Testing Independence Between Linear Combinations for Causal Discovery

Hao Zhang[1], Kun Zhang[3], Shuigeng Zhou[2*], Jihong Guan[4], Ji Zhang[5]

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

Abstract

Recently, regression based conditional independence (CI) tests have been employed to solve the problem of causal discovery. These methods provide an alternative way to test for CI by transforming CI to independence between residuals. Generally, it is nontrivial to check for independence when these residuals are linearly uncorrelated. With the ability to represent high-order moments, kernel-based methods are usually used to achieve this goal, but at a cost of considerable time. In this paper, we investigate the independence between two linear combinations under linear non-Gaussian structural equation model (SEM). We show that generally the 1-*st* to 4-*th* moments of the two linear combinations contain enough information to infer whether or not they are independent. The proposed method provides a simpler but more effective way to measure CIs, with only calculating the 1-*st* to 4-*th* moments of the input variables. When applied to causal discovery, the proposed method outperforms kernel-based methods in terms of both speed and accuracy. which is validated by extensive experiments.

Introduction

In the problem of causal discovery, statistical independence and conditional independence (CI) tests are usually used for checking CIs among variables. For example, in the implementation of the PC algorithm (Spirtes, Glymour, and Scheines 2000), we use independence and CI tests to remove the edges that violate the joint distribution of given data. If two variables x and y that are conditional independent given a set of variables Z ($x, y \notin Z$), denoted by $x \perp\!\!\!\perp y|Z$, then given Z, further knowing x (or y) does not provide any additional information about y (or x). Therefore, we can deduce that there is no direct causal relationship between x and y if the faithfulness assumption holds (Pearl 2009).

In practice, independence and CI tests play a central role in causal discovery. In constraint-based methods (Pearl and Mackenzie 2018), the CI relationship $x \perp\!\!\!\perp y|Z$ allows us to separate $x-y$ when constructing a probabilistic model based on the joint distribution, which results in a parsimonious representation (Zhang et al. 2011). By using CI tests, constraint-based methods can generally return a partial directed acyclic graph (DAG) (Pearl 2009). In the causal functional model (Velikova et al. 2014; Peters et al. 2012; Zhang et al. 2016), there is a solution to infer causal directions by testing the independence between the set of independent variables x and the corresponding residual $R_{x\to y}$ (or the causal process of $P(y|x)$).

Without given any assumption or precondition, CI testing is generally more difficult than independence testing. Many existing methods are based on explicit estimation of conditional densities or their variants, or first discretizing the conditional set Z to a set of bins, and then transforming CI to independence in each bin (Diakonikolas and Kane 2016; Su and White 2008). For example, the method presented in (Su and White 2008) uses a characterization of CI, $P_{x|yZ}=P_{x|Z}$, to check CI by measuring the weighted Hellinger distance between estimates of conditional density. However, due to the curse of dimensionality, inevitably the required sample size increases dramatically when the conditional set becomes very large, which makes accurate estimation of conditional density or related quantity hard to be accomplished. Consider that the controlling set Z takes a finite number of values $\{z_1, ..., z_k\}$, 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 are distributed evenly on the values of Z, we must ensure that the independence within each subset of the sample with the same Z value by using only approximately n/k data points in each subset. When Z is 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. Here, "neighboring" is quantified by some distance metric, but finding neighboring points becomes more difficult as the dimensionality of Z grows.

As kernel methods are able to represent high order moments, kernel-based CI tests were developed to solve the above problems. In practice, mapping variables into reproducing kernel Hilbert spaces (RKHSs) allows us to infer properties of distributions, such as independence and homogeneity (Gretton et al. 2006). In (Fukumizu et al. 2007), the authors proposed 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. If the RKHSs are characteristic kernels, the zero operator norm is equivalen-

*Corresponding author.
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

t to $x \perp\!\!\!\perp y|Z$. (Daudin 1980) presented a characterization of CI that transforms CI to a set of zero correlations of regression functions. (Zhang et al. 2011) developed a method, called KCIT, by following this characterization of CI. They showed that $x \perp\!\!\!\perp y|Z$ if and only if 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). KCIT relaxes the spaces of functions f, g, r_f and r_g to RKHSs. (Doran et al. 2014) introduced the PKCIT method that utilizes permutation to convert the CI test problem into an easier two-sample test problem. (Strobl, Zhang, and Visweswaran 2017) used random Fourier features to approximate KCIT. (Lee and Honavar 2017) employed a modified unbiased estimate of maximum mean discrepancy to measure CI. Compared to discretization-based CI testing methods, kernel methods exploit more complete information of the data and incur less random error. It was showed that causal learning based on kernel methods can discover more accurate causalities.

Recently, regression-based tests were proposed for CI testing. (Grosse-Wentrup et al. 2016) transformed the CI of $x \perp\!\!\!\perp y|Z$ to independence between $x - f(Z)$ and (y, Z). (Zhang et al. 2017) used $x - f(Z) \perp\!\!\!\perp (y - g(Z), Z)$ to test $x \perp\!\!\!\perp y|Z$. These two methods infer the function f (or g) by regressing x (or y) on Z, then relax a CI test to a set of independence tests. One major drawback of these methods is that the independence conditions are just sufficient but not necessary to meet CI. Note that, $x - f(Z) \perp\!\!\!\perp Z$ is a strong condition, as $x - E(x|Z) \perp\!\!\!\perp Z \Rightarrow Z$ causes x in many cases (Zhang and Hyvärinen 2009). Moreover, when the dimensionality of Z becomes large, checking whether $x - f(Z)$ is independent from a set of variables (y, Z) or $(y - g(Z), Z)$ (joint distribution) tends to be prohibitively expensive. For example, in linear non-Gaussian cases, we often conduct $|y| + |Z|$ independence tests to check whether $x - f(Z) \perp\!\!\!\perp (y, Z)$ holds. (Flaxman, Neill, and Smola 2016) showed that given structural faithfulness and Markov assumptions (Pearl 2009), if Z causes x or y, $x \perp\!\!\!\perp y|Z$ is equivalent to $x - E(x|Z) \perp\!\!\!\perp y - E(y|Z)$. Similarly, here a strong condition that Z causes x or y is assumed. It can be seen that if these conditions are given, then it is easy to derive the corresponding causal relations.

