Differentially Private Nonlinear Causal Discovery from Numerical Data

Hao Zhang[1], Yewei Xia[1], Yixin Ren[1], Jihong Guan[2], Shuigeng Zhou[1]*

[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Department of Computer Science & Technology, Tongji University, China
{haoz15, ywxia21, yxren21, sgzhou}@fudan.edu.cn; jhguan@tongji.edu.cn

Abstract

Recently, several methods such as private ANM, EM-PC and Priv-PC have been proposed to perform differentially private causal discovery in various scenarios including bivariate, multivariate Gaussian and categorical cases. However, there is little effort on how to conduct private nonlinear causal discovery from numerical data. This work tries to challenge this problem. To this end, we propose a method to infer nonlinear causal relations from observed numerical data by using regression-based conditional independence test (RCIT) that consists of kernel ridge regression (KRR) and Hilbert-Schmidt independence criterion (HSIC) with permutation approximation. Sensitivity analysis for RCIT is given and a private constraint-based causal discovery framework with differential privacy guarantee is developed. Extensive simulations and real-world experiments for both conditional independence test and causal discovery are conducted, which show that our method is effective in handling nonlinear numerical cases and easy to implement. The source code of our method and data are available at https://github.com/Causality-Inference/PCD.

Introduction

Causal discovery is to reason the causal relations among observed variables, which is generally a challenging task if no controlled experiments are available. From the computational perspective, causal discovery is usually formulated as a probabilistic graphical model on a set of variables such that each directed edge between two variables represents a causal link. In constraint-based methods (Pearl and Mackenzie 2018), the conditional independence (CI) $X \perp\!\!\!\perp Y|\boldsymbol{Z}$ enables us to separate two nodes $X - Y$ when constructing a probabilistic model based on the joint distribution, which leads to a parsimonious representation (Zhang et al. 2011).

Up to now, a series of independence and CI tests have been proposed to support constraint-based causal discovery, including conventional tests such as Spearman's ρ (Spiegelman 2010), Kendall's τ (Kendall 1938), $\mathcal{G}$-test (McDonald 2009) and $\mathcal{X}^2$ test (Rao 2002), Kernel-based tests such as Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005) and KCIT (Zhang et al. 2011), and regression-based CI tests (RCIT) (Ramsey 2014; Zhang et al. 2017). With these tests, constraint-based causal discovery can be effectively fulfilled in different scenarios.

Recently, some methods are proposed to address differential private causal discovery due to increasing privacy violation concerns. (Kusner et al. 2016) studies the problem of causal inference using the additive noise model (ANM) while simultaneously ensuring privacy of the users, and conducts sensitivity analysis for different dependence scores, including Spearman's ρ, Kendall's τ, interquartile range (IQR) and HSIC. As causal direction learning with ANM can be achieved by comparing the two scores between $X \to Y$ and $Y \to X$, there is no need to consider the sensitivity of independence statistics, such a private ANM is easy to implement but works only in bivariate cases. (Xu, Yuan, and Wu 2017) presents the EM-PC algorithm to handle differentially private causal discovery in multivariate cases. This algorithm is a modification to the well-known PC algorithm (Spirtes, Glymour, and Scheines 2000) to guarantee differential privacy by using the exponential mechanism. Instead of perturbing each independence test with noise, EM-PC randomly decides how many and which edges to delete using the exponential mechanism. In this way, EM-PC manages to achieve a relative balance between utility and privacy. The state-of-the-art solution to differentially private causal discovery is Priv-PC (Lun, Qi, and Dawn 2020), which achieves better utility and efficiency than EM-PC. Sensitivity analyses for conditional Kendall's τ and conditional Spearman's ρ are also performed, showing that conditional Spearman's ρ can be used only to large datasets due to the large coefficient in its sensitivity while conditional Kendall's τ works better in generally cases. For treatment effect analysis, (Lee et al. 2019) proposes a differentially private inverse probability weighting method for average treatment effect, and (Niu et al. 2022) provides a meta-algorithm for differentially private estimation of the conditional average treatment effect.

These works above significantly advance differential privacy causal discovery in different scenarios. However, this area has not yet been extensively explored. Particulary, as (Lun, Qi, and Dawn 2020) points out, an important research issue in this area is how to perform private causal discovery in nonlinear numerical cases where neither Spearman's ρ nor Kendall's τ and other existing tests with differential privacy guarantee are effective. In fact, nonlinear numerical cases are more common than the linear, categorical cases in real-world

*Corresponding author.

scenarios. So in this work, we aim to explore private causal discovery techniques for handling nonlinear numerical cases.

Our contributions are summarized as follows:

- We propose an effective method for private causal discovery from observed nonlinear numerical data by using regression-based CI test (RCIT) that consists of kernel ridge regression (KRR) (An, Liu, and Venkatesh 2007) and HSIC with permutation approximation.
- We perform the sensitivity analysis of RCIT and show that the time complexity of evaluating the bound of its local sensitivity is $O(k)$ (k is the number of permutations). Furthermore, we develop a private constraint-based causal discovery framework with differential privacy guarantee.
- We validate the effectiveness and advantage of our method by extensive experiments on both synthetic and real-world datasets. Experimental results show that our method is effective in handling nonlinear numerical cases and easy to implement.

Preliminaries

Here, we briefly review some important concepts about differential privacy, CI test and private causal discovery.

Differential Privacy

The concept of differential privacy is first formally introduced by (Dwork et al. 2006), which is now considered the golden standard for private analysis.

Definition 1. *((ϵ, δ)-differential privacy). A randomized algorithm $\mathcal{A}$ with input domain $\mathbb{D}$ and output range Rand($\mathcal{A}$) is (ϵ, δ)-differential private for $\epsilon, \delta \geq 0$, if for all $O \subseteq$ Rand($\mathcal{A}$) and for any two neighboring datasets $\mathcal{D}$ and $\tilde{\mathcal{D}} \in \mathbb{D}$, we have*

$$\mathbb{P}[\mathcal{A}(\mathcal{D}) \in O] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(\tilde{\mathcal{D}}) \in O] + \delta.$$

Particularly, if $\delta = 0$, it is called ϵ-differential privacy.

