Causal Discovery Using Regression-Based Conditional Independence Tests

Hao Zhang,[†] **Shuigeng Zhou,**[†*] **Kun Zhang,**[‡] **Jihong Guan**[§]
[†]Shanghai Key Lab of Intelligent Information Processing, Fudan University, China.
[‡]Department of Philosophy, Carnegie Mellon University, USA.
[§]Department of Computer Science & Technology, Tongji University, China
[†]{haoz15, sgzhou}@fudan.edu.cn; [‡]kunz1@cmu.edu; [§]jhguan@tongji.edu.cn

Abstract

Conditional independence (CI) testing is an important tool in causal discovery. Generally, by using CI tests, a set of Markov equivalence classes w.r.t. the observed data can be estimated by checking whether each pair of variables x and y is d-separated, given a set of variables Z. Due to the curse of dimensionality, CI testing is often difficult to return a reliable result for high-dimensional Z. In this paper, we propose a regression-based CI test to relax the test of $x \perp y|Z$ to simpler unconditional independence tests of $x - f(Z) \perp y - g(Z)$, and $x - f(Z) \perp Z$ or $y - g(Z) \perp Z$ under the assumption that the data-generating procedure follows additive noise models (ANMs). When the ANM is identifiable, we prove that $x - f(Z) \perp y - g(Z) \Rightarrow x \perp y|Z$. We also show that 1) f and g can be easily estimated by regression, 2) our test is more powerful than the state-of-the-art kernel CI tests, and 3) existing causal learning algorithms can infer much more causal directions by using the proposed method.

Introduction

Statistical independence and conditional independence (CI) are important concepts in statistics, artificial intelligence (AI) and other related fields. In causal discovery, we consider such a scenario: let X, Y and Z denote sets of random variables, if the CI between X and Y given Z holds, denoted by $X \perp Y|Z$, then it means that given Z, further knowing X (or Y) does not provide any additional information about Y (or X), thus we can deduce that X and Y have no directed causality. Independence and CI play a central role in causal discovery. Generally speaking, the CI relationship $X \perp Y|Z$ allows us to separate $X - Y$ when constructing a probabilistic model for $P(X, Y, Z)$, which results in a parsimonious representation. By using CI tests, the PC algorithm (Spirtes, Glymour, and Scheines 2000), for example, can determine a set of Markov equivalence classes (Pearl 2009).

However, CI testing is much more difficult than unconditional independence testing (Bergsma 2004). For CI tests, traditional methods either focus on the discrete cases (the conditional set can be combined into a variable according to the corresponding conditional probability table, hence it is easier to handle discrete cases), or impose simplified assumptions to deal with the continuous cases. For example, under the assumption of Gaussian variables with linear dependence relationships, partial correlation can be used to test CI (Baba, Shibata, and Sibuya 2004). In such a situation, $X \perp Y|Z$ reduces to zero partial correlation or zero correlation between X and Y given Z, which can be easily tested. Nevertheless, nonlinearity and non-Gaussian noises are popular in practice, hence this assumption is not always reasonable, and often leads to incorrect results.

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 unconditional independence in each bin. Due to the curse of dimensionality, the conditional set becomes very large, inevitably the required sample size increases dramatically. For example, in (Su and White 2008) the authors used a characterization of CI, $P_{X|YZ} = P_{X|Z}$, to determine CI by measuring the distance between estimates of conditional densities. However, accurate estimation of conditional densities or related quantities is not easy, which deteriorates the testing result, especially when the conditional set is too large.

Generally speaking, CI testing is a nontrivial task, and the "curse of dimensionality" of the conditional variable Z makes it even more challenging. When Z takes a finite number of values $\{z_1, ..., z_k\}$, then $X \perp Y|Z$ iff $X \perp Y|Z = z_i$ for each value z_i. Given a sample of size n, even if the data points are distributed evenly on the values of Z, we must show the independence within each subset of the sample with the same Z value by using only approximately n/k points in each subset. When Z is real-valued and P_z is continuous, 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.