In practice, given the faithfulness assumption, $x - E(x|Z) \perp\!\!\!\perp y - E(y|Z)$ and $x \perp\!\!\!\perp y|Z$ have significant correlations. For example, in (Ramsey 2014), the authors suggested to use $x - E(x|Z) \perp\!\!\!\perp y - E(y|Z)$ to test $x \perp\!\!\!\perp y|Z$ under the faithfulness assumption. In (Zhang et al. 2017), the authors further conjectured that $x - f(Z) \perp\!\!\!\perp y - g(Z)$ can lead to $x \perp\!\!\!\perp y|Z$ under nonlinear and faithfulness conditions, where f and g are nonlinear functions, x, y and Z are generated by nonlinear additive noise model. (Zhang, Zhou, and Guan 2018) showed that $x - E(x|Z) \perp\!\!\!\perp y - E(y|Z)$ is sufficient to support $x \perp\!\!\!\perp y|Z$ if the data is generated by following the linear non-Gaussian structural equation model (SEM) under the faithfulness assumption. As the residuals can be easily calculated by linear regression, the performance mainly depends on the independence test. Note that in this case, $cov(x - E(x|Z), y - E(y|Z)) = 0$ often holds. Therefore, it is difficult to detect the common component shared by $x - E(x|Z)$ and $y - E(y|Z)$. To get the best performance, this method (denoted by ReCIT) uses KCIT to achieve this goal, but it is computationally rather demanding.

In this work, we aim to measure the independence between the two residuals $x - E(x|Z)$ and $y - E(y|Z)$ in the case of ReCIT, where $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$, $z_j = \sum_{i=1}^{l} c_i s_i$ ($\forall z_j \in Z$) and $s_{1,...,l}$ are noise mutually independent. We show that the 1-*st* to 4-*th* moments of the two residuals contain enough information to infer whether they are independent or not. In general cases, the kurtosis of $x - E(x|Z) + y - E(y|Z)$ is not equal to that of $x - E(x|Z) + r$ where r has the same distribution as $y - E(y|Z)$ and satisfies $r \perp\!\!\!\perp x - E(x|Z)$ and $r \perp\!\!\!\perp y - E(y|Z)$. The only exception is that s_i is related to the causal data generating process. However, causal functional model requires that the distribution of noise is independent from the causal data generating process (Zhang et al. 2016). With this conclusion, we design an efficient independence test based on kurtosis and correlation, instead of spending time inferring the properties of distributions as kernel-based methods do. Extensive experiments show that our method performs better in testing independence, which makes ReCIT (Zhang, Zhou, and Guan 2018) much faster and get a better performance in causal discovery.

Independence Test Between Uncorrelated Linear Combinations

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

Consider the task as follows: given two randomly selected nodes x' and y', we want to test whether x' and y' are conditionally independent given a set of variables Z. According to the mechanism of ReCIT (Zhang, Zhou, and Guan 2018), the CI test of $x' \perp\!\!\!\perp y'|Z$ can be relaxed to an independence test between two residuals $x = x' - E(x'|Z)$ and $y = y' - E(y'|Z)$ in the linear non-Gaussian case. As the residuals x and y can be easily calculated by linear regression, the task turns to testing the independence between x and y. Concretely, the two variables (residuals) x and y are linear combinations of independent noise s_i ($i = 1, ..., l$) such that $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$. When x and y are correlated, we know $x \not\perp\!\!\!\perp y$ holds. However, if x and y are uncorrelated, then it is difficult to check whether x and y are independent or not. In what follows, we try to develop a low complexity method (compared to kernel-based methods) to measure independence between

two uncorrelated linear combinations.

Motivation

Given three linear combinations $x = s_1 + s_2$, $y = s_1 - s_2$ and $r = s_3 - s_4$, where $s_1, ..., s_4$ are independent non-Gaussian variables with the same distribution of zero mean. One can see that 1) x and y are uncorrelated but not independent, 2) x and r are uncorrelated and independent, 3) y and r are drawn from the same distribution. Assume all s_i are uniformly distributed, then we can see that $x + y$ and $x + r$ have different shapes of probability density function (PDF), as shown in Fig. 1. $x + y = 2s_1$ is still uniform distribution, while $x + r = s_1 + s_2 + s_3 - s_4$ is a linear combination of four i.i.d. noise variables, its PDF tends to be normally distributed according to Central Limit Theorem.