Typically, two datasets $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are considered to be neighbors when they differ by only one tuple. Here, we denote this by $\mathcal{D} \simeq \tilde{\mathcal{D}}$.

A common approach to achieving differential privacy is to perturb the output with additive random noise. The noise is carefully calibrated to appropriately hide the maximum difference of the output, which is defined as *sensitivity*.

Definition 2. *(Global sensitivity). The global sensitivity of a algorithm $\mathcal{A} : \mathbb{D} \to \mathbb{R}$ is:*

$$\triangle_{\mathcal{A}} := \max_{\mathcal{D},\tilde{\mathcal{D}} \in \mathbb{D} \ s.t. \ \mathcal{D} \simeq \tilde{\mathcal{D}}} |\mathcal{A}(\mathcal{D}) - \mathcal{A}(\tilde{\mathcal{D}})|$$

It may be that the global sensitivity of an algorithm $\mathcal{A}$ is unbounded in general, but can be bounded in the context of a specific data set $\mathcal{D}$ over all its neighbors $\tilde{\mathcal{D}}$. For such datasets we can bound the local sensitivity.

Definition 3. *(Local sensitivity). The local sensitivity of a algorithm $\mathcal{A} : \mathbb{D} \to \mathbb{R}$ is:*

$$\triangle(\mathcal{D})_{\mathcal{A}} := \max_{\tilde{\mathcal{D}} \in \mathbb{D} \ s.t. \ \mathcal{D} \simeq \tilde{\mathcal{D}}} |\mathcal{A}(\mathcal{D}) - \mathcal{A}(\tilde{\mathcal{D}})|$$

If an algorithm has bounded global sensitivity it certainly has bounded local sensitivity. Local sensitivity can also lead to differential privacy for datasets (Nissim, Raskhodnikova, and Smith 2007; Jain and Thakurta 2013).

Hypothesis Conditional Independence Testing

Generally, independence and CI tests are the key tools for constraint-based causal discovery. Assume two variables X, Y and a set of variables $\mathbf{Z}$ constitute n *i.i.d.* pairs $(X_i, Y_i, \mathbf{Z}_i)$, the problem of testing CI between X and Y given $\mathbf{Z}$ can be written in the form of a hypothesis testing:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y|\mathbf{Z} \quad \text{versus} \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y|\mathbf{Z}.$$

Hypothesis CI testing generally consists of the following steps: First, define the relevant CI test statistic T and calculate the value of T from the observational data. Then, given a user-selected significance level α (typically set to 0.05), which indicates the lower bound of the probability threshold for rejecting $\mathcal{H}_0$, i.e., the null hypothesis $\mathcal{H}_0$ is rejected if the p-value $\leq \alpha$.

There are two types of errors that may occur during hypothesis testing. Type I error means the rejection of $\mathcal{H}_0$ when it is actually true, and Type II error is the acceptance of $\mathcal{H}_0$ although it is not true. A well performed CI test requires that Type I error rate is not greater than the chosen significance level, while Type II error rate is as low as possible (Zhang et al. 2011).

Private Causal Discovery

In statistics, causal discovery is usually formulated as a causal graph — directed acyclic graph (DAG) to encode the assumptions about the generating process of a set of variables, such that any directed edge between two variables indicates a causal relationship. Causal discovery refers to the process of discovering the hidden causal graph from a given observed dataset, and if differential privacy is required to be preserved during causal discovery, we call this process *private causal discovery*.

In constraint-based causal discovery (Pearl and Mackenzie 2018), the causal graph can be (partially) recovered by applying CI tests to the observed variables (Zhang et al. 2022a). The CI of two nodes v_i and v_j allows us to separate them by constructing a probabilistic model, based on the joint distribution of observed variables under the faithfulness assumptions (Spirtes, Glymour, and Scheines 2000):

$$v_i \perp\!\!\!\perp v_j|\tilde{V} \Longrightarrow \tilde{V} \ dsep. \ v_i - v_j,$$

where $\tilde{V}$ is the controlling set and $dsep.$ denotes d-separation (Pearl 2009). The performance of constraint-based causal discovery is usually heavily affected by Type II error of CI test, as it may lead to edges being falsely removed (Zhang et al. 2017, 2022b). Comparatively speaking, Type I error has less impact on causal discovery, because once a Type I error occurs, the CI test will continue with another controlling set.

In this work, we try to design a new constraint-based method to achieve private causal discovery. Our focus is on nonlinear causal relationships among numerical data. We aim to ensure that 1) the designed CI tests are able to handle the concerned scenario with good control over Type I & II error rates; 2) either global sensitivity or local sensitivity of the statistic is bounded for the CI tests, and the tighter the bound, the better. In practice, infinite sensitivity (or bounded by 1) would fail to achieve private causal discovery.

Sensitivity Analysis of Regression-Based Conditional Independence Test

In this section, we introduce RCIT (Ramsey 2014; Zhang et al. 2017) that is used for causal discovery in this work. RCIT relaxes CI test to independence test:

$$X - \mathbb{E}[X|\mathbf{Z}] \perp\!\!\!\perp Y - \mathbb{E}[Y|\mathbf{Z}] \Longrightarrow X \perp\!\!\!\perp Y|\mathbf{Z},$$

which contains two regressions and one independence test. In fact, the independence of the two residuals is not exactly equivalent to CI without additional assumption (Flaxman, Neill, and Smola 2016; Zhang et al. 2019). However, such a method works well in many scenarios like causal discovery. To derive a smaller bound for the sensitivity of RCIT and be more effective in dealing with nonlinear numerical data, here we use kernel ridge regression (KRR) for regression and HISC with permutation approximation for independence test. To achieve private RCIT, in what follows we conduct the sensitivity analysis in two steps: first bound the sensitivity of empirical RCIT score, and then bound the sensitivity of the statistic for RCIT.

Bound Sensitivity of Empirical RCIT Score

We consider the empirical RCIT score for statistical dependence measurement between two residuals $X - \mathbb{E}[X|\mathbf{Z}]$ and $Y - \mathbb{E}[Y|\mathbf{Z}]$ w.r.t. RCIT.

