp y | Z$, where Z is an arbitrary set of random variables within the given variables set. If $x \perp y | Z$, generally X and Y have no directed causal relationship. By using CI tests, the existing causal structure learning method likes PC algorithm [3], can determine a rough graph with respect to the given variables set, and the rough graph might contain a set of graphs, which is called Markov equivalence class [4].

In practice, it is easy to conduct statistical independence test while CI test is much difficult [5]. Traditional methods can only be applied to discrete cases, then the CI results can be derived by $P(X, Y|Z) = P(X|Z)P(Y|Z)$ on the basis of conditional probability table. Hence there are no technical setbacks to solve discrete situations. Another way to measure CI is providing some simplified assumptions on the given variables with the continuous property. For example, the zero partial correlation is equivalent to CI under the assumption of joint Gaussian distribution [6], and the former one could

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

be easily tested. But non-linearity and non-Gaussianity are common in real-world cases, therefore these methods are often unreliable and incompetent.

Most of the existing methods use discretizing technique to solve the problems motioned above [7]–[9]. When Z contains a large set of variables, the required sample size should be very large. For example, CI is tested through the distance calculation of conditional densities estimate $P_{X|YZ} = P_{X|Z}$ [10]. However, if the conditional set became sufficient large, then it would be very difficult to estimate the conditional densities.

Recently, researchers resorted to solve the problem of CI test by using kernel-based techniques. For the test of unconditional CI, many kernel-based testing methods were proposed. The most important reason was that kernel-based techniques could represent high order moments, while the reproducing kernel Hilbert spaces [11] (RKHSs) can measure the properties of high-dimensional distributions like independence [12]. [13] used the Hilbert-Schmidt norm [14] of conditional cross-covariance operator. It is a measure of the conditional covariance of X and Y images under RKHSs correspondence function, if RKHSs were feature kernels [15], the operator norm must be zero if and only if $X \perp Y|Z$. Among the recent kernel-based methods, KCIT should be one of the most excellent methods proposed in [16]. KCIT used the relevance among regression functions to measure $X \perp Y|Z$. Let $f \in L^2_{XZ}$ and $g \in L^2_Y$ (L^2_{XZ} and L^2_Y) denoted the square integral functions spaces of (X, Z) and Y), then $E\left(\tilde{f}\tilde{g}\right) = 0$ can be inferred when $\tilde{f}(X,Z) = f(X,Z) - r_f(Z)$ and $\tilde{g}(Y,Z) = g(Y) - r_g(Z)$ held, where $r_f, r_g \in L^2_Z$ denote the corresponding regression functions. Therefore, KCIT could relax the kernel space defined by functions f, g, r_f and r_g in RKHSs. Kernel-based methods could generally detect more complete information from the given variables than discretizing-based CI tests. It had been shown from related literatures that causal inference based on kernel-based methods is able to obtain more accurate causal relationship [17]–[19].

On the issue of causal inference, researchers used to consider the causal functional model in the first place. And the additive noise model (ANM), which includes three types: linear, non-linear and discrete, is one of the widely used functional models [20]–[22]. It had been shown that many real-world datasets were likely to follow a certain ANM model [22]–[24]. Technically, ANM requires that the observed variables are generated by following a directed acyclic graph with a set of causal model functions: $X = f(Y) + \varepsilon$, where Y is the parent node of X, and is independent of the noise term ε. [25] presented a method to measure CI based on the assumption of ANM. Inside, $x - f(Z) \perp Y - g(Z)$ and $X - f(Z) \perp Z$ (or $Y - g(Z) \perp Z$) were used to check CI in non-linear cases, they proved if the two conditions hold, then $X \perp Y|Z$ holds. However, when Z contained more than one variable, $X - f(Z) \perp Z$ was difficult to measured, thus one has to consider the interaction among Z. Moreover, in Non-linear cases, there need to be high time cost to measure the regression functions therefore this method was difficult to be applied to cases with more than 10 variables [26].

In this work, an effective CI test method for causal inference was designed from the point of view of linear non-Gaussian additive noise model (LNANM) [27]. Assuming that the data generation process of a given set of variables follows LNANM, and $X \perp Y|Z$ could be simplified to $x - f(Z) \perp y - g(Z) + \Sigma H(Z)$ or $x - f(Z) + \Sigma H(Z) \perp y - g(Z)$, in which f and g could be obtained by using least squares regression, and $H(*)$, meeting condition $\sum H(Z) = h_i\left(z_j\right) +, \cdots, +h_k(z_l)$ $\left(h_i \in H, z_j \in Z\right)$, was a linear function of Z. We showed that $x \perp y|Z$ can be derived from $x - f(Z) \perp y - g(Z) + \Sigma H(Z)$ or $x - f(Z) + \Sigma H(Z) \perp y - g(Z)$. In practice, f and g can be estimated independently by minimizing the residuals w.r.t. (x, Z) and (y, Z). H can be randomly fixed at a set of linear function.

The proposed conditional independence test method was denoted by Residual Independence Test (RIT). RIT provided a way to simplify conditional independence testing into a simpler set of unconditional independence testing. Finally, the RIT method was applied to causal inference, at the same time, it showed that $x - f(Z) \perp y - g(Z) + \Sigma H(Z)$ (or $x - f(Z) + \Sigma H(Z) \perp y - g(Z)$) can lead to $x \leftarrow Z$ (or $y \leftarrow Z$) . Therefore causal discovery methods, like PC algorithm using RIT to check CI, can detect more causal directions rather than returning a set of Markov equivalent classes. Our experiments showed that on various real-world causal structures, the capability of our method is superior to the state-of-the-art approaches, and our method was high-efficient that can handle high-dimensional cases of more than 400 variables.

This paper is organized as follows: Section II describes Preliminaries include causal network, conditional independence test, D-separated criterion and Markov equivalent. Section III describes the framework of residual independence test (RIT). Section IV proposes the details of causal inference based on RIT. Section V completes performance evaluations by comparing the proposed algorithm with other approaches in the literature. Section VI concludes the paper.