Here, we use kurtosis as the descriptor of the shape of a probability distribution. We investigate in what case the kurtosis of $x + y$ equals to that of $x + r$, i.e., $Kurt(x + y) = Kurt(x + r)$. We have

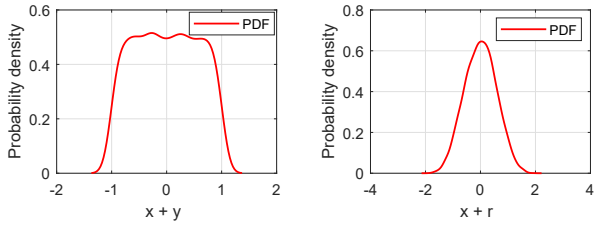
$$\begin{aligned} &Kurt(x+y) - Kurt(x+r) \\ =&Kurt(2*s_1) - Kurt(s_1+s_2+s_3-s_4) \\ =&\frac{16E(s_1^4)}{(var(x+y))^2} - \frac{E(\sum_{i=1}^4 s_i^4 - 6\sum_{i\neq j} s_i^2 s_j^2)}{(var(x+r))^2} \\ =&\frac{16E(s_1^4)}{(var(x+y))^2} - \frac{4E(s_1^4)+36(E(s_1^2))^2}{(var(x+r))^2} \end{aligned} \tag{1}$$

As $(var(x+y)))^2 = (var(x)+var(y)-2Cov(x,y))^2$ and $(var(x+r))^2 = (var(x)+var(r)-2Cov(x,r))^2$, we have $(var(x+y)))^2 = (var(x+r))^2$. Then in Eq. (1), if $Kurt(x + y)$ - $Kurt(x + r) = 0$, there must be $E(s_1^4) = 3(E(s_1^2))^2$. This means $Kurt(s_1) = 3$. That is, all the noise s_i have the same kurtosis as the normal distribution. This is a very strict condition for making $Kurt(x + y) = Kurt(x + r)$ hold.

In practice, we have to consider a more general case that $x = \sum_{i=1}^l a_i s_i$, $y = \sum_{i=1}^l b_i s_i$ and $R = \sum_{i=1}^l c_i s_i$, does there exist a similar strict condition to make $Kurt(x+y) = Kurt(x+r)$? This is exactly the motivation of this work.

Kurtosis Between Two Uncorrelated Variables

Here, we first reinstate the Darmois-Skitovitch theorem (Darmois 1953; Skitovich 1953), as it is fundamental to prove the subsequent theorem.



(a) (b)

Figure 1: (a) Probability density function of $x + y$, (b) probability density function of $x + r$.

Darmois-Skitovitch theorem (DST) Define two random variables x and y as linear combinations of independent random variables s_i $(i = 1, ..., l)$, $x = \sum_{i=1}^l a_i s_i$, $y = \sum_{i=1}^l b_i s_i$. Then, if $x \perp\!\!\!\perp y$, 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. We have the following theorem:

Theorem 1. *Given three linear combinations* $x=\sum_{i=1}^l a_i s_i$, $y=\sum_{i=1}^l b_i s_i$ *and* $r=\sum_{i=1}^l b_i e_i$ *where* x *and* y *are uncorrelated,* s_i, e_i *are independent non-Gaussian variables with the same distribution of zero mean, there is at least one* i *such that* $a_i b_i \neq 0$. *If* $Kurt(x + y)=Kurt(x + r)$, *then* $Kurt(s_i) = \frac{6\sum a_i^2 b_j^2 - \sum_{i\neq j}(4a_i^2 a_j b_j + 4b_i^2 b_j a_j + 6a_i^2 b_j^2 + 6a_i a_j b_i b_j)}{\sum(4a_i^3 b_i + 4b_i^3 a_i + 6a_i^2 b_i^2)}$.

Proof. Consider the kurtosis of $x + y$, we have

$$\begin{aligned} &Kurt(x+y) \\ =&\frac{E((x+y)^4)}{(var(x+y))^2} \\ =&\frac{E(x^4+y^4+4x^3y+4xy^3+6x^2y^2)}{(var(x+y))^2}, \end{aligned} \tag{2}$$

$$\begin{aligned} &Kurt(x+r) \\ =&\frac{E(x^4+r^4+4x^3r+4xr^3+6x^2r^2)}{(var(x+r))^2}, \end{aligned} \tag{3}$$

$$\begin{aligned} &E(4x^3r+4xr^3+6x^2r^2) \\ =&6E(x^2r^2) \\ =&6\sum a_i^2 b_j^2 E(s_i^2)E(e_j^2) \\ =&6\sum a_i^2 b_j^2((E(s_i^2))^2, \end{aligned} \tag{4}$$

Therefore, if $Kurt(x + y)$ - $Kurt(x + r) = 0$, the necessary and sufficient condition is $E(4x^3y+4xy^3+6x^2y^2) = 6E(x^2r^2)$.

We can see that

$$\begin{aligned} &E(4x^3y+4xy^3+6x^2y^2) \\ =&4E\{(\sum a_i^3 s_i^3 + \sum_{i\neq j} a_i^2 a_j s_i^2 s_j + \sum_{i\neq j\neq k} a_i a_j a_k s_i s_j s_k)\sum b_i s_i\} \\ &+4E\{(\sum b_i^3 s_i^3 + \sum_{i\neq j} b_i^2 b_j s_i^2 s_j + \sum_{i\neq j\neq k} b_i b_j b_k s_i s_j s_k)\sum a_i s_i\} \\ &+6\sum a_i^2 b_i^2 E(s_i^4) + \sum_{i\neq j} a_i^2 b_j^2 E(s_i^2 s_j^2) + 4\sum_{i\neq j} a_i a_j b_i b_j E(s_i^2 s_j^2) \\ =&\sum(4a_i^3 b_i + 4b_i^3 a_i + 6a_i^2 b_i^2)E(s_i^4) \\ &+\sum_{i\neq j}(4a_i^2 a_j b_j + 4b_i^2 b_j a_j + 6a_i^2 b_j^2 + 6a_i a_j b_i b_j)((E(s_i^2))^2 \end{aligned} \tag{5}$$