Definition 4. *(Empirical RCIT score) Let* $(X, Y, \mathbf{Z}) := \{(X_1, Y_1, \mathbf{Z}_1), ..., (X_m, Y_m, \mathbf{Z}_m)\}$ *be a series of m independent observations drawn from the joint distribution* $\mathbb{P}_{XY\mathbf{Z}}$*, an empirical estimator of RCIT score is given by*

$$\widehat{RCIT}(X, Y, \mathbf{Z}) := \widehat{HSIC}_{k,l}(U, V) := \frac{1}{(m-1)^2}\boldsymbol{tr}(KHLH)$$

where $\widehat{HSIC}$ *is the empirical HSIC score,* $U = X - \mathbb{E}[X|\mathbf{Z}]$*,* $V = Y - \mathbb{E}[Y|\mathbf{Z}]$*),* $H = I - 1/m$*,* $K_{ij} = k(u_i, u_j)$*,* $L_{ij} = l(v_i, v_j)$*, k and l are the given kernels. Particularly, if* $\mathbf{Z} = \emptyset$*, then*

$$\widehat{HSIC}_{k,l}(U, V) = \widehat{HSIC}_{k,l}(X, Y).$$

In this work, we use Gaussian kernel for HSIC and KRR. Following the mechanism of HSIC, $\widehat{HSIC}_{k,l}(U, V) = 0$ can approximate $U \perp\!\!\!\perp V$ with sufficient samples (Large m) (Rindt, Sejdinovic, and Steinsaltz 2020). A bound of global sensitivity for HSIC with two random variables X and Y has been studied in previous work (Kusner et al. 2016). For readability, in what follows we simply denote $\widehat{HSIC}_{k,l}$ by $\widehat{HSIC}$.

Theorem 1. *(Kusner et al. 2016) The empirical HSIC score given in Def. 4 has the global sensitivity bounded by*

$$|\widehat{HSIC}(X, Y) - \widehat{HSIC}(\tilde{X}, \tilde{Y})| \leq \frac{12m - 11}{(m-1)^2} \tag{1}$$

where $\tilde{X} \simeq X$ *and* $\tilde{Y} \simeq Y$*.*

We can see that the bound given in Equ. (1) can be used for private dependence measurement (not test statistic) between two random variables X and Y, but it is only applicable to the case involving controlling set $\mathbf{Z} = \emptyset$, so we need to further consider the more complex case of $\mathbf{Z} \neq \emptyset$.

In this work, we use KRR to calculate the two residuals U and V, so the regression function regarding $(X, \mathbf{Z})$ can be written as $f(\mathbf{w}, \mathbf{Z}) = \mathbf{w}^T\phi(\mathbf{Z})$, where $\phi(\mathbf{Z})$ is a feature space mapping to the Hilbert space $\mathcal{H}$ regarding a chosen kernel function. Then, the KRR used in RCIT can be written as

$$\mathbf{w} = arg\min_{\mathbf{w}\in\mathcal{H}} \frac{\lambda}{2}\|\mathbf{w}\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\phi(\mathbf{Z}_i) - X_i)^2, \tag{2}$$

where λ is the regularization parameter. Let $\tilde{f}(\mathbf{w}^*, \cdot)$ and $\tilde{f}(\tilde{\mathbf{w}}^*, \cdot)$ be the regression functions obtained from the optimization problem in Equ. (2) with input variables $(X, \mathbf{Z})$ and $(\tilde{X}, \tilde{\mathbf{Z}})$, respectively, where $\tilde{X} \simeq X$ and $\tilde{\mathbf{Z}} \simeq \mathbf{Z}$. We then have the following result:

Theorem 2. *Say* $\lambda \leq 1$*, given the regression function* $\tilde{f}(\boldsymbol{w}^*, \cdot)$ *and* $\tilde{f}(\tilde{\boldsymbol{w}}^*, \cdot)$ *obtained from the optimization problem in Equ. (2), denote the residuals of the two functions by U and* $\tilde{U}$*, respectively, then we have*

$$|u_i - \tilde{u}_i| \leq \frac{8}{m\lambda^{3/2}}$$

for all i, where u_i *and* $\tilde{u}_i$ *are the i-th elements of U and* $\tilde{U}$*, and m is the sample size.*

The proof of Theorem 2 can be derived from Theorem 5 in (Kusner et al. 2016), which is also presented in Appendix A (in the Supplementary Material) for the interested readers. Now we put Theorem 1 and Theorem 2 together to obtain the bound of sensitivity of RCIT score (not the test statistics of p-value) as follows:

Theorem 3. *Let* $(X, Y, \mathbf{Z}) := \{(X_1, Y_1, \mathbf{Z}_1), ..., (X_m, Y_m, \mathbf{Z}_m)\}$ *be a series of m independent observations drawn from* $\mathbb{P}_{XY\mathbf{Z}}$*, U and V are the residuals obtains from Equ. (2) with* $(X, \mathbf{Z})$ *and* $(Y, \mathbf{Z})$*, respectively,* $\tilde{U}$ *and* $\tilde{V}$ *are similarly obtained by neighbors* $(\tilde{X}, \tilde{\mathbf{Z}})$ *and* $(\tilde{Y}, \tilde{\mathbf{Z}})$*, then the sensitivity of RCIT score is bounded by*

if $\mathbf{Z} \neq \emptyset$*, then*

$$\begin{aligned} &|\widehat{HSIC}(U, V) - \widehat{HSIC}(\tilde{U}, \tilde{V})| \\ &\leq \frac{2(m^2 - m)(2 - e^{-\frac{128}{m^2\lambda^3\delta_k^2}} - e^{-\frac{128}{m^2\lambda^3\delta_l^2}})}{(m-1)^2}, \end{aligned} \tag{3}$$

if $\mathbf{Z} = \emptyset$*, then*

$$|\widehat{HSIC}(U, V) - \widehat{HSIC}(\tilde{U}, \tilde{V})| \leq \frac{12m - 11}{(m-1)^2}, \tag{4}$$

where δ_k *and* δ_l *are the parameters of bandwidth regarding Gaussian kernels k and l, respectively,* λ *is the regularization parameter in Equ. (2).*

Proof of Theorem 3 is given in Appendix B.