II. PRELIMINARIES

A. CAUSAL NETWORK

Causal network is generally denoted by a directed acyclic graph (DAG), which can represents the probability dependency between variables. Let $X = (x_1, x_2, \ldots, x_n)$ denote the nodes contained in DAG, $E = \left\{e\left(x_i, x_j\right)|x_i, x_j \in X\right\}$ denote the edges between two nodes in DAG, where $e\left(x_i, x_j\right)$ stands for dependencies $x_i \rightarrow x_j$ between x_i and x_j. $P = \left\{P\left(x_i|pa_{x_i}\right)|x_i, pa_{x_i} \in X\right\}$ is a set of conditional probabilities, where $P(x_1, x_2, \ldots, x_n)$ stands for the probabilistic impact of x_i's parent node set on x_i. We can see that a causal network is essentially a graphical representation of all the conditional independence with respect to the joint probability distribution $P = \left\{P\left(x_i|pa_{x_i}\right)|x_i, pa_{x_i} \in X\right\}$.

B. CONDITIONAL INDEPENDENCE TEST

Conditional independence (CI) is an very useful concept in statistics. Let X_i, X_j and Z denote the sets of random variables. CI between X_i and X_j given Z can be denoted by $X_i \perp X_j | Z$, which reflects given the values of Z, further knowing the values of X_i (or X_j) would not provide any additional information about X_j (or X_i). Traditional CI testing methods include G-test, Chi-squared test and kernel-based test. Throughout this work, we use $\perp$ to denote (conditional) independent. In practice, CI testing plays a central role in causal discovery. For one, d-separation criterion is usually used in CI test. The main relations between d-separation and CI are the Markov condition and faithfulness condition. The joint distribution $P(X)$ is said to be Markov with respect to the DAG G in case X_i, X_j d-separated by $Z \implies X_i \perp X_j | Z$, for all disjoint sets X_i,X_j and Z. $P(X)$ is said to be faithful to the DAG G in case X_i, X_j d-separated by $Z \implies X_i \perp X_j | Z$, for all disjoint sets X_i,X_j and Z.

C. D-SEPARATED CRITERION

D-separated criterion is an important graph property to describe the relation among nodes in a causal network. Let X, Y, Z be the set of any three disjoint nodes in the causal undirected graph G, Z d-separate node sets X and Y in graph G, P is blocked If any path from a node of X to a node of Y is blocked by Z, That is, there is a node x_i which on the path P satisfies one of the following conditions:

- x_i has a collision arrow on P, that is $\rightarrow x_i \leftarrow$, x_i and its descendant nodes are not in Z.
- x_i has no collision arrow on P, that is $\rightarrow x_i \rightarrow$ or $\leftarrow x_i \rightarrow$, and $x_i \in Z$.

According to the probability density implication of the desperation criterion, X and Y are independent by given Z in case X and Y are d-separated by Z. Conversely, if X and Y are not d-separated by Z, then X and Y are dependent given Z.

D. MARKOV EQUIVALENT

We said two DAGs are Markov equivalent (or they are Markov equivalent classes) iff they have the same skeleton and V-structures (if a node is a child of two other adjacent nodes, as shown in Fig.1). According to this definition, if a DAG G_1 has no V-structure, another arbitrary DAG G_2 with the same skeleton as G_2 is Markov equivalent to G_1.

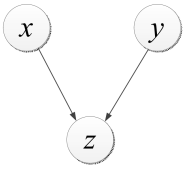


FIGURE 1. ***V*-structure.**

III. THE FRAMEWORK OF RESIDUAL INDEPENDENCE TEST (RIT)

Generally, Linear non-gaussian additive noise model (LNANM) consists of a joint distribution $(S, P(X))$, where $S = \{S_1, S_2, \cdots, S_n\}$ denotes n equations, $S_i : x_i = f_i\left(pa_{x_i}\right) + \varepsilon_i$, $i = 1, 2, \cdots, n$, inside, pa_{x_i} is the direct parents set of x_i in the corresponding DAG G, f_i is a set of linear functions, and the noise variables ε_i have Non-Gaussian distributions and satisfy $\varepsilon_i \perp pa_{x_i}$. LNANM reflects the data-generating process of X in directed acyclic graph G. LNANM is identifiable if it can distinguish asymmetric causal variables [26], [28]. In fact, LNANM is usually recognizable.

In this study, we first study such a situation: Given a DAG, in which the data-generating process of G follows LNANM and two nodes x_i and x_j are randomly selected, we aim to test if x_i and x_j are (conditionally) independent under the given Z, where $Z \cup V \backslash_{x_i x_j}$. If not, all variables in this chapter follow LNANM by default.

In the next part, the foundation of CI testing method was laid by the theoretical results describing CIs (like $x_i \perp x_j | Z$) under the assumption of LNANM. We first presented the Darmois-Skitovitch theorem [29], [30], that is used for deriving the subsequent contents, and as follows:

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

Theorem 1. It is assumed that the data-generating procedure follows linear Non-Gaussian additive noise models. In case of two variables x_i and x_j $(x_i, x_j \in V)$ being neither adjacent nor marginally independent, there must be $x - f(Z) \perp y - g(Z) + \sum H(Z)$ or $x - f(Z) + \sum H(Z) \perp y - g(Z)$ caused by a set of variables Z and two functions f and g .