Recently, kernel-based tests were proposed for conditional and unconditional independence testing. With the ability to represent high order moments, mapping of vari-

[*]Correspondence author.

ables into reproducing kernel Hilbert spaces (RKHSs) allows us to infer properties of distributions, such as independence and homogeneity (Gretton et al. 2006). In (Fukumizu et al. 2007), the authors proposed 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 iff $X \perp Y|Z$. We denote this method by CI_{PERM}. A more recent method (denoted by KCIT in short) proposed in (Zhang et al. 2011), uses partial association of regression functions to measure CI, $X \perp Y|Z$ iff for all $f \in L^2_{XZ}$ and $g \in L^2_Y$ (L^2_{XZ} and L^2_Y denote the spaces of square integrable functions of (X, Z) and Y, respectively) such that $E(\tilde{f}\tilde{g}) = 0$ where $\tilde{f}(X, Z) = f(X, Z) - r_f(Z)$ and $\tilde{g}(Y, Z) = g(Y) - r_g(Z)$ ($r_f, r_g \in L^2_Z$ are regression functions). This method relaxes the spaces of functions f, g, r_f and r_g to RKHSs, corresponding to kernels defined on these variables. Compared to discretization-based CI testing methods, kernel methods exploit more complete information of the data and involve less random error. It was showed that causal learning methods based on kernel methods can discover more accurate causalities.

In causal discovery, the mechanism of data generation is often assumed. A widely used model is the additive noise model (ANM) (Shimizu et al. 2006; Hoyer et al. 2009; Peters, Janzing, and Schölkopf 2011), because many real-world data are regarded to be generated by following ANM (Peters, Janzing, and Schölkopf 2011). Concretely, ANM assumes that the observed variables follow a directed acyclic graph (DAG) with the structure function: $Y = f(X) + \varepsilon$ where X is the parent of Y and ε is a random noise term that $X \perp \varepsilon$. So CI tests in ANM not only use the three sets of variables X, Y and Z, but also consider random noise that may be small but really exists.

In this paper, we try to develop a new CI testing method for causality discovery from the perspective of ANM. Consider a set of variables Z and other two variables x and y, we show that if the data-generating procedure follows ANM, then we can relax $X \perp Y|Z$ to two conditions $x - f(Z) \perp y - g(Z)$, and $x - f(Z) \perp Z$ or $y - g(Z) \perp Z$, where functions f and g are estimated by regression. When the ANM is identifiable (Zhang and Hyvärinen 2009; Peters, Janzing, and Schölkopf 2011), we further prove that $x - f(Z) \perp y - g(Z)$ implies $x - f(Z) \perp Z$ or $y - g(Z) \perp Z$, which means $x - f(Z) \perp y - g(Z)$ is sufficient to support $X \perp Y|Z$. We also show that f and g can be easily calculated independently by minimizing the residuals of (x, Z) and (y, Z). With this result, we propose the regression-based conditional independence test method, which is denoted by RCIT. RCIT provides a way to relax CI tests to simpler unconditional independence tests. Finally, we apply RCIT to causality discovery.

It is well known that existing causal discovery methods based on CI tests usually return a set of Markov equivalence classes. In our RCIT, $x - f(Z) \perp y - g(Z)$ and $x - f(Z) \perp Z$ or $y - g(Z) \perp Z$ implies $Z \rightarrow x$ or y. This means that causal discovery methods (e.g. the PC algorithm) using RCIT for CI testing can detect more causal directions (details are in the next section). In our experiments, we show that on synthetic datasets the proposed method is more powerful than state-of-the-art approaches, and it can accurately estimate the distribution of the test statistic under the null hypothesis when the dimensionality of Z grows to produce a well-calibrated test. We also validate the practicability of the new test for inferring CI relationships on real-world datasets.

Regression-based conditional independence test (RCIT)

Generally, ANM is defined as a tuple $(S, P(X))$, where $S = \{S_1, S_2, \cdots, S_n\}$ is a collection of n equations, S_i : x_i=$f_i(pa_{x_i})$+ε_i, i=1, 2, $\cdots$, n, where 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 $\varepsilon_i \perp pa_{x_i}$. ANM reflects the data-generating processes of X in the DAG G. We say a ANM is identifiable if it is asymmetrical in cause and effect and is capable of distinguishing between them. In fact, ANM is generally identifiable in nonlinear cases, all the non-identifiable cases are summarized in (Zhang and Hyvärinen 2009) (let the invertible mapping in Post-Nonlinear causal model (Zhang and Hyvärinen 2009) be identity mapping).

We consider such a scenario: given a DAG G where the data-generating procedure follows ANM, 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 paper we assume that all variables follow ANM.

In what follows, we present the theoretical results for characterizing CIs (i.e., $x_i \perp x_j|Z$) from the perspective of ANM, which underlie the proposed new method.