Combining Eq. (2) ~ (5), to ensure that $Kurt(x + y) = Kurt(x + r)$, there must be

$$\frac{E(s_i^4)}{(E(s_i^2))^2} = Kurt(s_i)$$
$$= \frac{6\sum a_i^2 b_j^2 - \sum_{i\neq j}(4a_i^2 a_j b_j + 4b_i^2 b_j a_j + 6a_i^2 b_j^2 + 6a_i a_j b_i b_j)}{\sum(4a_i^3 b_i + 4b_i^3 a_i + 6a_i^2 b_i^2)}. \quad (6)$$

□

Proposition 1. *Given three random variables x, y and r that are generated by the linear non-Gaussian structural equation model such that* $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$ *and* $r = \sum_{i=1}^{l} b_i e_i$ *where x and y are uncorrelated,* s_i *and* e_i *are independent with the same distribution of zero mean, there is at least one i such that* $a_i b_i \neq 0$. *If* $Kurt(x+y) = Kurt(x+r)$, *then* $Kurt(s_i) = \frac{6\sum a_i^2 b_j^2 - \sum_{i\neq j}(4a_i^2 a_j b_j + 4b_i^2 b_j a_j + 6a_i^2 b_j^2 + 6a_i a_j b_i b_j)}{\sum(4a_i^3 b_i + 4b_i^3 a_i + 6a_i^2 b_i^2)}$.

It is straightforward to derive this conclusion from **Theorem 1**. According to the mechanism of causal functional model, the distribution shape of disturbance should not be related to data generating process. This means that in the linear non-Gaussian case, if $Kurt(x+y) = Kurt(x+r)$, then we generally can conclude that x is independent of y.

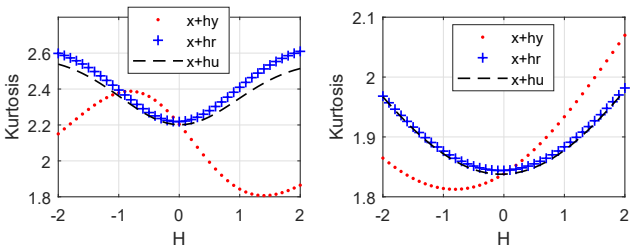
Proposition 2. *Given three random variables x′, y′, r and* $Z = \{z_1, ..., z_m\}$ *that are generated by the linear non-Gaussian structural equation model such that* $x = x' - E(x'|Z) = \sum_{i=1}^{l} a_i s_i$, $y = y' - E(y'|Z) = \sum_{i=1}^{l} b_i s_i$ *and* $r = \sum_{i=1}^{l} b_i e_i$ *where x and y are uncorrelated,* s_i *and* e_i *are independent with the same distribution of zero mean, there is at least one i such that* $a_i b_i \neq 0$. *If* $Kurt(x+y) = Kurt(x+r)$, *then* $Kurt(s_i) = \frac{6\sum a_i^2 b_j^2 - \sum_{i\neq j}(4a_i^2 a_j b_j + 4b_i^2 b_j a_j + 6a_i^2 b_j^2 + 6a_i a_j b_i b_j)}{\sum(4a_i^3 b_i + 4b_i^3 a_i + 6a_i^2 b_i^2)}$.

Similar to **Proposition 1**, this conclusion can be straightforwardly derived from **Theorem 1**. As x is the residual of linear regression of (x', Z), the coefficients a_i and b_j depend on the linear non-Gaussian structural equation model, the 1-*st* and 2-*nd* moments of s_i. Generally, it is difficult to deduce the 4-*th* moment from the 1-*st* and 2-*nd* moments in linear non-Gaussian cases, which means the distribution of noise is related to the data generating process. That is, in general linear non-Gaussian cases, if $Kurt(x+y) = Kurt(x+r)$, we can conclude that $x' - E(x'|Z) \perp\!\!\!\perp y' - E(y'|Z)$, i.e., $x' \perp\!\!\!\perp y'|Z$ if the corresponding conditions in RCIT holds.

Fast Regression Based Conditional Independence Test

In this section, we study how to measure the difference between $Kurt(x+y)$ and $Kurt(x+r)$. According to DST, if x and y are uncorrelated but not independent, they must share at least two common noise variables. However, as we do not know the coefficients of x and y, $Kurt(x+y)$ can be very close to $Kurt(x+r)$. For example, we scale y and r by $10^{10} * y$ and $10^{10} * r$ respectively, then $Kurt(x+y)$ and $Kurt(x+r)$ have almost the same shape of distribution.

To explain the main idea of testing independence, here we give an example. We generate x, y and R such that $x = \sum_{i=1}^{l} a_i s_i$, $y = \sum_{i=1}^{l} b_i s_i$, $r = \sum_{i=1}^{l} b_i e_i$, and $u = \sum_{i=1}^{l} b_i e_i'$ where s_i, e_i and e_i' are independent disturbances. There are at least two values of i such that $a_i b_i \neq 0$. We scale y, r and u by a set of weights $H = \{h_1 = -2, ..., h_m = 2\}$. As shown in Fig. 2(a), in this case the PCC of x and y is $\rho_{x,y} = -0.28$. We can deduce that x and y are not independent. Fig. 2(a) shows the curves of $Kurt(x+hy)$ and $Kurt(x+hr)$, they are definitely different. In the second scenario, we generate x and y with very small $\rho_{x,y} = 0.007$ such that one cannot determine whether x and y are independent or not. As shown in Fig. 2(b), the curve of $Kurt(x+hy)$ is still different from $Kurt(x+hr)$.



(a) $\rho_{x,y} = -0.28$ (b) $\rho_{x,y} = 0.007$

Figure 2: The curves of PCC and kurtosis with distinct $\rho_{x,y}$, (a) $|\rho_{x,y}| > 0.2$; (b) $|\rho_{x,y}| < 0.01$.