Proof. Generally, there is an assumption that x_j become the ancestor to x_i, and pa_{x_i} denotes the parents of x_i. Based on the mechanism of ANM, we have $x_i = f\left(pa_{x_i}\right) + \varepsilon_i$ and $\varepsilon_i \perp pa_{x_i}$. Thereby, $x_i - f\left(pa_{x_i}\right) \perp pa_{x_i}$ can be concluded. Since ε_i is an exogenous additive noise independent of x_i and all non-descendant nodes, then we have $\varepsilon_i \perp \left(x_j, pa_{x_i}\right)$. So it is certain that any function g can result in $x_i - f\left(pa_{x_i}\right) \perp x_j - g\left(pa_{x_i}\right)$. According to DST, we can deduce $x_i - f\left(pa_{x_i}\right) \perp x_j - g\left(pa_{x_i}\right) + \sum pa_{x_i}$, and let $\sum pa_{x_i}$ denotes $\sum H(Z)$, then $x - f(Z) \perp y - g(Z) + \sum H(Z)$ is obtained. Similarly, on the condition that x_i is an ancestor (not parent) of x_j, it will be proved that $x - f(Z) + \sum H(Z) \perp y - g(Z)$. Meanwhile, if in the other case that a common ancestor (including parent) of x_i and x_j, then we have $x - f(Z_1) \perp y - g(Z_1) + \sum H(Z_1)$ or $x - f(Z_2) + \sum H(Z_2) \perp y - g(Z_2)$, where Z_1 and Z_2 are denoted as different Z sets. The proof is completed.

In the next section, we will prove that, in the case of conditions $x - f(Z) \perp y - g(Z) + \sum H(Z)$ or $x - f(Z) + \sum H(Z) \perp y - g(Z)$, two variables x_i and x_j are independent to Z, i.e., $x_i \perp x_j | Z$ under given condition set Z.

Theorem 2. It is assumed that the data-generating procedure of the dataset V and an arbitrary linear function $H(*)$ follow linear Non-Gaussian additive noise models. If there exist $x_i, x_j \left(x_i, x_j \in V\right)$, $Z \left(Z \cup V_{\backslash x_i x_j}\right)$ and f and g such that $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$, then $x_i \perp x_j | Z$ holds.

Proof. We first consider the conditional mutual information [31]x_i,x_j of condition set Z, as

$$\begin{aligned} &I\left(x_i, x_j | Z\right) \\ &\quad = I\left(x_i - f(Z); y - g(Z) | Z\right) \\ &\quad = I\left(x_i - f(Z); \left(x_j - g(Z), Z\right)\right) - I\left(x_i - f(Z); Z\right) \\ &\quad = I\left(x_j - g(Z); \left(x_i - f(Z), Z\right)\right) - I\left(x_j - g(Z); Z\right). \end{aligned} \tag{1}$$

In linear Non-Gaussian additive noise case, there exists $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$, as $H(*)$ is an arbitrary linear function, then there must be one $H(*)$ such that $-g(Z) + \sum H(Z)$ cannot eliminate any additive noise variables of Z. The reason is listed as follows:

x_i, x_j and Z are generated by following a linear Non-Gaussian additive noise model, that is $x_i = \sum_{t=1}^{l} a_t s_t$, $y = \sum_{t=1}^{l} b_t s_t$, $z_1 = \sum_{t=1}^{l} c_t s_t, \cdots, z_w = \sum_{t=1}^{l} w_t s_t$ where $s_i \, (i = 1, 2, \cdots, l)$ is the i.i.d. Non-Gaussian additive noises and $Z = \{z_1, \cdots, z_w\}$. Then, for two arbitrary variables z_1 and z_w, there must exist two functions h_1 and h_w such that

$$h_1(z_1) + h_w(z_w) = \sum_{t=1}^{l} d_t s_t. \tag{2}$$

$z_1 = \sum_{t=1}^{l} c_t s_t$ and $z_w = \sum_{t=1}^{l} w_t s_t$ meet the condition, if c_l or w_l is not 0, then d_l is not 0. From the perspective of the function model of causal graph, if two arbitrary functions h_1 and h_w such that $h_1(z_1)$ plus $h_w(z_w)$ are able to eliminate the common additive noise of z_1 and z_w, then the model is not faithfulness, because it is bound that any of their common child must lose at least one additive noise term about its ancestors. This means that $h_1(z_1)$ plus $h_w(z_w)$ will not eliminate any additive noise term of z_1 and z_w . Such a result can be extended to: $\sum H(Z), -g(Z) + \sum H(Z), \cdots, y - g(Z) + \sum H(Z)$. Therefore we have $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp \sum H(Z)$ according to DST. Similarly, given $x_i - f(Z) \perp \sum H(Z)$, since $H(*)$ is an arbitrary linear function, we can deduce that $x_i - f(Z) \perp Z$ according to DST. There are two conditions $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$, such that

$$\begin{cases} I\left(x_i - f(Z); \left(x_j - g(Z), Z\right)\right) = 0 \\ I\left(x_i - f(Z); Z\right) = 0 \end{cases} \tag{3}$$

thus $I\left(x_i; x_j, Z\right) = 0$ is obtained, i.e., $x_i \perp x_j | Z$. On the other side, we can also obtain $I\left(x_j - g(Z); \left(x_i - f(Z), Z\right)\right) - I\left(x_j - g(Z); Z\right) = 0$ on the similar conditions. This completes the proof.

Theorem 2 means that $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$ are sufficient to support $x_i \perp x_j | Z$. It can be found out from the combination of Theorem 1 and Theorem 2 that the CI test of $x_i \perp x_j | Z$ can be replaced by two unconditionally independent tests $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$. Hence, we can simplify the CI test into a set of marginal independent tests according to the above two theorems. When the method is applied to causal discovery, in the worst case, we need at most $2 * k * \sum_{i=1}^{|S|} C_{|S|}^{i}$ (S denotes the maximum conditional set, $Z \in S$, k is the times of choosing $H(*)$) unconditionally independent tests to determine whether x_i and x_j are conditionally independent. The existing CI test methods, by contrast, need $\sum_{i=1}^{|S|} C_{|S|}^{i}$ times in CI test.

From the perspective of LNANM, it is assumed that we randomly choose two variables, if they are not adjacent; it is easy to find the two linear functions f and g and choose k times to meet the condition $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$ according to Theorem 1. On the other side, $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ (or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$) leads to $x_i \perp x_j | Z$ according to Theorem 2. So there is only need to check whether these three variables (functions f, g and k times $H(*)$) can be found to achieve $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$.