Theorem 1. *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 $x_i - f(Z) \perp x_j - g(Z)$, and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$.*

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 $\varepsilon_i \perp pa_{x_i}$, i.e., $x_i - f(pa_{x_i}) \perp pa_{x_i}$. For the reason that ε_i is an exogenous additive noise that is independent of x_i and all its non-descendant nodes, we have $\varepsilon_i \perp (x_j, pa_{x_i})$. Thus, given an arbitrary function g, we have $x_i - f(pa_{x_i}) \perp x_j - g(pa_{x_i})$. Similarly, if x_i is an ancestor of x_j, or x_i and x_j share common ancestors, we can also obtain $x_i - f(pa_{x_i}) \perp (pa_{x_i}, x_j - g(pa_{x_i}))$. 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 sufficient condition to complete Theorem 1. In many cases, we need not restrict $Z = pa_{x_i}$ or $Z = pa_{x_j}$ to meet $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$. For example, given a DAG of $x_1 \rightarrow z_1 \rightarrow x_2$ and $z_2 \rightarrow x_2$, if x_2 can be expressed by $x_2 = f_1(z_1) + f_2(z_2) + \varepsilon$, let $Z = z_1$ and g be an arbitrary function, we can also obtain $x_2 - f_1(Z) \perp x_1 - g(Z)$

and $x_2 - f_1(Z) \perp Z$. In what follows, we show that x_i and x_j are independent given such a Z.

Theorem 2. *Given two variables x_i and x_j, there is a set of variables Z and two functions f and g such that $x_i - f(Z) \perp x_j - g(Z)$, and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$, if the corresponding ANM is 1) linear, or 2) identifiable and faithfulness then $x_i \perp x_j|Z$.*

Proof. As $I(x_i; x_j|Z)$
$=I(x_i - f(Z); y - g(Z)|Z)$
$= I(x_i - f(Z); (x_j - g(Z), Z)) - I(x_i - f(Z); Z)$
$= I(x_j - g(Z); (x_i - f(Z), Z)) - I(x_j - g(Z); Z)$.
In linear case, if $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$, then $I(x_i - f(Z); (x_j - g(Z), Z)) = 0$ and $I(x_i - f(Z); Z) = 0$, hence $I(x_i; x_j, Z) = 0$, i.e., $x_i \perp x_j|Z$. Similarly, we can also deduce $I(x_j - g(Z); (x_i - f(Z), Z)) - I(x_j - g(Z); Z) = 0$ in the similar conditions. In the case that ANM is identifiable and faithfulness, let $\varepsilon_i = x_i - f(Z)$, if $x_i - f(Z) \perp Z$, then $\varepsilon_i \perp Z$. Assume that $x_i \not\perp x_j|Z$, then there must be at least one path P from x_i to x_j, or x_j to x_i, and P must via ε_i. Hence, $\varepsilon_i \not\perp x_j$, as $\varepsilon_i \perp Z \Rightarrow \varepsilon_i \perp g(Z)$, we have $\varepsilon_i \not\perp x_j - g(Z)$, that is contradictory. □

In (Zhang and Hyvärinen 2009), the authors showed that only very carefully chosen parameters can lead to a non-identifiable ANM in nonlinear cases. Therefore, Theorem 2 means that $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$ are generally sufficient to support $x_i \perp x_j|Z$. Combining Theorem 1 and Theorem 2, we can see that the CI test of $x_i \perp x_j|Z$ can be replaced by two unconditional independent tests $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$.

Therefore, according to the above two theorems, we can relax a CI test to at most three unconditional independent tests. In causal discovery, we need at most $3 * \sum_{i=1}^{|S|} C_{|S|}^i$ (S denotes the maximum conditional set, $Z \in S$) unconditional independent tests to determine whether x_i and x_j are CI in the worst case, while the existing CI testing methods need $\sum_{i=1}^{|S|} C_{|S|}^i$ CI tests. In what follows, we try to further simplify the conditions.

Let us consider the scenario that the data-generating procedure follows nonlinear ANM. Inspired by plenty of empirical results, we find that in practice only one unconditional independence test is enough. Thus, we have the following conjecture.

Conjecture 1. *Given a set of variables Z, two variables x_i and x_j and two functions f and g such that $x_i - f(Z) \perp x_j - g(Z)$, a necessary condition for $x_i \not\perp x_j|Z$ is that the corresponding ANM is not identifiable*[1].