As mentioned above, the difference between $Kurt(x+hy)$ and $Kurt(x+hr)$ can be used to test CI. With these theoretical results, we design a new method for **F**ast **R**egression based **C**onditional **I**ndependence **T**est (FRCIT in short). The details of FRCIT are given in Alg. 1.

We aim to test the CI of $x' \perp\!\!\!\perp y'|Z$. We first follow the general process of RCIT, do the linear regression on (x', Z) and (y', Z), and calculate the residuals $x = x'\text{-}E(x'|Z)$ and $y = y'\text{-}E(y'|Z)$ (Line 1). In the second step, we generate the data of counterparts that have the same distribution as y (like the variables r and k shown in Fig. 2). However, the noise variables and coefficients contained in y are unobservable. Alternatively, we directly resample y by using permutation or bootstrap methods, and denote the t new variables by $r_1, ..., r_t$ (Line 2). We calculate $t+1$ kurtosis vectors of $x+h_i(y, r_1, ..., r_t)$ with different weights $H = \{h_1, ..., h_m\}$, these vectors are denoted by K_y, K_{r_1}, ..., K_{r_t} respectively (Line 3). K_y, K_{r_1}, ..., K_{r_t} are curves like those shown in Fig. 2 (a) and (b). In the final step, we calculate the average vector of K_{r_1}, ..., K_{r_t}, and denote it by K. Intuitively, if y is independent of x, the gap between K_y and K would not be bigger than that between K_{r_i} and K. Here, PCC is used to measure the similarity. If $\rho_{k,k_y} > \rho_{k,k_{r_i}}$, we return $x' \perp\!\!\!\perp y'|Z$, otherwise return $x' \not\perp\!\!\!\perp y'|Z$ (Lines 4-10).

As FRCIT is used for linear CI testing, therefore FRCIT can be directly applied to the PC algorithm for linear causality discovery. For more details about using regression based CI test in the PC algorithm, the readers can refer to (Zhang, Zhou, and Guan 2018). In the following section, we will evaluate the performance of FRCIT in causal discovery.

Performance Evaluation

We first compare FRCIT with ReCIT (Zhang, Zhou, and Guan 2018) by extensive simulated experiments. To the best of our knowledge, ReCIT is one of the best CI test-

Algorithm 1 Fast regression based conditional independence test (FRCIT)

Input: variables: x', y', Z, weights: $H = \{h_1, ..., h_m\}$, the number of counterparts: t.
Output: the decision of $x' \perp\!\!\!\perp y'|Z$ or $x' \not\perp\!\!\!\perp y'|Z$.

Calculate the residuals x=(x'-$E(x'|Z)$ and y=(y'-$E(y'|Z)$.
Resample y t times, and the new variables are denoted by $r_1, ..., r_t$.
Let $K_y = \{Kurt(x + h_1 y), ..., Kurt(x + h_m y)\}$ and $K_{r_i} = \{Kurt(x+h_1 r_i), ..., Kurt(x+h_m r_i)\}$
Let K=$\{\sum_{i=1}^{k} Kurt(x+h_1 r_i)/m, ..., \sum_{i=1}^{k} Kurt(x+h_m r_i)/m\}$
for $\forall K_{r_i}$ $(i = 1, ..., k)$ **do**
    if $\rho_{K,K_y} > \rho_{K,K_{r_i}}$ **then**
        Return $x' \perp\!\!\!\perp y'|Z$.
    end if
end for
Return $x' \not\perp\!\!\!\perp y'|Z$

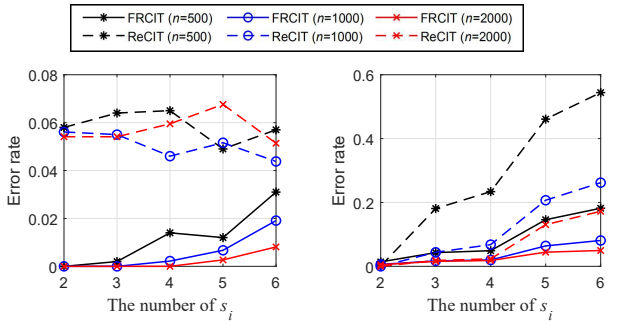
ing methods in linear cases. There are many comparisons among ReCIT, KCIT and other CI testing methods presented in the previous works (Zhang et al. 2011, 2017; Zhang, Zhou, and Guan 2018). We then illustrate the advantage of FRCIT in causal structure learning. We compare our method with existing causal learning methods, including PC$_{ReCIT}$, SADA-LiNGAM (Cai, Zhang, and Hao 2017) and DirectLiNGAM (Shimizu et al. 2011), over various real-world causal structures. Note that all these methods can distinguish Markov equivalence classes, among them PC$_{ReCIT}$ stands for the state of the art in these cases.

Effect of the Number of s_i and Sample Size

We first examine how the probabilities of Type I error (where the CI hypothesis is incorrectly rejected) and Type II error (where the CI hypothesis is not rejected although being false) of FRCIT and ReCIT change with the number of noise variables involved in x and y ($d = 2, ..., 6$) and the sample size (n = 500, 1000 and 2000) by simulation. Here, we consider two cases as follows:

In Case I, x and y are independent. We generate x and y according to the linear non-Gaussian SEM data generating procedure: $x = \sum_{i=1}^{l} a_i * s_i$ and $y = \sum_{i=1}^{l} b_i * s_i$ where $a_i, b_i \sim U(-1, -0.2) \cup U(0.2, 1)$ are different for x and y, s_i is i.i.d. sampled from $\sim U(-0.5, 0.5)$, and $a_i b_i = 0$ holds.