In causal discovery, we often use V-structure learning and consistent propagation [4] to learn causal directions. Recall that $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$ (or $x_i - f(Z) + \sum H(Z) \perp x_j - g(Z)$) $\Rightarrow x_i - f(Z) \perp Z$ (or $x_j - g(Z) \perp Z$) according to Theorem 2. Compared with a series of Markov equivalence classes alone, RIT can capture more information about the causal direction, not just the deterministic V-structure. This is because in LNANM, if $x - f(Z) \perp Z$, then Z cannot contain a child of x. Compared with concretizing-based and kernel-based tests, RIT can detect more causal directions even without V-structure. Here is a simple example, given a causal structure of $x_1 \leftarrow x_2 \rightarrow x_3$, it is easy to find two linear functions f and g such that $x_i - f(Z) \perp x_j - g(Z) + \sum H(Z)$, eventually $x_1 \leftarrow x_2$ and $x_2 \rightarrow x_3$ can be inferred. However, it is difficult for concretizing-based and kernel-based tests 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 they have the same conditional and unconditional independence.

IV. CAUSAL INFERENCE BASED ON RIT

In this section, we will introduce a new causal inference method based on the combination of PC algorithm and RIT, this method is denoted by PC_{RIT} for simplicity. Concretely, PC_{RIT} is based on the standard PC algorithm [3], where RIT is used for testing CI, and any existing methods like KCIT can be used for testing marginal independence. RIT is performed simply by estimating $\tilde{f}$ of f and $\tilde{g}$ of g by using least square regression. Therefore it is easy to test whether $x - \tilde{f}(Z) \perp y - \tilde{g}(Z) + \sum H(Z)$ or $x - \tilde{f}(Z) + \sum H(Z) \perp y - \tilde{g}(Z)$ holds with k times $H(*)$. The corresponding pseudo-code of our proposed method is listed in Algorithm 1. The first step (Lines 1-7) aims to reconstruct the causal skeleton based on RIT. The procedure is the same to that of PC algorithm. Concretely, we first construct a fully connected undirected graph G over the given variables set X, then we check whether every two variables x_i and x_j can be conditional independent given a variable set Z. We delete the edge $x_i - x_j$ from G if the CI holds. After obtained the causal skeleton, the edges can be oriented according to $x_i - \tilde{f}(Z) \perp \sum H(Z)$ (or $x_i - \tilde{f}(Z) \perp Z$) and $x_j - \tilde{g}(Z) \perp \sum H(Z)$ (or $x_j - \tilde{g}(Z) \perp Z$) (Lines 8-12). Finally, the remaining undirected edges will be checked and direction-deduced by V-structures learning and consistent propagation, this process is the same as that in the PC algorithm (Line 17). For example, to check whether $x_i - x_j - x_k$ forms V-structure. If it does, then $x_i \rightarrow x_j \leftarrow x_k$ will be accepted.

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

```
1:  Input: variables set $X = \{x_1, \ldots, x_n\}, k$.
2:  Output: partial DAG $G$.
3:  Form the complete undirected graph $G$ on the variables set $X$.
4:  for $\forall x_i, x_j \in X$ and adjacent in $G$ do
5:      for $\forall Z \in X \setminus \{x_i, x_j\}$ do
6:          do linear regression to measure $\tilde{f}$ of $f$ and $\tilde{g}$ of $g$
7:          randomly generate $k$ times $H(*)$ as $H_1, \ldots, H_k$.
8:          if $x_i - \tilde{f}(Z) \perp x_j - \tilde{g}(Z) + \sum H_t(Z)$ or $x_i - \tilde{f}(Z) + \sum H_t(Z) \perp x_j - \tilde{g}(Z)$ for $\forall H_i \in H$ holds then
9:              delete edge $x_i - x_j$ from $G$
10:             if $x_i - \tilde{f}(Z) \perp \sum H(Z)$ (or $x_i - \tilde{f}(Z) \perp Z$) then
11:                 orient $Z$ to $x_i$.
12:             end if
13:             if $x_j - \tilde{g}(Z) \perp \sum H(Z)$ (or $x_j - \tilde{g}(Z) \perp Z$) then
14:                 orient $Z$ to $x_j$.
15:             end if
16:             break
17:         end if
18:     end for
19: end for
20: orient the remaining un-oriented edges based on $V$-structure and do consistent propagation.
```

V. PERFORMANCE EVALUATIONS

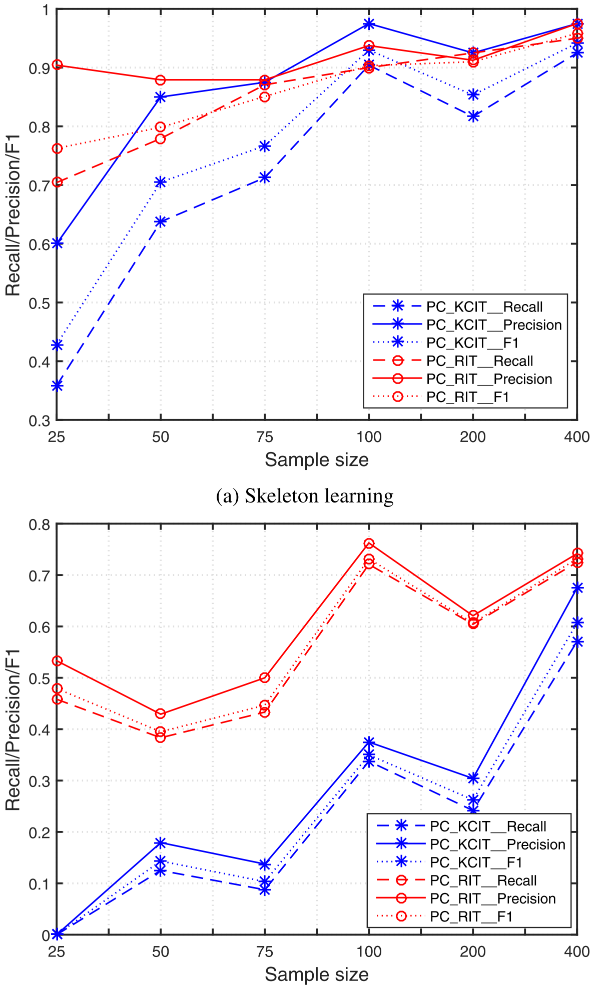
In this section, we first conduct experiments to evaluate RIT, and make comparison of the RIT and KCIT [16] integrated with the PC algorithm framework [3], i.e., PC_{RIT} vs. PC_{KCIT}. There are many results on the comparisons between KCIT and the other CI testing methods in the previous works [16], [17], [25], [32]. In this implementation of RIT, the least square regression method is used to measure the linear functions f and g. In order to obtain the best performance or result of KCIT, the inner bootstrap step and Gaussian process are active. In all the following procedure of RIT, the parameter k is fixed at 10, which means we randomly choose 10 times $H(*)$ to measure CIs. In practice, because the coefficients in generating simulated model are randomly chosen, k can be as small as 10 according to the proof in Theorem 2.