To rationalize this conjecture, without loss of generality we assume that x_i is the ancestor (or parent) of x_j. If $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$, then we have $x_i \perp x_j|Z$ according to Theorem 2. If $x_i - f(Z) \not\perp Z$ and $x_j - g(Z) \not\perp Z$, we can generate two nodes v_i and v_j to extend the corresponding DAG with $v_i = x_i - f(Z) + \varepsilon_i$ and $v_j = x_j - g(Z) + \varepsilon_j$ where ε_i and ε_j are additive noise generated randomly, thus we have $v_i \not\perp Z$ and $v_j \not\perp Z$. If $x_i - f(Z) \perp x_j - g(Z)$, then we obtain $v_i \perp v_j$.

Combining $v_i \not\perp Z$ and $v_j \not\perp Z$ with $v_i \perp v_j$, we can deduce that v_i, Z and v_j form a V-structure (Cai, Zhang, and Hao 2013), i.e., both v_i and v_j are the parents of Z. Recall that $v_i = x_i - f(Z) + \varepsilon_i$ and $v_j = x_j - g(Z) + \varepsilon_j$ imply $Z \to v_i$ and $Z \to v_j$ respectively, then there are two cases: 1) the distribution is faithful to the original DAG with added edges $Z \to v_i$ or $Z \to v_j$, 2) the distribution is neither faithful to the original DAG with added edges $Z \to v_i$ nor that with added edges $Z \to v_j$ such that v_i and v_i can be expressed by a function of the other nodes but Z.

Therefore, in case 1, Z can be both the cause and the effect of v_i (or v_j). According to the mechanism of ANM (Hoyer et al. 2009), this can occur only in case that the ANM is not identifiable. All non-identifiable cases of ANM (let the invertible mapping in Post-Nonlinear causal model (Zhang and Hyvärinen 2009) be identity mapping) are summarized in (Zhang and Hyvärinen 2009), and they showed that only very carefully chosen parameters can lead to a non-identifiable ANM in nonlinear cases. That is, if the ANM is not identifiable, then the ANM is generally linear, strictly speaking, the causal relationship between v_i and z is linear, i.e., $v_i = x_i - f(Z) + \varepsilon_i$ where f is a linear function. In case 2, if both $Z \to v_i$ and $Z \to v_j$ can be removed, considering that $v_i = x_i - f(Z) + \varepsilon_i$ and $v_j = x_j - g(Z) + \varepsilon_j$, then x_i (x_j) is the parent of v_i (v_j). As x_i is the ancestor (or parent) of x_j, v_i cannot be independent of v_j.

Thus, if $x_i - f(Z) \perp x_j - g(Z)$ and $x_i - f(Z) \not\perp Z$ and $x_j - g(Z) \not\perp$, the corresponding ANM is not identifiable and generally we have either f or g is linear.

For example, consider a DAG of $x \to (y, Z)$ and $Z \to y$, where $Z = f(x) + \varepsilon_z$ (f is a linear function) and $y = h(x) + g(Z) + \varepsilon_j$. Then $x - f(Z) = \varepsilon_z$ is independent of $y - g(Z) = h(x) + \varepsilon_y$. In practice, even in such a case, it is not easy to find appropriate f and g to guarantee $x - f(Z) \perp y - g(Z)$.

In the context of nonlinear ANM, for two arbitrary variables, if they are not directly connected, we can surely find two nonlinear functions f and g to meet the condition $x_i - f(Z) \perp x_j - g(Z)$ according to Theorem 1. As shown in Theorem 3, $x_i - f(Z) \perp x_j - g(Z)$ covers $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$, which leads to $x_i \perp x_j|Z$ according to Theorem 2. Thus, instead of testing $x_i \perp x_j|Z$, we need only to check whether there exist two nonlinear functions f and g such that $x_i - f(Z) \perp x_j - g(Z)$.

A consequent advantage is that it is easy to find the target functions independently. We need only do nonlinear regression to find the minimum residuals of (x_i, Z) and (x_j, Z) respectively. Moreover, in causal discovery, causal directions are usually detected by determining V-structure and consistent propagation (Pearl 2009). For the reason that $x_i - f(Z) \perp x_j - g(Z)$ implies $x_i - f(Z) \perp Z$ or $x_j - g(Z) \perp Z$ according to Theorem 3, RCIT can capture more information about causal directions than determining only V-structure. This is because in nonlinear ANM, if $x - f(Z) \perp Z$, it rarely occurs that Z contains a child of x, which was also suggested in (Mooij et al. 2009):

[1] We found a flaw in the proof of this result given in an earlier version, and therefore, here it is treated as a conjecture.

whenever $\{z_1, ..., z_l\}$ contains a child of x, independence of $Residuals(\{z_1, ..., z_l\}, x)$ (fits x as a function of Z and returns the residuals) and $\{z_1, ..., z_l\}$ is rejected.

