RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders

Takashi Nicholas Maeda **Shohei Shimizu**
RIKEN Center for Advanced Intelligence Project / Shiga University

Abstract

Causal discovery from data affected by latent confounders is an important and difficult challenge. Causal functional model-based approaches have not been used to present variables whose relationships are affected by latent confounders, while some constraint-based methods can present them. This paper proposes a causal functional model-based method called repetitive causal discovery (RCD) to discover the causal structure of observed variables affected by latent confounders. RCD repeats inferring the causal directions between a small number of observed variables and determines whether the relationships are affected by latent confounders. RCD finally produces a causal graph where a bi-directed arrow indicates the pair of variables that have the same latent confounders, and a directed arrow indicates the causal direction of a pair of variables that are not affected by the same latent confounder. The results of experimental validation using simulated data and real-world data confirmed that RCD is effective in identifying latent confounders and causal directions between observed variables.

1 Introduction

Many scientific questions aim to find the causal relationships between variables rather than only find the correlations. While the most effective measure for identifying the causal relationships is controlled experimentation, such experiments are often too costly, unethical, or technically impossible to conduct. Therefore, the development of methods to identify causal relationships from observational data is important.
Many algorithms that have been developed for constructing causal graphs assume that there are no latent confounders (e.g., PC [Spirtes and Glymour, 1991], GES [Chickering, 2002], and LiNGAM [Shimizu et al., 2006]). They do not work effectively if this assumption is not satisfied. Conversely, FCI [Spirtes et al., 1999] is an algorithm that presents the pairs of variables that have latent confounders. However, since FCI infers causal relations on the basis of the conditional independence in the joint distribution, it cannot distinguish between the two graphs that entail exactly the same sets of conditional independence. Therefore, to understand the causal relationships of variables where latent confounders exist, we need a new method that satisfies the following criteria: (1) the method should accurately (without being biased by latent confounders) identify the causal directions between the observed variables that are not affected by latent confounders, and (2) it should present variables whose relationships are affected by latent confounders. Compared to the constraint-based causal discovery methods (e.g., PC [Spirtes and Glymour, 1991] and FCI [Spirtes et al., 1999]), causal functional model-based approaches [Hoyer et al., 2009, Mooij et al., 2009, Yamada and Sugiyama, 2010, Shimizu et al., 2011, Peters et al., 2014] can identify the entire causal model under proper assumptions. They represent an effect Y as a function of direct cause X. They infer that variable X is the cause of variable Y when X is independent of the residual obtained by the regression of Y on X but not independent of Y. Most of the existing methods based on causal functional models identify the causal structure of multiple observed variables that form a directed acyclic graph (DAG) under the assumption that there is no latent confounder. They assume that the data generation model is acyclic, and that the external effects of all the observed variables are mutually independent. Such models are called additive noise models (ANMs). Their methods discover the causal structures by the following two steps: (1) identifying the causal order of variables and (2) eliminating unnecessary edges. DirectLiNGAM [Shimizu

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

et al., 2011], which is a variant of LiNGAM [Shimizu et al., 2006], performs regression and independence testing to identify the causal order of multiple variables. DirectLiNGAM finds a *root* (a variable that is not affected by other variables) by performing regression and independence testing of each pair of variables. If a variable is exogenous to the other variables, then it is regarded as a root. Thereafter, DirectLiNGAM removes the effect of the root from the other variables and finds the next root in the remaining variables. DirectLiNGAM determines the causal order of variables according to the order of identified roots. RESIT [Peters et al., 2014], a method extended from Mooij et al. [Mooij et al., 2009] identifies the causal order of variables in a similar manner by performing an iterative procedure. In each step, RESIT finds a *sink* (a variable that is not a cause of the other variables). A variable is regarded as a sink when it is endogenous to the other variables. RESIT disregards the identified sinks and finds the next sink in each step. Thus, RESIT finds a causal order of variables. DirectLiNGAM and RESIT then construct a complete DAG, in which each variable pair is connected with the directed edge based on the identified causal order. Thereafter, DirectLiNGAM eliminates unnecessary edges using AdaptiveLasso [Zou, 2006]. RESIT eliminates each edge $X \rightarrow Y$ if X is independent of the residual obtained by the regression of Y on $Z/\{X\}$ where Z is the set of causes of Y in the complete DAG.