A. PERFORMANCE IN SIMULATED MODEL

In this group of experiments, we evaluate the proposed method by the simulated datasets generated by following a set of simulated causal network structures under a linear Non-Gaussian additive noise model. In fact, it is hard to find a large group of datasets with respect to causal inference problems with ground truth. At present, simulated data on given structures are popularly used in many causality inference methods [33]. Here we make an assumption that the actual causal network structure for n random variables $x_1, \cdots, x_n$ can be denoted as a graph (DAG) G. An additional hypothesis of faithfulness is widely used in the constraint-based methods (such as PC algorithm); faithfulness means that the joint distribution cannot access by any CI information not contained in Markov conditions. We thus can recover the graph structure by checking the CIs and independence in the data. Evidently, it is only possible for one to restore a set of Markov equivalence classes. Therefore, if PC algorithm use the existing CI test methods to detect causal directions like KCIT based on V-structures learning and CI test method for consistent propagations [4], then PC algorithm can find only a set of Markov equivalence classes. In the following experiments, we will show that PC_{RIT}, proposed in the paper, can deduce much more causal directions than those of PC_{KCIT}.

The simulated dataset is generated by following a random DAG G. Specifically, we random chooses four variables $x_1, \cdots, x_4$ and makes the arrows among them by following x_i to x_j only for $i < j$. And the arrow is either present or absent with probability 0.5. We generate the root variables by following $U(0, 1)$ and the leaf variables x_i came from $\sum_i a_i * pa_{x_i} + \varepsilon$ where $a_i \sim U(0.2, 1)$ and $\varepsilon \sim U(-0.2, 0.2)$ are independent from pa_{x_i}. In addition, 1000 sample data are simulated and generated in a sample data set whose size is controlled in 25, 50, 75, 100, 200, 400 and the ability of PC_{RIT} and PC_{KCIT} to infer causal skeleton and PDAGs (including identifiable causal direction) respectively is evaluated. the ability of PC_{RIT} to infer causal skeleton is evaluated, and which of PC_{KCIT} (including identifiable causal direction) is also evaluated.

We can see that from Fig.2(a), when the sample is less (e.g. less than 100), PC_{RIT} performs much better than PC_{KCIT} on causal skeleton learning. On the one hand we increase the sample, the performance of PC_{KCIT} is close to that of PC_{RIT}, on the other hand the sample size is up to 400, the PC_{RIT} and PC_{KCIT} tend to overlap in term of the F1 curves. In practice, we have collected many other simulation datasets generated



(a) Skeleton learning

(b) PDAG learning

FIGURE 2. The performance with respect to PC_{RIT} and PC_{KCIT} in term of (a) causal skeleton learning and (b) PDAG learning.

by following a similar procedure to test PC_{RIT} by inputting different parameters, and we finally obtained the results similar to those shown in Fig.2(a).

On the other side, the two methods are also evaluated in term of PDAG learning. We presented the corresponding results in Fig.2(b). It can be seen that PC_{RIT} can obtain better results in these cases. With enough samples, PC_{KCIT}'s performance in causal skeleton learning is similar to that of PC_{RIT}. The reason is that PC_{KCIT} detect the causal directions only by V-structure and the corresponding consistent propagation [4] that is PC_{KCIT} outputs only Markov equivalence classes, while PC_{RIT} can learn more information over the causal directions.

B. PERFORMANCE ON REAL-WORLD STRUCTURES

In the above experiments, we compared our method to PC_{KCIT} in causal skeleton learning and causal direction learning, and the experimental results have shown that RIT is able to break Markov equivalence classes. Therefore PC_{RIT} can recover more information about the causal directions. In this subsection, further comparisons were made between PC_{RIT} and three other causal inference methods, including LiNGAM [20], DLiNGAM [27] and Spase-ICA LiNGAM [34]. Since all these methods can distinguish Markov equivalence classes, we can further evaluate our causal inference methods, including causal direction learning. The implementations of LiNGAM and DLiNGAM strictly follow the original papers [20], [27]. Sparse-ICA and LiNGAM use [34] Spase-ICA algorithm and [20] pruning algorithm. In the next step, we will collect eight real-world causal network structures to evaluate PC_{RIT}, LiNGAM, DLiNGAM and Sparse-ICA LiNGAM. These causal network structures cover a variety of applications in causality discovery, including medicine system (Alarm and Pathfinder), insurance system evaluation (Insurance), agricultural industry (Barley), weather forecasting (Hailfinder), the pedigree of breeding pigs (Pigs dataset) and system troubleshooting (Win95pts and Andes). The structural statistics of these causal network structures are summarized in Table 1. It is noted that the three baselines (Nodes, Avg.degree, Max degree) are highly impacted by the the sample size and the number of nodes, and LiNGAM cannot work if the samples size is smaller than $|V|$. Consequently, in what follows, we fix the sample size at $2|V|$ to compare PC_{RIT} with the other three existing methods, LiNGAM, DLiNGAM and Sparse-ICA LiNGAM.