Compared with CI tests, RCIT can detect more causal directions even though there is no V-structure contained in the corresponding DAG. Consider a simple example, give a DAG: $x_1 \leftarrow x_2 \rightarrow x_3$, it is easy to find two functions f and g such that $x_1 - f(x_2) \perp x_2$ and $x_3 - g(x_2) \perp x_2$, so we can infer $x_1 \leftarrow x_2$ and $x_2 \rightarrow x_3$. However, it is difficult for CI 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 all of them fit the observed conditional and unconditional independence, though obviously with completely different structures.

Causal discovery based on RCIT

In this section, we present a new method of causality discovery based on the PC algorithm and RCIT. For convenience, the method is called PC$_{RCIT}$, which means PC algorithm based on RCIT. PC$_{RCIT}$ is developed based on the PC algorithm, in which we use RCIT to replace CI, and use existing methods (e.g., KCIT (Zhang et al. 2011)) to test unconditional independence.

We perform RCIT simply by estimating $\tilde{f}$ of f and $\tilde{g}$ of g, then we can test whether $x - \tilde{f}(Z) \perp y - \tilde{g}(Z)$, $x - \tilde{f}(Z) \perp Z$ or $y - \tilde{g}(Z) \perp Z$. If there is sufficient priori knowledge showing that the corresponding data-generating procedure follows nonlinear ANM, we can test only $x - \tilde{f}(Z) \perp y - \tilde{g}(Z)$ according to Conjecture 1.

Our method is outlined in Algorithm 1. The first step (Line 1 – 6) is to construct the causal skeleton by employing RCIT. The procedure follows the PC algorithm. That is, we form the complete undirected graph G on the set X of variables, then check whether every two variables x_i and x_j are conditional independent, given a set Z of variables, while saving the results of independence (e.g. recording that Z is the ancestor or parent of x_i if $x_i - f(Z) \perp Z$).

After obtain the causal skeleton, we orient the edges according to the results of independence and do consistent propagation (Line 7).

Finally, as a refinement step, we deal with the remaining unoriented edges by detecting V-structures and doing consistent propagation as in the PC algorithm. That is, to check whether a local structure $x_i - x_j - x_k$ can form a V-structure. If it is, orient it as $x_i \rightarrow x_j \leftarrow x_k$ (Line 8).

Note that in the case of nonlinear ANM, the performance of testing $x - \tilde{f}(Z) \perp y - \tilde{g}(Z)$, $x - \tilde{f}(Z) \perp Z$ or $y - \tilde{g}(Z) \perp Z$ is slightly different from that of testing only the first term. For example, the ground true is $x_i \not\perp x_j | Z$, let a and b denote the error rate of regression of f and g respectively. If we assume that unreliable $\tilde{f}$ (or $\tilde{g}$) will cause $x_i - \tilde{f}(Z) \perp x_j - \tilde{g}(Z) | Z$ and $x_i - \tilde{f}(Z) \perp Z$ (or $x_j - \tilde{g}(Z) \perp Z$), then the error rate of RCIT of testing 2 (or 3) terms is $a * b$, while the error rate of RCIT of testing 1 term is $a + b$. However, different assumptions may lead to different results, which is difficult to generalize. According to our observation, in many cases, the two kinds of RCIT testing are very close to each other in performance.

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

```
Input: variables set $X = \{x_1, ..., x_n\}$, threshold $k$.
Output: 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 $\exists Z \in X \setminus \{x_i, x_j\}$ and $(|Z| < k)$ such that $x_i \perp x_j | Z$ (estimated by RCIT) then
4:         delete edge $x_i - x_j$ from $G$ and save the results of independence (e.g. record '$Z$ to $x_i$' if $x_i - f(Z) \perp Z$).
5:     end if
6: end for
7: orient the edges of skeleton according the results of independence and then do consistent propagation.
8: orient the remaining un-oriented edges and do consistent propagation.
```

Performance evaluation

We apply the proposed method to both synthetic and real data to evaluate its practical performance and compare it with KCIT, CI$_{PERM}$ and partial correlation and their applications of PC algorithm. In our implementation, we perform the regression using Gaussian Processes (Rasmussen 2006) and the unconditional independence tests of RCIT using KCIT (Zhang et al. 2011).