For readability, we denote the bounds of $|\widehat{HSIC}(U, V) - \widehat{HSIC}(\tilde{U}, \tilde{V})|$ in Equ. (3) and Equ. (4) by "$\triangle_{\mathcal{R}}$".

Note that Equ. (4) can be derived from Theorem 1 that corresponds to the case of independence test, we mainly fucus on the proof of Equ. (3) w.r.t. CI test. On the other hand, if δ_k, δ_l and λ can be given in advance by prior knowledge, then Equ. (3) leads to a bound for global sensitivity, otherwise gives a bound for local sensitivity. In practice, the kernel

bandwidth generally heavily relies on the input data, thus a global sensitivity with fixed bandwidth may reduce the robustness of private RCIT. Therefore, a local global sensitivity seems to be more useful in general cases.

Bear in mind that we cannot test CI by using RCIT score directly, but need to further design a test statistic based on this score, and finally compare the p-value with the significant level. In what follows, we will show how to bound the sensitivity of CI test statistic.

Bound Sensitivity of Test Statistic of RCIT

In fact, the test statistics for many CI tests including RCIT is general unbounded or bounded by 1 in many cases, such a loose bound is useless when compared with the significant level. Alternatively, we try to bound the local sensitivity.

There are two general ways to produce the test statistic of HSIC: Gamma approximation and permutation/bootstrap approximation (Gretton et al. 2005). As we find that it is very time-consuming to calculate a bound for the sensitivity w.r.t. Gamma approximation (we need to solve an optimization problem that requires considerable extra time and might lead to poor performance. More details are presented in Appendix C for the interested readers), in what follows we focus on how to bound the sensitivity w.r.t. permutation approximation.

We follow the notation of (Pfister et al. 2018) for the permutation approximation of HSIC. Define maps $\psi_i : \{1, ..., m\} \to \{1, ..., m\}$ for $i = 1, ..., k$, then ψ_i maps U (or V) by

$$\psi_i(U) := \psi_i(u_1, ..., u_m) \tag{5}$$

It can be seen that each ψ_i is a random permutation of $\{1, ..., m\}$. Then, we can calculate the k new scores of $\widehat{HSIC}(\psi_i(U), V), ..., \widehat{HSIC}(\psi_k(U), V)$, and each of which corresponds to the hypothesis test $\mathcal{H}_0 : U \perp\!\!\!\perp V$, we thus can compare the k scores with $\widehat{HSIC}(U, V)$ to derive the independence test statistic:

Reject $\mathcal{H}_0$ if

$$\frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V)\}}{k} \leq \alpha,$$

otherwise reject $\mathcal{H}_1$, where $\mathbf{1}$ is the indicator function, α is the given significant level.

We now consider the sensitivity of test statistic of HSIC between two residuals U and V, e.g., RCIT, using permutation approximation. We have the following result:

Theorem 4. *Let p and $\tilde{p}$ be the test statistic of HSIC with permutation approximation of (U, V) and $(\tilde{U}, \tilde{V})$, respectively, then*

$$|p - \tilde{p}| \leq \max\{p - \mathcal{A}, \mathcal{B} - p\} \tag{6}$$

where

$$\mathcal{A} = \frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V) - 2\triangle_{\mathcal{R}}\}}{k} \tag{7}$$

and

$$\mathcal{B} = \frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V) + 2\triangle_{\mathcal{R}}\}}{k}. \tag{8}$$

Proof. Consider the test statistic of $\widehat{HSIC}(U, V)$ with permutation approximation:

$$p = \frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V)\}}{k}.$$

There are two cases, (1) reject $\mathcal{H}_1 : U \not\perp\!\!\!\perp V$ if $p > \alpha$, and (2) reject $\mathcal{H}_0 : U \perp\!\!\!\perp V$ if $p \leq \alpha$.

Case (1). In this case, we need to consider how much p will decrease after changing one sample. We have

$$\widehat{HSIC}(\tilde{U}, \tilde{V}) - \triangle_{\mathcal{R}} \leq \widehat{HSIC}(U, V) + \triangle_{\mathcal{R}}$$

and

$$\widehat{HSIC}(\psi_k(\tilde{U}), \tilde{V}) + \triangle_{\mathcal{R}} \geq \widehat{HSIC}(\psi_k(U), V) - \triangle_{\mathcal{R}}.$$

Then yield

$$\begin{aligned} &|p - \tilde{p}| \\ &\leq |p - (\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) + \triangle_{\mathcal{R}} \\ &\quad < \widehat{HSIC}(\psi_k(U), V) - \triangle_{\mathcal{R}}\})/k| \\ &= \left| p - \frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V) - 2\triangle_{\mathcal{R}}\}}{k} \right|. \end{aligned} \tag{9}$$

Case (2). In this case, similarly we need to consider how much p will increase after changing one sample. Therefore we have

$$\begin{aligned} &|p - \tilde{p}| \\ &\leq \left| \frac{\sum_{i=1}^{k} \mathbf{1}\{\widehat{HSIC}(U, V) < \widehat{HSIC}(\psi_k(U), V) + 2\triangle_{\mathcal{R}}\}}{k} - p \right|. \end{aligned} \tag{10}$$

The two equations yield $|p - \tilde{p}| \leq \max\{p - \mathcal{A}, \mathcal{B} - p\}$. □

For simplicity, we denote the bound of $|p - \tilde{p}|$ regarding permutation approximation in Equ. (6) by "$\triangle_{\mathcal{P}}$". Since $\widehat{HSIC}(U, V)$ and $\widehat{HSIC}(\psi_i(U), V)$ are calculated in testing independence between U and V with a time complexity of $O(km^2)$, therefore the time cost of calculating $\triangle_{\mathcal{P}}$ is trivial in private (conditional) independence test.