Causal functional model-based methods effectively discover the causal structures of observed variables generated by an additive noise model when there is no latent confounder. However, the results obtained by these methods are likely disturbed when there are latent confounders because they cannot find a causal function between variables affected by the same latent confounders. Furthermore, the causal functional model-based approaches have not been used to show variables that are affected by the same latent confounder, as FCI does.

This paper proposes a causal functional model-based method called repetitive causal discovery (RCD) to discover the causal structures of the observed variables that are affected by latent confounders. RCD is aimed at producing causal graphs where a bi-directed arrow indicates the pair of variables that have the same latent confounders, and a directed arrow indicates the direct causal direction between two variables that do not have the same latent confounder. It assumes that the data generation model is linear and acyclic, and that external influences are non-Gaussian. Many causal functional model-based approaches discover causal relations by identifying the causal order of variables and eliminating unnecessary edges. However, RCD discovers the relationships by finding the direct or indirect causes (*ancestors*) of each variable, distinguishing direct causes (*parents*) from indirect causes, and identifying the pairs of variables that have the same latent confounders.

Our contributions can be summarized as follows:

- We developed a causal functional model-based method that can present variable pairs affected by the same latent confounders.
- The method can also identify the causal direction of variable pairs that are not affected by latent confounders.
- The results of experimental validation using simulated data and real-world data confirmed that RCD is effective in identifying latent confounders and causal directions between observed variables.

2 Problem definition

2.1 Data generation process

This study aims to analyze the causal relations of observed variables confounded by unobserved variables. We assume that the relationship between each pair of (observed or unobserved) variables is linear, and that the external influence of each (observed or unobserved) variable is non-Gaussian. In addition, we assume that (observed or unobserved) data are generated from a process represented graphically by a directed acyclic graph (DAG). The generation model is formulated using Equation 1.

$$x_i = \sum_j b_{ij} x_j + \sum_k \lambda_{ik} f_k + e_i \qquad (1)$$

where x_i denotes an observed variable, b_{ij} is the causal strength from x_j to x_i, f_k denotes a latent confounder, λ_{ik} denotes the causal strength from f_k to x_i, and e_i is an external effect. The external effect e_i and the latent confounder f_k are assumed to follow non-Gaussian continuous-valued distributions with zero mean and nonzero variance and are mutually independent. The zero/nonzero pattern of b_{ij} and λ_{ik} corresponds to the absence/existence pattern of directed edges. Without loss of generality [Hoyer et al., 2008], latent confounders f_k are assumed to be mutually independent. In a matrix form, the model is described as Equation 2:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{\Lambda}\mathbf{f} + \mathbf{e} \qquad (2)$$

where the connection strength matrices $\mathbf{B}$ and $\mathbf{\Lambda}$ collect b_{ij} and λ_{ik}, and the vectors $\mathbf{x}$, $\mathbf{f}$ and $\mathbf{e}$ collect x_i, f_k and e_i.

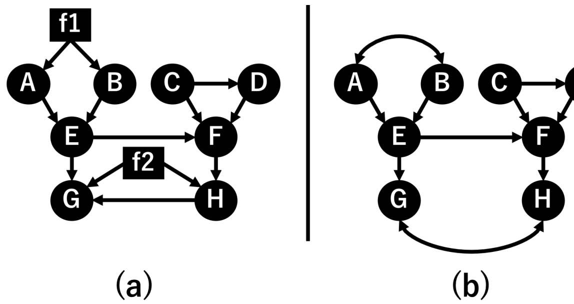

Figure 1: (a) Data generation model (f_1 and f_2 are latent confounders). (b) Causal graph that RCD produces. A bi-directed arrow indicates that two variables are affected by the same latent confounders.