In Case II, x and y are dependent. Similarly, x and y are generated by $x = a_1 s_1 + a_2 s_2 + \sum_{i=3}^{l} a_i * s_i$ and $y = b_1 s_1 + b_2 s_2 + \sum_{i=3}^{l} b_i * s_i$ where $a_i, b_i \sim U(-1, -0.2) \cup U(0.2, 1)$ are different for x and y, s_i is i.i.d. sampled from $\sim U(-0.5, 0.5)$, and $a_i b_i = 0, b_2 = -a_1 b_1 / a_2$ hold. We can see that in this case x and y are not correlated but $x \not\perp\!\!\!\perp y$. As the controlling set $Z = \emptyset$, ReCIT uses KCIT to test the independence of $x - E(x|Z)$ and $y - E(y|Z)$. This is the major difference between FRCIT and ReCIT. In this group of experiments we aim to compare FRCIT with ReCIT in terms of both types of error. The significance level of ReCIT is fixed at $\alpha = 0.05$. We check how the errors change when increasing the number of s_i (with $a_i b_i \neq 0$) and the sample size n. For each parameter setting, we randomly repeat the testing 1000 times and average their results.



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

Figure 3: (a) Type I error rate in Case I with the ground truth of $x \perp\!\!\!\perp y$; (b) Type II error rate in Case II with the ground truth of $x \not\perp\!\!\!\perp y$.

As shown in Fig. 3(a), Type I error rate of FRCIT is much smaller than that of ReCIT, and is close to zero. As d (the number of s_i) increases, the Type I error rate increases slightly. On the other hand, the Type I error rate of ReCIT is close to the significance level $\alpha = 0.05$.

As shown in Fig. 3(b), for Case II the Type II error rate of FRCIT is also lower than that of ReCIT. As d increases, the Type II error rates of both FRCIT and ReCIT increases, but FRCIT increases more slowly than ReCIT. We can also see that the increase of sample size (from 500 to 2000) obviously reduces the Type II error rate. Generally, it is difficult to detect the common component between x and y in Case II, as the sum of s_i makes both x and y tend to be Gaussian.

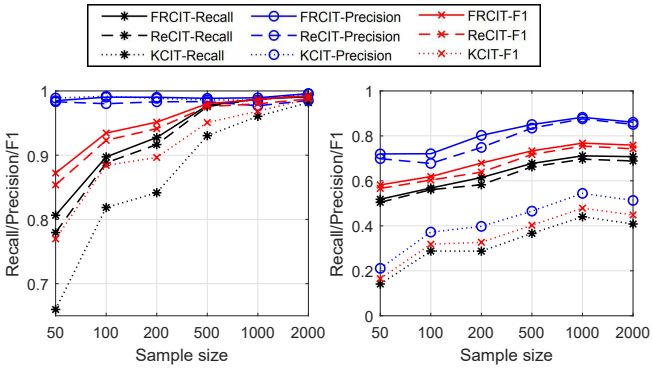
We then compare the efficiency between FRCIT and ReCIT in terms of elapsed time.As presented in Table 1, the elapsed time of FRCIT increases slowly with sample size, and when the sample size is small, the elapsed time difference between FRCIT and ReCIT is not obvious. However, the time consumed by ReCIT is up to 30 times of that of FRCIT when the sample size is larger than 2000.

Performance on Causal Discovery

CI tests are frequently used in causal discovery where we usually assume that the true causal structure of n random variables $x_1, ..., x_n$ can be represented by a DAG G. Concretely, the causal Markov condition assumes that the joint distribution satisfies all CIs that are imposed by the true causal graph. The CI testing-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 con-

Sample size	Elapsed time (s)	
	FRCIT	ReCIT
100	0.0404	0.0433
500	0.0554	0.1585
1000	0.0645	0.5108
2000	0.0970	2.9821

Table 1: Efficiency comparison between FRCIT and ReCIT.



(a) Skeleton learning (b) Structure learning

Figure 4: Performance comparison among PC_{FRCIT}, PC_{ReCIT} and PC_{KCIT} with various sample sizes in discovering (a) causal skeleton and (b) PDAG.

dition, and recover the graph structure by exploiting the CIs and independence that can be found in the data. Therefore, the performance of CI testing-based methods almost depends on the performance of CI tests.

In this group of experiment, the datasets are generated from a random DAG G. In particular, we sample four random variables $x_1, ..., x_4$ and allow arrows from x_i to x_j only for $i < j$. With probability 0.5 each possible arrow is either present or absent. The root variables are generated by $U(0, 1)$ and the leaf variables x_i are generated by $\sum_i a_i * pa_{x_i} + \varepsilon$ where $a_i \sim U(0.2, 1)$ and $\varepsilon \sim U(-0.2, 0.2)$ independent across pa_{x_i}. For significance level 0.05 and sample sizes between 50 and 2000, we simulate 1000 DAGs and evaluate the performance of the three methods PC_{FRCIT}, PC_{ReCIT} and PC_{KCIT} on discovering causal skeleton and PDAG (including identifiable causal directions). To save time, we employ a faster partial correlation testing method (Cai, Zhang, and Hao 2017) before checking CI, if x and y are correlated given Z, we output $x \not\perp\!\!\!\perp y|Z$; if x and y are not correlated given Z, we further use CI testing methods to check CI.

As shown in Fig. 4(a), when the sample size is small (e.g. less than 500), the performances of PC_{FRCIT} and PC_{ReCIT} are very close to each other, and are significantly better than PC_{KCIT}. As the sample size increases, the performance of PC_{KCIT} tends be close to those of the other two methods. When the sample size up to 1000, the $F1$ curves of the three methods tend to overlap, but the $F1$ of PC_{FRCIT} is still slightly (about 0.007) better than that of the PC_{ReCIT} and about 0.014 better than that of the PC_{KCIT}. As the regression coefficient $Z(Z^TZ)^{-1}Z^T$ in FRCIT and 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_{FRCIT} performs significantly better than PC_{ReCIT} in discovering causal skeleton according to the results presented Fig. 3.