Corollary 1. *Given $1 + k$ scores of $\widehat{HSIC}(U, V)$ and $\widehat{HSIC}(\psi_i(U), V)$ $(i = 1, ..., k)$, the time complexity of calculating the bound of sensitivity in Equ. (6) is $O(k)$.*

Proof of Corollary 1 is straightforward according to Equ. (7) and Equ. (8), we need only to call the indicator function $2k$ times to return the bound in Equ. (6). Generally, k can be set to 100 (Gretton et al. 2005).

In fact, the bound $\triangle_{\mathcal{P}}$ can be further improved. Generally, we do not need to consider both cases mentioned in the proof of Theorem 4 at the same time. Concretely, assume the test rejects $\mathcal{H}_0$ with $p \leq \alpha$, then the decreasing of p after changing one sample would not affect the decision of rejecting $\mathcal{H}_0$, i.e., we need only to consider the Case (2) that how much p will increase after changing one sample. Similarly, if the test rejects $\mathcal{H}_1$ with $p > \alpha$, we need only to consider Case (1)

that how much p will decrease after changing one sample. Consequently, we can improve $\triangle_{\mathcal{P}}$ by $\triangle_{\mathcal{P}} = \mathcal{B} - p$ in the case of $p \leq \alpha$, and by $\triangle_{\mathcal{P}} = p - \mathcal{A}$ in the case of $p > \alpha$.

Next, we will develop an effective framework for private causal discovery based on RCIT.

Private Causal Discovery Based on RCIT

In this work, we use $\triangle_{\mathcal{P}}$ to achieve private causal discovery. We call the proposed framework Differentially **P**rivate Nonlinear **C**ausal **D**iscovery (PCD in short). The details are outlined in Alg. 1. PCD is similar to the PC algorithm in non-private part, which starts with constructing a complete graph on V (Line 1). We then test the independence and CI between any two nodes $v_i, v_j \in V$ given the controlling set $\mathbf{Z}$ by using RCIT, here simply denote the corresponding statistic by $r(ij|\mathbf{Z})$. There are two cases: $r(ij|\mathbf{Z}) \geq \alpha$ (Lines 4-11) and $r(ij|\mathbf{Z}) \leq \alpha$ (Lines 12-17). Consider the first case, we need to judge whether $r(ij|\mathbf{Z})$ minus its bound of sensitivity is small enough to affect the result. If not, we directly accept $\mathcal{H}_0$ and delete the edge $v_i - v_j$ from $\mathcal{G}$ (Line 11). If yes, we apply sparse vector technique (SVT) (Dwork and Roth 2014) to achieving differentially privacy (Lines 8&16). Similarly, for the second case, we need to decide whether $r(ij|\mathbf{Z})$ plus its bound of sensitivity is large enough to affect the result, and then determine whether or not to apply SVT (Lines 11-14). Line 15 follows the PC algorithm to orient causal directions by determining V-structure and performing consistent propagation. Finally, we output the causal graph $\mathcal{G}$ (generally a partial DAG)(Line 19). In what follows, we give the differential privacy guarantee for PCD.

Algorithm 1: Differentially Private Nonlinear Causal Discovery (PCD)

Input: V: vertex set, $\mathcal{D}$: dataset, ϵ: privacy parameter, ϵ_t: total privacy and is initialized to 0, α: significant level.
Output: $\mathcal{G}$: causal graph of V.

1: Construct a complete graph on V, denote by $\mathcal{G}$
2: **for** $\forall v_i, v_j \in V$ and $\forall \mathbf{Z} \subseteq V \backslash_{v_i, v_j}$ **do**
3: calculate the statistic $r(ij|\mathbf{Z})$
4: **if** $r(ij|\mathbf{Z}) > \alpha$ **then**
5: calculate $\triangle_{\mathcal{P}} = r(ij|\mathbf{Z}) - \mathcal{A}$, $\mathcal{A}$ follows Equ. (7)
6: **if** $r(ij|\mathbf{Z}) - \triangle_{\mathcal{P}} \leq \alpha$ **then**
7: $\epsilon_t = \epsilon_t + \epsilon$
8: **if** $r(ij|\mathbf{Z}) + Lap(\frac{4\triangle_{\mathcal{P}}}{\epsilon}) \geq \alpha + Lap(\frac{2\triangle_{\mathcal{P}}}{\epsilon})$ **then**
9: delete edge $v_i - v_j$ from $\mathcal{G}$
10: **else**
11: delete edge $v_i - v_j$ from $\mathcal{G}$
12: **else**
13: calculate $\triangle_{\mathcal{P}} = \mathcal{B} - r(ij|\mathbf{Z})$, $\mathcal{B}$ follows Equ. (8)
14: **if** $r(ij|\mathbf{Z}) + \triangle_{\mathcal{P}} \geq \alpha$ **then**
15: $\epsilon_t = \epsilon_t + \epsilon$
16: **if** $r(ij|\mathbf{Z}) + Lap(\frac{4\triangle_{\mathcal{P}}}{\epsilon}) \geq \alpha + Lap(\frac{2\triangle_{\mathcal{P}}}{\epsilon})$ **then**
17: delete edge $v_i - v_j$ from $\mathcal{G}$
18: Orientating causal directions of $\mathcal{G}$ by V-structure and consistent propagations.
19: Return $\mathcal{G}$.

Theorem 5. *Causal discovery following Alg. 1 is ϵ-differentially private.*

Proof Sketch. As we directly use SVT to achieve differentially privacy (Lines 7,13), and the rest part follows the PC algorithm and is not related to privacy, Alg. 1 does not change ϵ-differentially private of SVT according to the composition theorem (Dwork and Roth 2014). □

Performance Evaluation

We first compare RCIT with conditional Kendall's τ (Lun, Qi, and Dawn 2020) on testing CIs with differentially private setting, then evaluate their performance on causal discovery with PCD by simulations and real-world datasets. Note that there is no existing CI tests with differential privacy guarantee can handle the nonlinear numerical case, but conditional Kendall's τ can work in the case of the controlling set $\mathbf{Z} = \emptyset$. Our RCIT consists of KRR and HSIC with permutation approximation. We fix the regularization parameter $\lambda = 1$ for KRR, and set the kernel size to median distance between points and $k = 100$ permutations for HSIC. These are normal parameter settings for KRR and HSIC.