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 errors (where the CI hypothesis is not rejected although being false) of RCIT change along with the size of the conditioning set Z ($D = 1, 2, ..., 5$) and the sample size ($n = 100$ and 200) in particular situations 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., other conditioning variables are independent of X, Y, and Z_1. We generate X and Y from Z_1 according to the ANM data generating procedure: they are generated as $f(g(Z_1)) + \varepsilon$ where f and g are randomly selected from sin, cos, tanh, square and cubic functions and are different for X and Y, and $\varepsilon \sim U(-0.2, 0.2)$. Hence, $X \perp Y | Z$ holds. In our simulations, Z_i is i.i.d. uniform $U(0, 1)$.

In Case II, all variables in the conditioning set Z are effective in generating X and Y. We first generate the independent variables Z_i, then X and Y are generated as $\sum_i f_i(g_i(Z_i)) + \varepsilon$ where f_i and g_i are randomly selected from sin, cos, tanh, square and cubic functions.

We compare RCIT with KCIT, CI$_{PERM}$ (with the standard setting of 500 bootstrap samples) and partial correlation test in terms of both types of errors. The significance level is fixed at 0.01. 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. To see how large they are for RCIT, we increase the dimensionality of Z and the sample size n, and repeat the tests 1000 random times.

We calculate Type I and II errors like this: for example $D = 3$, in Case I x should be independent of y given (Z_1), (Z_1, Z_2), (Z_1, Z_3) and (Z_1, Z_2, Z_3), then Type I error $=$1-$the\ number\ of\ CIs/4$. On the other side, x is independent of y given $\emptyset$, (Z_2), (Z_3) and (Z_2, Z_3), then Type II error $=$ $the\ number\ of\ CIs/4$. Similarly, we can calculate Type I and II errors in Case II.

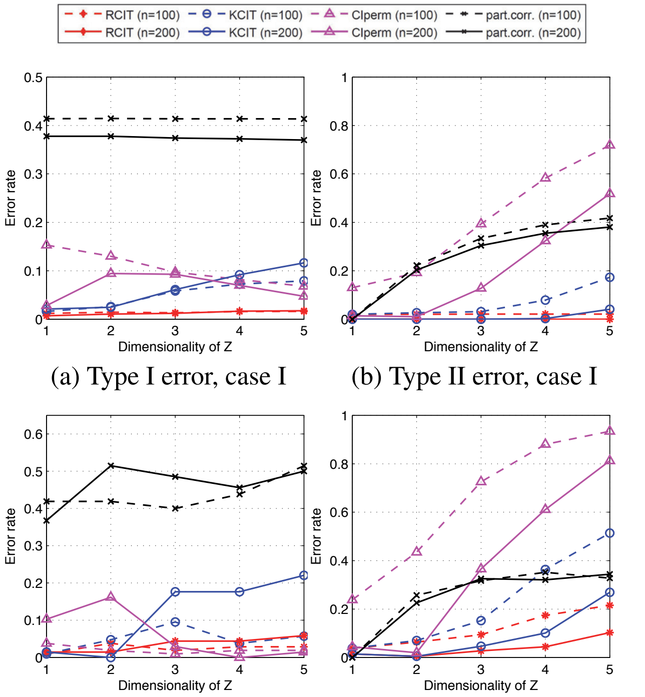
We first examine the Type I error in both case I and II. As shown in Fig. 1(a) and 1(c), the Type I error of RCIT is close to the significance level, and as D increases, the probability of Type I error slightly increases. One can see that in Case I, even when $D = 3$, the probability of Type I error of the other three methods is clearly larger than the significance level. Furthermore, KCIT and CI$_{PERM}$ are very sensitive to D. The curve of partial correlation tends to be a straight line parallel to the x axis, for the reason that all the experimental data follow nonlinear generating procedure aforementioned. A significant observation is that increasing sample size (from 100 to 200) does not reduce the Type I errors of the four methods.

In many scenarios, two disjoined variables are CI given a set of variables, thus a good test is expected to have a small probability of Type II error. As shown in Fig. 1(b) and 1(d), RCIT obtains clearly the best result. As D increases, the probability of Type II error always increases. Intuitively, this is reasonable: due to the finite sample size effect, as the conditioning set becomes larger and larger, X and Y tend to be considered as conditionally independent. On the other hand, as the sample size increases from 100 to 200, the probability of Type II error quickly approaches zero. In particular, as shown in Fig. 1(b), the curves of RCIT under 200 sample size keep close to zero. In contrast to the case of Type I error, the increasing sample size (from 100 to 200) can dramatically reduce Type II error.