We also evaluate the three methods in discovering PDAG. The results are presented in Fig. 4(b). We can see that PC_{FRCIT} achieves better result in all cases, though the performance of PC_{KCIT} in discovering causal skeleton is very close to that of PC_{ReCIT} when the sample size is up to 1000. The reason is that PC_{KCIT} orients causal directions only based on V-structure and consistent propagation (Pearl 2009), in other words, returns only a set of Markov equivalence classes, while PC_{FRCIT} and PC_{ReCIT} can uncover more causal directions (for more details please refer to the *Algorithm* 1 presented in (Zhang, Zhou, and Guan 2018)).

Performance on Causal Structures

Here we compare PC_{FRCIT} with three existing causal structure learning methods, including PC_{ReCIT} (Zhang, Zhou, and Guan 2018), SADA-LiNGAM (Cai, Zhang, and Hao 2017) and DirectLiNGAM (Shimizu et al. 2011). As all these methods can break Markov equivalence classes, we therefore can evaluate the advantage of our method in causal direction learning. The implementation of the three existing methods strictly follow the corresponding original papers. All methods are evaluated on six real-world causal network structures[1] that cover a variety of applications, including insurance evaluation (*Insurance*), medicine (*Alarm*), decision support system (*Barley*), weather forecasting (*Hailfinder*), system troubleshooting (*Win95pts*) and expert system (*Pathfinder*). These causal networks contain nodes from 22 to 109, and have been used in many related works like SADA-LiNGAM and ReCIT. Because there are not large-scale causal inference problems with ground truth, simulated data on real-world structures are used in most causal structure learning works (Kalisch and Bühlmann 2007). The structures and data generating processes are similar to those presented in (Cai, Zhang, and Hao 2017). We use 500 samples in the experiments.

The results are shown in Table 2, where for saving space in the table, PC_{FRCIT}, PC_{ReCIT}, SADA-LiNGAM and DirectLiNGAM are simply denoted as PCF, PCR, SL and DL, respectively. We can see that PC_{FRCIT} achieves the best *Recall*, *Precision* and $F1$ score on almost all structures. In many cases, the accuracy of PC_{ReCIT} is very close to that of PC_{FRCIT}, as the only difference between the two methods is the independence tests. As the performance of SADA-LiNGAM and DirectLiNGAM are heavily impacted by the rate of $\frac{Sample\ size}{The\ number\ of\ nodes}$, if we fix the sample size, their accuracy values decrease with the number of nodes. For performance comparison among PC_{ReCIT}, SADA-LiNGAM, DirectLiNGAM and other causal structure learning methods, we refer the readers to (Cai, Zhang, and Hao 2013, 2017; Zhang, Zhou, and Guan 2018).

Performance on Real-World Gene Expression Data

In this section, we evaluate our method on real-world gene expression data in term of causal genes identification (Ruichu et al. 2013). This data (Golub and R. 1999) is a collection of 72 samples from leukemia patients, with each sample giving the expression levels of 7129 genes. According to pathological/histological criteria, these sample include 47 type I Leukemias (called ALL) and 25 type II Leukemias (called AML).

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

Dataset	*Recall*				*Precision*				*F*1			
	PCF	PCR	SL	DL	PCF	PCR	SL	DL	PCF	PCR	SL	DL
Insurance	**0.79**	0.71	0.73	0.63	0.69	0.61	0.66	**0.81**	**0.74**	0.65	0.69	0.71
Alarm	**0.74**	0.65	0.57	0.37	**0.88**	0.86	0.43	0.57	**0.79**	0.74	0.49	0.45
Barley	**0.65**	0.61	0.59	0.60	**0.64**	0.63	0.41	0.61	**0.64**	0.62	0.48	0.61
Hailfinder	**0.72**	0.68	0.53	0.40	**0.75**	0.67	0.51	0.50	**0.74**	0.68	0.52	0.44
Win95pts	**0.89**	0.87	0.51	0.38	**0.86**	0.82	0.47	0.50	**0.87**	0.85	0.49	0.43
Pathfinder	**0.96**	**0.96**	0.89	0.50	**0.62**	0.61	0.39	0.34	**0.76**	0.75	0.54	0.41

Table 2: Performance of four causal learning methods on real-world causal structures.

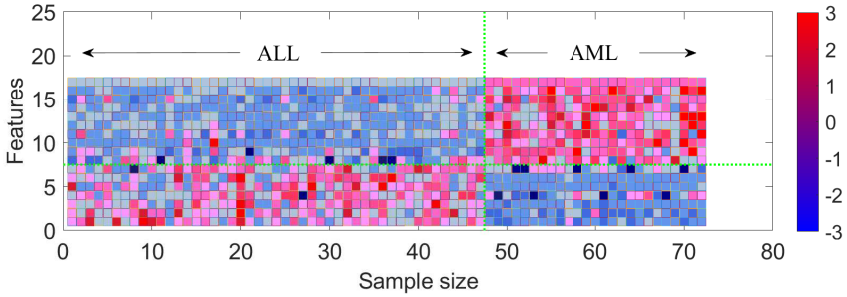


Figure 5: The discovered genes distinguish ALL from AML.

We evaluate the results returned by FRCIT (combined with PC algorithm) from two aspects:
1. How many causal genes discovered by FRCIT can be verified by the previous works that they are really directly related to the occurrence/treatment of human acute leukemia?
2. Can these genes differentiate AML and ALL?