Simulation on Conditional Independence Test

Here, we consider the performance of RCIT and conditional Kendall's τ on private CI tests. As CI tests are heavily affected by the sample size of Z, we examine how the probabilities of Type I & II errors of the two methods change with the sample size of Z. We consider the following two cases:

1. (**CI test**) All variables in $\mathbf{Z} = \{Z_1, ..., Z_5\}$ are effective in generating X and Y with the causal structure $X \leftarrow \mathbf{Z} \rightarrow Y$, where $X = sin(\sum Z_i) + c_1 \cdot \varepsilon_x$, $Y = sin(\sum Z_i) + c_2 \cdot \varepsilon_y$, Z_i, ε_x and ε_y are i.i.d. sampled from $U(-0.5, 0.5)$, c_1 and c_2 are two independent coefficients randomly chosen from $U(0.5, 1)$. Type I error rate is evaluated by testing whether X and Y are independent given $\mathbf{Z}$, while Type II error rate is calculated by testing whether X and Y are independent given S where $S \subset \mathbf{Z}$ and $|S| = |\mathbf{Z}| - 1$. Obviously, the ground truth is $X \not\perp\!\!\!\perp Y|S$.
2. (**Independence test**) Type I error rate is calculated by testing whether Z_1 and Z_2 are independent, and Type II error rate is evaluated by testing whether X and Y are independent.

For each experiment setting, we randomly generate $\{200, 500\}$ samples, the privacy parameter ϵ is set to $\{0.1, 0.2, 0.4, 0.8, 1, 2, 4, 8\}$. We repeat the test 100 times and average the corresponding results.

Type I error rate. The results of CI tests are presented in Fig. 1 (a). We simply use "Inf" to denote non-private test. It is clear that Type I error rate of RCIT is heavily affected by ϵ, it decreases as ϵ increases. Similarly, increasing the sample size from 200 to 500 can also reduce the error rate, as large sample size can improve the test performance and tighten the bound of sensitivity of RCIT. We can also see that conditional Kendall's τ fails to handle these cases, as each value for $\mathbf{Z}$ is unique, Kendall's τ cannot calculate the correlation for the case of only one sample.

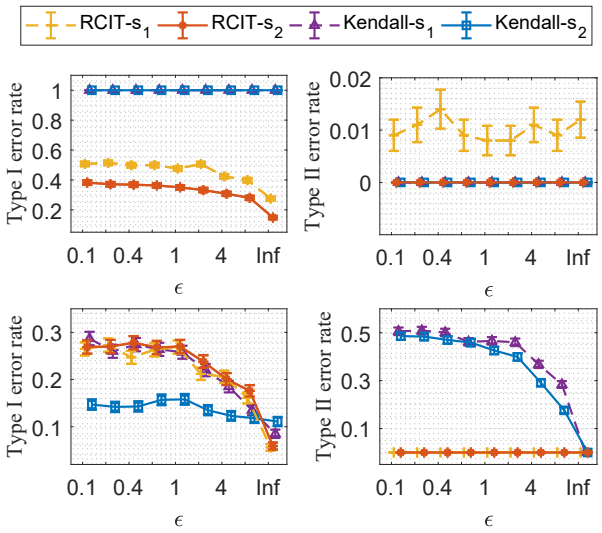

Figure 1: Performance comparison of RCIT and conditional Kendall's τ with sample size $s_1 = 200$, $s_2 = 500$ and privacy parameter $\epsilon = \{0.1, 0.2, 0.4, 0.8, 1, 2, 4, 8\}$. Here, "Inf" denotes non-private test. (a) Type I error rate of CI test $X \perp\!\!\!\perp Y|\mathbf{Z}$; (b) Type II error rate of CI test $X \not\perp\!\!\!\perp Y|S\,(|S| = |\mathbf{Z}| - 1)$; (c) Type I error rate of independence test $Z_1 \perp\!\!\!\perp Z_2$; (d) Type II error rate of independence test $X \not\perp\!\!\!\perp Y$.

The results of independence test are presented in Fig. 1 (c). RCIT and Kendall's τ perform similarly with 200 samples, but when the sample size increases to 500, Kendall's τ outperforms RCIT.

Type II error rate. In contrast to the performance on Type I error rate, Fig. 1 (d) shows that RCIT performs significantly better than Kendall's τ in terms of Type II error rate of independence test. The curve of RCIT stays stable at around 0 as ϵ increases for 200 samples. Different from conditional Kendall's τ, the bound of sensitivity for HSIC is calculated by permutation approximation, the HSIC score in independence test is generally far from the threshold of accepting $\mathcal{H}_0$ by permutation test, and it is hard to exceed this threshold even after subtracting its (small) bound of sensitivity. Therefore, in most cases, the bound of sensitivity of RCIT is very small, so it is hard to generate Type II error.

The results of CI tests are presented in Fig. 1 (b). As aforementioned, conditional Kendall's τ cannot handle CI tests, it directly treats all cases as being not conditional independent, thus Type II error rate is stable at 0. Type II error rate of RCIT is very small, around 0.01 given 200 samples. So there is not much difference between the two methods in CI tests in terms of Type II error rate.

Here, we do not compare the two methods in terms of running time, as HSIC needs to calculate the trace of the product of two kernel matrices and k times of permutations for approximation, the time complexity of RCIT is obviously higher than that of conditional Kendall's τ. Next, we will further demonstrate the advantage of RCIT in private nonlinear causal discovery.

Graph	#Nodes	#Arcs	Max in-degree
Cancer	5	4	2
Asia	8	8	2
Child	20	25	2
Alarm	37	46	4

Table 1: Statistics of four causal graphs.