Note that KCIT and RCIT have very similar performance when the dimensionality of Z is 1 and 2, which means that when a given DAG is very small, the two methods should perform similarly in discovering causal skeleton. However, RCIT can learn more information about the causal directions, which will be discussed in the next subsection.

Performance in causal discovery

CI tests are frequently used in causal inference where one assumes that the true causal structure of n random variables $x_1, ..., x_n$ can be represented by a directed acyclic graph (DAG) G. More specifically, the causal Markov condition assumes that the joint distribution satisfies all CIs that are imposed by the true causal graph (note that this is an assumption about the physical generating process of the data, not only about their distribution). The constraint-based methods like the PC algorithm make additional assumption of faithfulness (i.e., the joint distribution does not allow any CIs that are not entailed by the Markov condition) and recover the graph structure by exploiting the (conditional) 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 orients causal directions by finding V-structure and consistent propagations (Pearl 2009). In our experiments, we show that the PC algorithm based on RCIT can reveal much more causal directions as mentioned above.



(a) Type I error, case I

(b) Type II error, case I

(c) Type I error, case II

(d) Type II error, case II

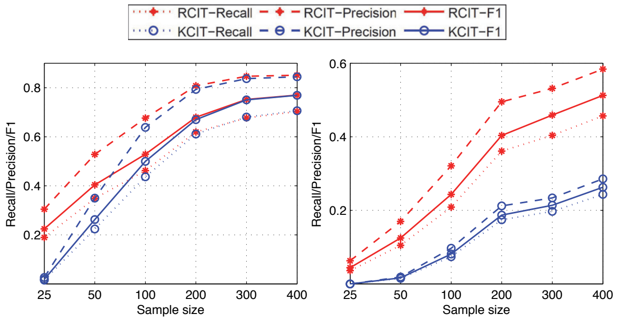
Figure 1: The probabilities of Type I and Type II errors obtained by simulation in various situations. Top: Case I (only one variable in Z is effective to X and Y). Bottom: Case II (all variables in Z are effective).

We generate data from a random DAG G. In particular, we sample four random variables $x_1, ..., x_4$ and allow arrows from x_i to x_j only for $i < j$. With probability 0.5 each possible arrow is either present or absent. The root variables are generated by $N(0, 1)$ and the leaf variables x_i are generated by $f(g(\sum_i f_i(g_i(pa_{x_i})))) + \varepsilon$ where f, g, f_i and g_i are randomly selected from sin, cos, tanh, square and cubic functions and are different for X and Y, and $\varepsilon \sim U(-0.2, 0.2)$ independent across pa_{x_i}. For significance level 0.01 and sample sizes between 25 and 400 we simulate 1000 DAGs, and evaluate the performance of different methods on discovering the causal skeleton and PDAG (including identifiable causal directions).

For the reason that RCIT and KCIT work significantly better than CI$_{PERM}$ and partial correlation as shown in Fig. 1 (for the performance comparison between KCIT, CI$_{PERM}$ and partial correlation in PC, see (Zhang et al. 2011)), here we compare PC$_{RCIT}$ with PC based on KCIT (denoted by PC$_{KCIT}$) for performance evaluation. To the best of our knowledge, in generic cases KCIT outperforms the other existing methods in term of discovering causality with PC when the input data are continuous.

As shown in Fig. 2(a), we can see that when the sample size is small (e.g. less than 200), PC$_{RCIT}$ performs significantly better than PC$_{KCIT}$. As the sample size increases, the performance of PC$_{KCIT}$ tends close to that of PC$_{RCIT}$. When the sample size up to 400, the F1 curves of PC$_{RCIT}$ and PC$_{RCIT}$ tend to overlap, but the former is still slightly (about 0.016) better than that of the latter. That is, although both RCIT and KCIT utilize regression, PC$_{RCIT}$ performs significantly better than PC$_{KCIT}$ in term of discovering causal skeleton when the sample size is small, which is the frequently-encountered case in reality. Obviously, our method is advantageous over the existing CI tests where a larger conditional set with a smaller sample will always lead to an incorrect conclusion, while in our method regression with unconditional test can perform significantly better.

We also evaluate the two methods in discovering PDAG. The results are presented in Fig. 2(b). We can see that PC$_{RCIT}$ achieves the best result in all cases, though the performance of PC$_{KCIT}$ in discovering causal skeleton is very close to that of PC$_{RCIT}$ when the sample size is large enough. The reason is that PC$_{KCIT}$ orients causal directions only based on V-structures and consistent propagations (Pearl 2009), in other words, returns only a set of Markov equivalence classes, while PC$_{RCIT}$ can uncover more causal directions by checking whether $x - f(Z) \perp Z$.