TABLE 1. The statistics of causal network structures

Dataset	Nodes#	Avg. degree	Max degree
$Cancer$	5	1	3
$Aisa$	8	1	4
$Insurance$	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

The experiment results are presented in Table 2, in which the three methods DLiNGAM, LiNGAM and Sparse-ICA LiNGAM are respectively denoted by LiGM, DLiGM and SICA due to space limit. We can see that PC_{RIT} almost achieves the top performance in all datasets and only LiNGAM outperforms PC_{RIT} in term of Recall in the case of Insurance. One of the reasons is that Insurance has only 27 nodes (see Table 1) that is the simplest one among the eight structures. In the other seven cases, PC_{RIT} outperforms LiNGAM, DLiNGAM and Sparse-ICA LiNGAM, especially in larger causal networks (with $|V| > 100$). The recall rates of the three methods (LiNGAM, DLiNGAM and Sparse-ICA LiNGAM) can distinguish the actual graph from the corresponding Markova equivalence classes, but they are not reliable in some cases. The sample size needs to be enlarged if in need of improvement for their learning accuracy. Due to high

TABLE 2. Results on four causal structure learning methods.

Dataset	*Recall*				*Precision*				*F1* Score			
	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA
Insurance	0.32	**0.54**	0.42	0.31	**0.65**	0.13	0.09	0.09	**0.41**	0.20	0.15	0.14
Alarm	**0.59**	0.24	0.22	0.20	**0.86**	0.30	0.20	0.33	**0.67**	0.27	0.20	0.25
Barley	**0.40**	0.37	0.24	0.37	**0.72**	0.24	0.16	0.23	**0.62**	0.24	0.15	0.32
Hailfinder	**0.60**	0.27	0.17	0.33	**0.70**	0.22	0.14	0.31	**0.62**	0.24	0.15	0.32
Win95pts	**0.45**	0.30	0.23	0.34	**0.49**	0.33	0.24	0.27	**0.42**	0.31	0.23	0.30
Pathfinder	**0.92**	0.34	0.34	0.33	**0.72**	0.23	0.23	0.20	**0.78**	0.28	0.27	0.25
Andes	**0.76**	0.21	0.12	0.26	**0.76**	0.46	0.26	0.51	**0.76**	0.28	0.16	0.34
Pigs	**0.54**	0.58	N.A.	N.A.	**0.58**	0.16	N.A.	N.A.	**0.55**	0.15	N.A.	N.A.

TABLE 3. Results on four causal structure learning methods with 200 samples.

Dataset	*Recall*				*Precision*			
	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA
Cancer	0.5	**1**	**1**	**1**	0.5333	**1**	**1**	**1**
Aisa	0.8125	0.875	0.875	**1**	0.8125	0.875	0.875	**1**
Insurance	0.6498	0.5926	0.5	**0.9167**	**0.62**	0.3077	0.2308	0.4231
Alarm	0.5856	0.5556	0.3023	**0.6667**	**0.727**	0.3261	0.2826	0.4783
Barley	0.4878	0.3846	0.2712	**0.6444**	**0.5974**	0.2381	0.1905	0.3452
Hailfinder	**0.5202**	0.3333	0.2603	0.3875	**0.558**	0.3182	0.2879	0.4697
Win95pts	**0.5976**	0.2759	0.2099	0.3158	**0.5878**	0.3429	0.2429	0.3429
Pathfinder	**0.7688**	0.3309	0.3252	0.2672	**0.8439**	0.2359	0.2051	0.1795
Andes	**0.7425**	0.1549	0.1287	0.1837	**0.5516**	0.2929	0.2456	0.2663
Dataset	*F1* Score				*Time(s)*			
	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA
Cancer	0.5143	**1**	**1**	**1**	0.2099	0.0476	0.0724	0.0737
Aisa	0.8125	0.875	0.875	**1**	1.4174	1.2696	0.4343	0.4061
Insurance	**0.6207**	0.4051	0.3158	0.5789	22.3079	0.5299	3.6307	7.4778
Alarm	**0.6267**	0.411	0.2921	0.557	34.9715	0.7345	16.0263	17.9295
Barley	**0.5059**	0.2941	0.2238	0.4496	**0.5974**	0.2381	0.1905	0.3452
Hailfinder	**0.5202**	0.3333	0.2603	0.3875	101.6749	0.9934	33.4152	41.9886
Win95pts	**0.5705**	0.3057	0.2252	0.3288	267	1.8969	153.0445	162.7144
Pathfinder	**0.792**	0.2754	0.2516	0.2147	951	4.2543	274.7252	479.5264
Andes	**0.6066**	0.2027	0.1689	0.2174	18600	22.6499	3900	4340

time-complexity of DLiNGAM and Sparse-ICA LiNGAM, in the case of Pigs network, their results are not presented.

It is clearly that the PC_{RIT} curves (Recall, Precision and F1 score) increased with the sample size instead of the ratio of the sample size to the number of nodes ($2|V|$), and the other three methods DLiNGAM, LiNGAM and Sparse-ICA LiNGAM are relatively stable. Simultaneously, it can be seen that in the case of larger network, like *Pathfinder*, *Andes* and *Pigs*, the F1 value of PC_{RIT} is 2 to 3 times higher than those of the other three methods. The reasons can be concluded in two perspectives,: 1) RIT is used to detect the CIs in PC_{RIT}, while causal skeleton inference of DLiNGAM, LiNGAM and Spase-ICA LiNGAM was based on functional modeling only. Given reliable CI tests methods, CI tests in skeleton learning has higher robustness than functional modeling, hence the causal skeleton recovered by PC_{RIT} is theoretically more accurate. 2) Breaking Markov equivalence classes by RIT means that our method can capture more information about causal directions than other three methods.

As a conclusion, there are two main reasons for why PC_{RIT} can work better than the state-of-the-art methods DLiNGAM, LiNGAM and Sparse-ICA LiNGAM, especially in large-scale cases, 1) RIT relaxes the CI test to two simple marginal independence test which can achieve a better performance on detecting CI. 2) PC_{RIT} can infer the corresponding causal directions during CI testing by using RIT, which enable PC_{RIT} to distinguish Markov equivalence classes.