Performance on Causal Discovery

Here, we compare RCIT and conditional Kendall's τ with PCD (Alg.1) in terms of causal discovery performance. The two methods are evaluated on four causal graphs [1] that cover different applications, including biomedicine (*Cancer* and *Asia*), expert systems (*Child*) and medicine (*Alarm*). The structural statistic data of these causal networks are summarized in Tab. 1. We generate data by following a nonlinear causal model: $x_i = \phi_i(\sum a_{ij} \cdot pa_{ij}) + b_i \cdot \varepsilon_i$, where pa_{ij} is the parent of x_i, ε_i is the noise term sampled from $U(-0.5, 0.5)$, a_{ij} and b_i are two coefficients randomly chosen from $U(0.5, 1)$, ϕ_i is a nonlinear function randomly chosen from *sin*, *cos*, *tanh*, quadratic and exponential functions with probability of 20%. Such a data generation process in widely used in many previous works such as (Zhang et al. 2011). For each experiment setting, we randomly generate $\{1000, 2000, 4000, 8000\}$ samples, repeat the experiments 100 times and average their results. We do not present the error bar as they are very small. Other parameters and settings follows the previous section.

The results are shown in Fig 2. We can see that *Precision* is higher than *Recall* in most cases. Note that $Recall = \frac{Discovered\ edges \cap Actual\ edges}{Actual\ edges}$ and $Precision = \frac{Discovered\ edges \cap Actual\ edges}{Discovered\ edges}$. Type I errors in RCIT and conditional Kendall's τ would not affect structure learning much, that is because CI tests will continue to test x and y given another controlling set Z when Type I error occurs. However, such a traversal search strategy will be greatly affected by Type II error. For example, assume that Type II error rate is r_i for each controlling set Z_i, then the rate of rejecting all CI hypotheses when they are really false is $\prod(1 - r_i)$, and we have $\lim_{k \to +\infty} \prod_{i=1,\ldots,k} (1 - r_i) = 0$. Therefore, the performance of constraint-based causal discovery is largely determined by Type II error rate of CI tests. Compared to conditional Kendall's τ, we can see that the performance of RCIT with PCD is more stable and more advantageous as the sample size increases.

Independence Test on Protein-Signaling Network

The experiments above show that the proposed method works well in private (conditional) independence tests and causal discovery on simulations. Here, we evaluate the proposed method on a well-known real-world causal protein-signaling network, *Sachs* (Sachs et al. 2005). The underlying causal graph of *Sachs* has 11 nodes and 18 arcs, which is usually regarded as the ground truth and widely used in previous works of causal discovery (He et al. 2021).

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

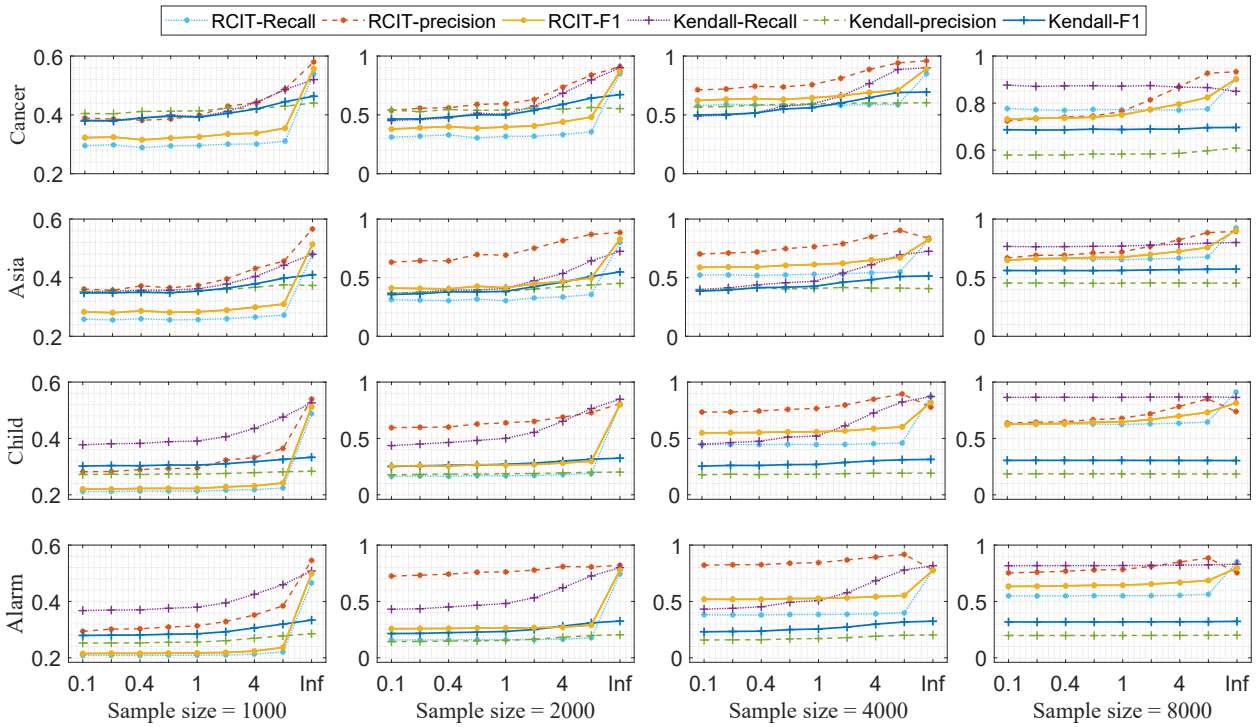

Figure 2: Performance of RCIT and conditional Kendall's τ with PCD on four causal graphs with $\epsilon = \{0.1, 0.2, 0.4, 0.8, 1, 2, 4, 8\}$ and sample size = {1000, 2000, 4000, 8000}, "Inf" denotes non-private test.

The results are shown in Fig. 3, which indicates that RCIT with PCD achieves better performance on this dataset. We can see that the F1 of conditional Kendall's τ is around 0.3 when $\epsilon \leq 1$, which is far below that of RCIT. The low *Recall* indicates that both the two methods are prone to Type II error, which is much more serious than the results of simulations in Fig. 1 and Fig. 2. This indicates that real-world data are more complex and challenging than simulations. Therefore, it is still a tough issue to reduce Type II error rate, especially in real-world scenarios of private causal discovery.