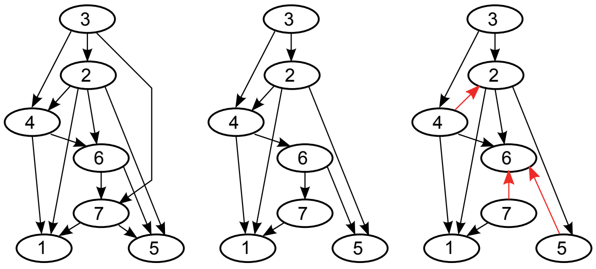


(a) Discovering skeleton (b) Discovering PDAG

Figure 2: Performance comparison between PC$_{RCIT}$ and PC$_{KCIT}$ for various sample sizes in term of discovering (a) causal skeleton and (b) PDAG.

Performance in causal direction inference We apply PC$_{RCIT}$ to the data set presented in (Mooij et al. 2009), which was generated following ANM w.r.t. a DAG consisting seven variables as shown in Fig. 3(a). For performance comparison in discovering causal direction, we choose two similar skeletons reconstructed by PC$_{RCIT}$ and PC$_{KCIT}$ with 1000 samples, which are shown in Fig. 3(b) and Fig. 3(c). We can see that all the causal edges discovering by PC$_{RCIT}$ are correct. However, as shown in Fig. 3(c), the directions of edges 2→4, 6→7 and 6→5 are incorrectly inferred by PC$_{KCIT}$. By taking the advantage of RCIT, existing constraint-based methods (e.g. the PC algorithm) can greatly improve the performance in causal discovery as RCIT helps to break the Markov equivalence classes.

Graphical modeling from medical data We finally apply RCIT to a real-word dataset used in a previous work (Fukumizu et al. 2007). The data consists of three variables, creatinine clearance (C), digoxin clearance (D), and urine flow (U). These were taken from 35 patients, and analyzed by graphical models in (Edwards 2012). Based on medical knowledge, D should be independent of U when controlling C, i.e., $D \perp U|C$ is known as the ground truth. The results are presented in Table 1. We can see that $D \perp U|C$ is strongly affirmed by using RCIT, while is not found by using partial correlation.



(a) (b) (c)

Figure 3: Performance comparison in causal direction inference. (a) ground truth causal model; (b) reconstructed DAG based on PC$_{RCIT}$; (c) reconstructed DAG based on PC$_{KCIT}$. Here, the red arrows indicate error directions.

Table 1: Results on real medical data.

Methods on testing $D \perp U\|C$	*P-value*
RCIT	0.1827
part.corr.	0.0037

Conclusion

In this paper, we propose a novel regression-based conditional independence testing approach based on additive noise model. In contrast to the existing CI testing methods, it makes use of the characterization of conditional independence in term of residuals (or exogenous noise) between variables. We show that once the causal process is assumed, the general CIs can be replaced by some weaker conditions. We relax the test of $x \perp y|Z$ to simpler unconditional independence tests of $x - f(Z) \perp y - g(Z)$, and $x - f(Z) \perp Z$ or $y - g(Z) \perp Z$ under the assumption that the data-generating procedure follows ANM. When the ANM is identifiable, we prove that $x - f(Z) \perp y - g(Z) \Rightarrow x \perp y|Z$. Compared to the exiting methods, our method is less sensitive to the dimensionality of Z. Experiments on both simulated and real world data show that the new method outperforms the existing techniques in discovering causality.

Acknowledgement This work was supported by the Key Projects of Fundamental Research Program of Shanghai Municipal Commission of Science and Technology under grant No. 14JC1400300. Jihong Guan was supported by the Program of Shanghai Subject Chief Scientist (15XD1503600).

References

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

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

Cai, R.; Zhang, Z.; and Hao, Z. 2013. Causal gene identification using combinatorial v-structure search. *Neural Networks* 43(7):63–71.

Edwards, D. 2012. *Introduction to graphical modelling*. Springer Science & Business Media.

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

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

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

Mooij, J.; Janzing, D.; Peters, J.; and Schölkopf, B. 2009. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, 745–752. ACM.

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

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

Rasmussen, C. E. 2006. Gaussian processes for machine learning.

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

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

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

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

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-based conditional independence test and application in causal discovery. 804–813. Corvallis, OR, USA: AUAI Press.