In order to test how the sample sizes affect the performance of these methods, we conduct another group of experiment that the sample size is fixed at 200 with different dimensionalities of networks. The results are presented in Table 3.

We can see that the performance of PC_{RIT} under two lower-dimensional datasets Asia and Cancer with 6 and 8 nodes respectively, are not as good as other these algorithms. Because the performance of our method is heavily impacted by the sample size rather than the rate of Sample size/The number of nodes. So given only 200 sample sizes, it is not enough to show the best performance of our method. But one thing that should be noted, our method becomes much more competitive with the growing dimensionalities, while the other methods require more and more sample. If we fixed the sample size, the accuracy of those methods tends to be unreliable. On the other side, as we knows there are a lot of methods can deal with low-dimensional cases, thus out method mainly aims to deal with high-dimensional cases with limit samples.

But the three data sets Insurance, Alarm and Barley in the table above with 27, 37 and 48 nodes, respectively, present the

TABLE 4. Results on four causal structure learning methods with Win95 data set.

Sample	*Recall*				*Precision*			
	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA
300	**0.6066**	0.2857	0.2674	0.5	**0.5455**	0.3714	0.3286	0.7
250	**0.5743**	0.2697	0.4684	0.1667	**0.527**	0.3429	0.5286	0.2286
200	**0.5976**	0.2759	0.2099	0.3158	**0.5878**	0.3429	0.2429	0.3429
150	**0.5094**	0.3229	0.2093	0.4259	**0.5192**	0.4429	0.2571	0.3286
100	**0.4622**	0.2464	0.1977	0.1846	**0.4914**	0.2429	0.2429	0.1714
Sample	*F1* Score				*Time(s)*			
	PC_{RIT}	LiGM	DLiGM	SICA	PC_{RIT}	LiGM	DLiGM	SICA
300	0.5556	0.323	0.2949	**0.5833**	373.8302	2.0668	180.488	180.0242
250	**0.5345**	0.3019	0.4966	0.1928	294.4153305	1.9871	166.1809	175.9603
200	**0.5705**	0.3057	0.2252	0.3288	266.6057	1.8969	153.0445	162.7144
150	**0.4875**	0.3735	0.2308	0.371	171.2773859	1.4052	131.1348	130.766
100	**0.454**	0.2446	0.2179	0.1778	120.4015211	1.0318	112.0104	109.7414

result that they are better than the other three algorithms in the evaluation criteria of Recall, Precision and F1. In these three cases, SICA worked better than the other three algorithms in term of Recall criterion, while the proposed algorithm is better than the other three algorithms in term of Precision and F1 criterion.

When the dimensionality of the structures becomes higher, such as in the cases of Hailfinder, Win95pts, Pathfinder and Andes with 56, 76, 109, 223 nodes, our method can get significantly better Recall, Precision and F1 that the other methods. In these cases, the rates of Sample size/The number of nodes became low enough, the other three methods are not able to get a competitive performance. It can be seen that the performance of PC_{RIT} is much stable than those of the counterparts when the sample size is fixed. Because the counterparts are heavily impacted by the rate of Sample size/The number of node. As the samples are generally limited, these methods are not easily to handle high-dimensional cases, and then we can choose PC_{RIT} to achieve the goal.

We further select one of the datasets to do a experiments with different sample sizes, and the experimental results are shown in table 4. We can see that the F1 score of the proposed algorithm is 0.45 with 100 samples, and is up to 0.55 when samples reaching 300, the gap is (0.55-0.45)/0.55=0.18. We can see that our method is much easier to obtain a better score when the sample size is very small under a higher-dimensional network, while other methods, especially SICA, one has to dramatically increase the sample size to get a better score.

VI. CONCLUSION

In this work, a novel residual-based conditional independence characterization testing method based on linear non-Gaussian additive noise model was proposed to solve the problem of distinguishing Markov equivalent classes. We showed that the CIs can be tested by some weaker conditions if the causal process is known as linear Non-Gaussian. Concretely, test of $x \perp y | Z$ can be reduced to a set of unconditional independence tests of $x - f(Z) \perp y - g(Z) + \sum H(Z)$ (or $x - f(Z) + \sum H(Z) \perp y - g(Z)$) under the assumption that the data-generating process follows linear Non-Gaussian additive noise model. We further use $x - f(Z) \perp y - g(Z) + \sum H(Z)$ (or $x - f(Z) + \sum H(Z) \perp y - g(Z)$) to infer causal directions. In contrast to the state-of-the-art kernel-based method KCIT, the proposal is less sensitive to the dimensionality of Z in causal skeleton and direction learning. Experiments on both simulated and real-world causal network structures verify that the new method outperforms KCIT in linear non-Gaussian cases.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest the authors declare they have no conflict of interest.

REFERENCES

[1] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, "Kernel-based tests for joint independence," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 80, no. 1, pp. 5–31, Jan. 2018.

[2] Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic, "Large-scale kernel methods for independence testing," *Stat Comput.*, vol. 28, no. 1, pp. 113–130, Jan. 2018.

[3] B. P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. New York, NY, USA: Springer, 1993.

[4] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[5] W. P. Bergsma, *Testing Conditional Independence for Continuous Random Variables*. Eindhoven, The Netherlands: Eurandom, 2004.

[6] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measure of conditional independence," *Austral. New Zealand J. Statist.*, vol. 46, no. 4, pp. 657–664, 2015.

[7] S. Richardson and W. R. Gilks, "Conditional independence models for epidemiological studies with covariate measurement error," *Statist. Med.*, vol. 12, no. 18, pp. 1703–1722, Sep. 1993.

[8] S. Richardson and W. R. Gilks, "A Bayesian approach to measurement error problems in epidemiology using conditional independence models," *Amer. J. Epidemiol.*, vol. 138, no. 6, pp. 430–442, Sep. 1993.