2.2 Research goals

This study has two goals. First, we extract the pairs of observed variables that are affected by the same latent confounders. This is formulated by $\mathbf{C}$ whose element c_{ij} is defined by Equation 3:

$$c_{ij} = \begin{cases} 0 & (\text{if } \forall k, \lambda_{ik} = 0 \vee \lambda_{jk} = 0) \\ 1 & (\text{otherwise}) \end{cases} \tag{3}$$

Element c_{ij} equals 0 when there is no latent confounder affecting variables x_i and x_j. Element c_{ij} equals 1 when variables x_i and x_j are affected by the same latent confounders.
The second goal is to estimate the absence/existence of the causal relations between the observed variables that do not have the same latent confounder. This is defined by a matrix $\mathbf{P}$ whose element p_{ij} is expressed by Equation 4:

$$p_{ij} = \begin{cases} 0 & (\text{if } b_{ij} = 0 \text{ or } c_{ij} = 1) \\ 1 & (\text{otherwise}) \end{cases} \tag{4}$$

$p_{ij} = 0$ when $c_{ij} = 1$ because we do not aim to identify the causal direction between the observed variables that are affected by the same latent confounders.

Finally, RCD produces a causal graph where a bi-directed arrow indicates the pair of variables that have the same latent confounders, and a directed arrow indicates the causal direction of a pair of variables that are not affected by the same latent confounder. For example, assume that using the data generation model shown in Figure 1-(a), our final goal is to draw a causal diagram shown in Figure 1-(b), where variables f_1 and f_2 are latent confounders, and variables A–H are observed variables.

3 Proposed Method

3.1 The framework

RCD involves three steps: (1) It extracts a set of *ancestors* of each variable. *Ancestor* is a direct or indirect cause. In this paper, M_i denotes the set of ancestors of x_i. M_i is initialized as $M_i = \emptyset$. RCD repeats the inference of causal directions between variables and updates M. When inferring the causal directions between observed variables, RCD removes the effect of the already identified common ancestors. Causal direction between variables x_i and x_j can be identified when the set of identified common causes (i.e. $M_i \cap M_j$) satisfies the back-door criterion [Pearl, 1993, Pearl, 2000] to x_i and x_j. The repetition of causal inference is stopped when M no longer changes. (2) RCD extracts *parents* (direct causes) from M. When x_j is an ancestor but not a parent of x_i, the causal effect of x_j on x_i is mediated through $M_i \setminus \{x_k\}$. RCD distinguishes direct causes from indirect causes by inferring conditional independence. (3) RCD finds the pairs of variables that are affected by the same latent confounders by extracting the pairs of variables that remain correlated but whose causal direction is not identified.

3.2 Finding ancestors of each variable

RCD repeats the inference of causal directions between a given number of variables to extract the ancestors of each observed variable. We introduce Lemmas 1 and 2, by which the ancestors of each variable can be identified when there is no latent confounder. Then, we extend them to Lemma 3 by which RCD extracts the ancestors of each observed variable for the case that latent confounders exist. The proofs of Lemmas 1, 2, and 3 are available in Appendices A.1, A.2, and A.3 in [Maeda and Shimizu, 2020]. After the introduction of Lemmas 1–3, we describe how RCD extracts the ancestors of each observed variable.

Lemma 1 Assume that there are variables x_i and x_j, and their causal relation is linear, and their external influences e_i and e_j are non-Gaussian and mutually independent. Let $r_i^{(j)}$ denote the residual obtained by the linear regression of x_i on x_j and $r_j^{(i)}$ denote the residual obtained by the linear regression of x_j on x_i. The causal relation between variables x_i and x_j is determined as follows: (1) If x_i and x_j are not linearly correlated, then there is no causal effect between x_i and x_j. (2) If x_i and x_j are linearly correlated and x_j is independent of residual $r_i^{(j)}$, then x_j is an ancestor of x_i. (3) If x_i and x_j are linearly correlated and x_j is dependent on $r_i^{(j)}$ and x_i is dependent on $r_j^{(i)}$, then x_i and x_j have a common ancestor. (4) There is no case that x_i and x_j are linearly correlated and x_j is