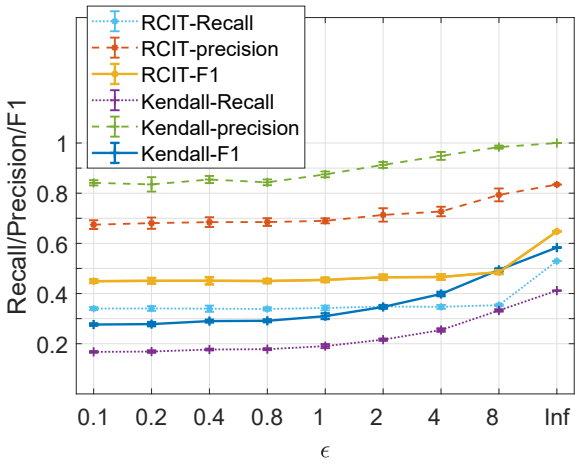

Figure 3: Performance on Sachs dataset.

Conclusion & Future Work

In this paper, we take an significant step towards differentially private causal discovery from nonlinear numerical data, which is an important research topic highlighted in previous studies. Concretely, we propose a method to infer nonlinear causal relations from observed numerical data by using regression-based conditional independence (CI) test (RCIT). To derive a small and feasible bound of sensitivity for private causal inference, we use kernel ridge regression and Hilbert-Schmidt independence criteria to implement RCIT. Furthermore, we develop a private constraint-based causal discovery framework with differential privacy guarantee. We conduct extensive experiments of both CI tests and causal discovery. To the best of our knowledge, there is no existing CI test methods with differential privacy guarantee that can handle nonlinear numerical data. In experiments, we compare our method with the latest Kendall's τ. Experiment results demonstrate the effectiveness and advantage of the proposed method in handling nonlinear numerical data.

However, there are still many challenges in differentially private causal discovery. For example, it is unclear how to reconcile independence tests with infinite/large bound of sensitivity, and it is challenging to derive privacy guarantee for the state-of-the-art causal discovery techniques such as some continuous optimization based methods. These issues can be promising future research topics.

Acknowledgements

This work was supported by National Key Research and Development Program of China (grant No. 2021YFC3340302), and partially by National Natural Science Foundation (NSFC) (U1936205, 61972100 and 62006051). Hao Zhang was also supported by China Postdoctoral Science Foundation (2022M720033).

References

An, S.; Liu, W.; and Venkatesh, S. 2007. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, 40(8): 2154–2162.

Dwork, C.; Mcsherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. *Proceedings of the VLDB Endowment*.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4): 211–407.

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.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *PROCEEDINGS ALGORITHMIC LEARNING THEORY*, 63–77. Springer-Verlag.

He, Y.; Cui, P.; Shen, Z.; Xu, R.; Liu, F.; and Jiang, Y. 2021. DARING: Differentiable Causal Discovery with Residual Independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD -21, 596–605.

Jain, P.; and Thakurta, A. 2013. Differentially Private Learning with Kernels. In Dasgupta, S.; and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 118–126. Atlanta, Georgia, USA: PMLR.

Kendall, M. G. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2): 81–93.

Kusner, M. J.; Sun, Y.; Sridharan, K.; and Weinberger, K. Q. 2016. Private Causal Inference. In Gretton, A.; and Robert, C. C., eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, 1308–1317. Cadiz, Spain.

Lee, S. K.; Gresele, L.; Park, M.; and Muandet, K. 2019. Privacy-Preserving Causal Inference via Inverse Probability Weighting. *arXiv: Learning*.

Lun, W.; Qi, P.; and Dawn, S. 2020. Towards practical differentially private causal graph discovery. In *Advances in Neural Information Processing Systems*, volume 33, 5516–5526.

McDonald, J. H. 2009. *Handbook of biological statistics*. sparky house publishing Baltimore.

Nissim, K.; Raskhodnikova, S.; and Smith, A. 2007. Smooth Sensitivity and Sampling in Private Data Analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, 75–84. New York, NY, USA: Association for Computing Machinery.

Niu, F.; Nori, H.; Quistorff, B.; Caruana, R.; Ngwe, D.; and Kannan, A. 2022. Differentially Private Estimation of Heterogeneous Causal Effects. In *CLeaR 2022*. First Conference on Causal Learning and Reasoning.

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.

Pfister, N.; Bühlmann, P.; Schölkopf, B.; and Peters, J. 2018. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 5–31.

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

Rao, C. R. 2002. *Karl Pearson Chi-Square Test The Dawn of Statistical Inference*. Goodness-of-Fit Tests and Model Validity.

Rindt, D.; Sejdinovic, D.; and Steinsaltz, D. 2020. Consistency of permutation tests for HSIC and dHSIC. *arXiv: Statistics Theory*.

Sachs, K.; Perez, O.; Pe'Er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308.

Spiegelman, D. 2010. Commentary: some remarks on the seminal 1904 paper of Charles Spearman 'The proof and measurement of association between two things'. *International Journal of Epidemiology*, 39(5): 1156–9.

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

Xu, D.; Yuan, S.; and Wu, X. 2017. Differential Privacy Preserving Causal Graph Discovery. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, 60–71.

Zhang, H.; Yan, C.; Xia, Y.; Guan, J.; and Zhou, S. 2022a. Causal Gene Identification Using Non-linear Regression-based Independence Tests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1.

Zhang, H.; Zhou, S.; Guan, J.; and Huan, J. L. 2019. Measuring Conditional Independence by Independent Residuals for Causal Discovery. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5): 1–19.

Zhang, H.; Zhou, S.; Yan, C.; Guan, J.; Wang, X.; Zhang, J.; and Huan, J. 2022b. Learning Causal Structures Based on Divide and Conquer. *IEEE Transactions on Cybernetics*, 52(5): 3232–3243.

Zhang, H.; Zhou, S.; Zhang, K.; and Guan, J. 2017. Causal Discovery Using Regression-Based Conditional Independence Tests. In *AAAI Conference on Artificial Intelligence*.

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.