[9] N. S. Revankar and M. J. Hartley, "An independence test and conditional unbiased predictions in the context of simultaneous equation systems," *Int. Econ. Rev.*, vol. 14, no. 3, p. 625, Oct. 1973.

[10] L. Su and H. White, "A nonparametric Hellinger metric test for conditional independence," *Econom. Theory*, vol. 24, no. 4, pp. 829–864, Aug. 2008.

[11] I. Steinwart and C. Scovel, "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs," *Constructive Approximation*, vol. 35, no. 3, pp. 363–417, 2012.

[12] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola, "A kernel method for the two-sample problem," 2008, *arXiv:0805.2368*. [Online]. Available: https://arxiv.org/abs/0805.2368

[13] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, vol. 20, no. 1, pp. 167–204.

[14] H. Araki and S. Yamagami, "An inequality for Hilbert–Schmidt norm," *Commun. Math. Phys.*, vol. 81, no. 1, pp. 89–96, Sep. 1981.

[15] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, "Universality, characteristic kernels and RKHS embedding of measures," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2389–2410, 2010.

[16] K. Zhang, J. Peters, D. Janzing, and B. Schoelkopf, "Kernel-based conditional independence test and application in causal discovery," *Comput. Sci.*, vol. 06, no. 08, pp. 895–907, 2012.

[17] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf, "A permutation-based kernel conditional independence test," in *Proc. UAI*, 2014, pp. 132–141.

[18] E. V. Strobl, K. Zhang, and S. Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," *J. Causal Inference*, vol. 7, no. 1, 2019.

[19] H. Zhang, S. Zhou, and J. Guan, "Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery," in *Proc. AAAI*, 2018, pp. 2029–2036.

[20] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, Dec. 2006.

[21] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Scholkopf, "Nonlinear causal discovery with additive noise models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2008.

[22] J. Peters, D. Janzing, and B. Scholkopf, "Causal inference on discrete data using additive noise models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2436–2450, Dec. 2011.

[23] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf, "Consistency of causal inference under the additive noise model," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 478–486.

[24] R. P. Bagozzi and Y. Yi, "On the evaluation of structural equation models," *J. Acad. Marketing Sci.*, vol. 16, no. 1, pp. 74–94, 1988.

[25] H. Zhang, S. Zhou, K. Zhang, and J. Guan, "Causal discovery using regression-based conditional independence tests," in *Proc. AAAI*, 2017, pp. 1250–1256.

[26] K. Zhang and A. Hyvarinen, "On the identifiability of the post-nonlinear causal model," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009.

[27] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-Gaussian structural equation model," *J. Mach. Learn. Res.*, vol. 12, no. 2, pp. 1225–1248, 2011.

[28] K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf, "On estimation of functional causal models: General results and application to the post-nonlinear causal model," *TISTACM Trans. Intell. Syst. Technol.*, vol. 7, no. 2, pp. 1–22, Dec. 2015.

[29] G. Darmois, "Analyse générale des liaisons stochastiques: Etude particulière de l'analyse factorielle linéaire," *Rev. Int. Stat. Inst.*, 1953.

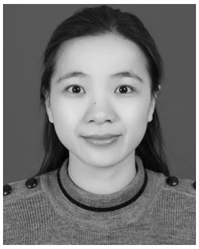
[30] V. Skitovich, "On a property of the normal distribution," *DAN SSSR*, vol. 89, pp. 217–219, 1953.

[31] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.

[32] E. V. Strobl, K. Zhang, and S. Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," 2017, *arXiv:1702.03877*. [Online]. Available: https://arxiv.org/abs/1702.03877

[33] M. Kalisch and P. Buehlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *J. Mach. Learn. Res.*, vol. 8, no. 2, pp. 613–636, 2012.

[34] K. Zhang and L. W. Chan, "ICA with sparse connections," in *Intelligent Data Engineering and Automated Learning—IDEAL*. Berlin, Germany: Springer, 2006.



GUIZHEN MAI received the B.S. degree in information and computing science from Hanshan Normal University, Chaozou, China, in 2007, and the M.Sc. and Ph.D. degrees from the Guangdong University of Technology, Guangzhou, China, in 2015 and 2018, respectively. She is currently a Postdoctoral Researcher with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou. Her research interests cover a variety of different topics, including causality, machine learning, cloud computing, and information systems.

YINGHAN HONG received the B.S. degree in information and computing science from Hanshan Normal University, Chaozou, China, in 2007, and the M.Sc. and Ph.D. degrees in computer science from the Guangdong University of Technology, Guangzhou, China, in 2010 and 2018, respectively. He is currently an Associate Professor and a Computer System Analysts with the School of Physics and Electronic Engineering, Hanshan Normal University. His research interests cover a variety of different topics including causality, machine learning, cloud computing, and data mining and their applications.

PINGHUA CHEN received the B.S. degree in industrial automation from Xiangtan University, Xiangtan, China, in 1989, and the M.Sc. degree in industrial automation from the South China University of Technology, Guangzhou, China, in 1992. He is currently a Professor and the Vice President with the School of Computer Science and Technology, Guangdong University of Technology. His research interests cover a variety of different topics including cloud computing, and data mining and their applications.

KEXI CHEN received the B.S. degree from Dalian Ocean University, China, in 2017. He is currently pursuing the M.S. degree with the College of Automation, Guangdong University of Technology. His interests include impulsive control and asynchronous consensus of multiagent systems.

HAN HUANG received the B.Man. degree in information management and information system from School of Mathematics, South China University of Technology, Guangzhou, China, in 2003, and the Ph.D. degree in computer science from the South China University of Technology, in 2008. He is currently a Professor with the School of Software Engineering, SCUT. His research interests include theoretical foundation and application of evolutionary computation and stochastic heuristics. He is also a Senior Member of CCF.

GENGZHONG ZHENG received the Ph.D. degree from the School of Computer Science and Technology, Xidian University, in 2012. He is currently working as a Professor with the School of Computer and Information Engineering, Hanshan Normal University. His current research interests include network optimization and wireless sensor networks.