There are 16 causal genes returned by FRCIT, their Gene Accession Number (GAN) are J05243_at, M11147_at, M11722_at, M23197_at, M55150_at, M63138_at, M84526_at, U46499_at, X17042_at, X63097_at, X95735_at, M31523_at, U05259_rna1_at, X98833_rna1_at, and HG2562HT2658_s_at, respectively. According to the previous works, it can be verified that 9/16 genes are related to the occurrence/treatment of leukemia, they are J05243_at (Zhou, Liu, and Wong 2004), M11722_at (Sasaki et al. 1996), M23197_at (Sievers et al. 2001), X63097_at (Gurda and Turowska 1970), X95735_at (Wang et al. 2003), M31523_at (Meriem et al. 2017), X98833_rna1_at (Joko et al. 2007), U05259_rna1_at (Thompson et al. 1997) and HG2562HT2658_s_at (Tang, Zhang, and Huang 2007), respectively.

We presented the expression levels of the 16 discovered causal genes in Fig. 5. The expression level of each gene is normalized across the samples such that the mean is zero and the standard deviation is one. For each gene, the expression level greater than the mean is shaded in red, and that below the mean is shaded in blue. One can see that the discovered 16 genes can distinguish ALL from AML, and there is not any gene uniformly expressed across the class.

Conclusion

In this paper, we propose a new and fast regression based conditional independence (CI) test method FRCIT to support effective and efficient causality discovery under the linear structural equation model (SEM) with non-Gaussian noise variables. Concretely, we provide a simple way to test the independence between two linear combinations x=x'-$E(x'|Z)$ and y=y'-$E(y'|Z)$ returned by linear regression. We show that if the kurtosis of $x + y$ (denoted by $Kurt(x + y)$) equals to that of $x + r$ where r is drawn from the same distribution as y and meets $r \perp\!\!\!\perp (x, y)$, then the distribution of noise variable in SEM is related to the coefficients. This implies that the two kurtosis cannot be the same in general cases or violate the mechanism of causal functional model.

As mentioned in Proposition 2, FRCIT requires that all the disturbances s_i are independent with the same distribution. Evidently it is a strict precondition, which makes it difficult to apply FRCIT to general cases. In future work, we aim to remove this precondition.

Acknowledgements

This work was sponsored by Zhejiang Lab (2019KB0AB05), and partially supported by National Natural Science Foundation (NSFC) (62006051, U1636205 and U1936205). Hao Zhang was also partially supported by Projects of Talents Recruitment of GDUPT (2020rc02). Kun Zhang was supported by the United States Air Force under Contract No. FA8650-17-C-7715. Shuigeng Zhou was also partially supported by 2019 Special Fund for Artificial Intelligence Innovation & Development, Shanghai Economy and Information Technology Commission (SHEITC).

References

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

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

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

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

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

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

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

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

Golub; and R., T. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286(5439): 531–537.

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

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

Gurda, M.; and Turowska, B. 1970. Distribution of the ABO and Rh D blood groups in patients with leukemia. *Polish Medical Science & History Bulletin* 13(2): 89.

Joko, T.; Nanba, D.; Shiba, F.; Miyata, K.; Shiraishi, A.; Ohashi, Y.; and Higashiyama, S. 2007. Effects of promyelocytic leukemia zinc finger protein on the proliferation of cultured human corneal endothelial cells. *Molecular Vision* 13(1): 649–658.

Kalisch, M.; and Bühlmann, P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8(Mar): 613–636.

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

Meriem; Ben-Ali; Jing; Yang; Koon; Wing; Chan; Imen; Ben-Mustapha; and and, N. 2017. Homozygous transcription factor 3 gene (TCF3) mutation is associated with severe hypogammaglobulinemia and B-cell acute lymphoblastic leukemia. *Journal of Allergy & Clinical Immunology* .

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

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

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

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

Ruichu; Cai; ; ; Zhenjie; Zhang; ; ; Zhifeng; and Hao. 2013. Causal gene identification using combinatorial V-structure search. *Neural Networks* .

Sasaki, R.; Fukushima, M.; Miura, Y.; Chang, L. M. S.; and Bollum, F. J. 1996. Sensitivity and applicability of different methods for detection of terminal transferase in leukemia. *Leukemia* 10(8): 1377–1382.

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

Sievers, E. L.; Larson, R. A.; Stadtmauer, E. A.; and Estey. 2001. Efficacy and Safety of Gemtuzumab Ozogamicin in Patients With CD33-Positive Acute Myeloid Leukemia in First Relapse. *Journal of Clinical Oncology* 19(13): 3244–3254.

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

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

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

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

Tang, Y.; Zhang, Y. Q.; and Huang, Z. 2007. Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis. *IEEE/ACM Transactions on Computational Biology & Bioinformatics* 4(3): 365–381.

Thompson, A. A.; Talley, J. A.; Do, H. N.; Kagan, H. L.; and Wall, R. 1997. Aberrations of the B-cell receptor B29 (CD79b) gene in chronic lymphocytic leukemia. *Blood* 90(4): 1387–1394.

Velikova, M.; van Scheltinga, J. T.; Lucas, P. J.; and Spaanderman, M. 2014. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare.

International Journal of Approximate Reasoning 55(1): 59–73.

Wang; Y.; Gilmore; T.; and D. 2003. Zyxin and paxillin proteins: focal adhesion plaque LIM domain proteins go nuclear. *Biochimica Et Biophysica Acta* .

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

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

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

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

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

Zhou, X.; Liu, K. Y.; and Wong, S. T. C. 2004. Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics* 37(4): 249–259.