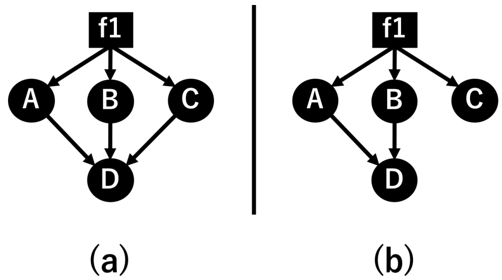

Figure 2: (a) Variables A, B, and C are the causes of variable D, and they have a common cause, f_1. (b) A and B are the causes of D, but C is not.

independent of $r_i^{(j)}$ and x_i is independent of $r_j^{(i)}$.

It is necessary to remove the effect of common causes to infer the causal directions between variables. When the set of the identified common causes of variables x_i and x_j satisfies the back-door criterion, the causal direction between x_i and x_j can be identified. The back-door criterion [Pearl, 1993, Pearl, 2000] is defined as follows:

Definition 1 A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (x_i, x_j) in a DAG G if no node in Z is a descendant of x_i, and Z blocks every path between x_i and x_j that contains an arrow into x_i.

Lemma 1 is generalized to Lemma 2 to incorporate the process of removing the effects of the identified common causes. Lemma 2 can also be used to determine whether the identified common causes are sufficient to detect the causal direction between the two variables.

Lemma 2 Let H_{ij} denote the set of common ancestors of x_i and x_j. Let y_i and y_j denote the residuals when x_i and x_j are regressed on H_{ij}, respectively. Let $r_i^{(j)}$ and $r_j^{(i)}$ denote the residual obtained by the linear regression of y_i on y_j, and y_j on y_i, respectively. The causality and the existence of the confounders are determined by the following criteria: (1) If y_i and y_j are not linearly correlated, then there is no causal effect between x_i and x_j. (2) If y_i and y_j are linearly correlated and y_j is independent of the residual $r_i^{(j)}$, then x_j is an ancestor of x_i. (3) If y_i and y_j are linearly correlated and y_j is dependent on $r_i^{(j)}$ and y_i is dependent on $r_j^{(i)}$, then x_i and x_j have a common ancestor other than H_{ij}, and H_{ij} does not satisfy the back-door criterion to (x_i, x_j) or (x_j, x_i). (4) There is no case that y_i and y_j are linearly correlated and y_j is independent of $r_i^{(j)}$ and y_i is independent of $r_j^{(i)}$.

Next, we consider the case that there are latent confounders. In Lemma 2, the direction between two variables is inferred by regression and independence tests. However, if there are two paths from latent confounder f_k to x_i, and x_j is only on one of the paths, then $M_i \cap M_j$ cannot satisfy the back-door criterion. For example, in Figure 2-(a), variables A, B, and C are the causes of variable D, and the causes are also affected by the same latent confounder f_1. The causal direction between A and D cannot be inferred only by inferring the causality between them because the effect of f_1 is mediated through B and C to D. Therefore, A, B, and C are the causes of D when they are independent of the residual obtained by the multiple regression of D on $\{A, B, C\}$. However, it is necessary to confirm that variables in each proper subset of $\{A, B, C\}$ are not independent of the residual obtained by the regression of D on the proper subset (i.e., no proper subset of $\{A, B, C\}$ satisfies the back-door criterion). For example, in Figure 2-(b), C is not a cause of D, but A, B, and C are all independent of the residual obtained by the multiple regression of D on $\{A, B, C\}$. C should not be regarded as a cause of D because A and B are also independent of the residual when D is regressed on $\{A, B\}$. This example is generalized and formulated by Lemma 3:

Lemma 3 Let X denote the set of all observed variables. Let U denote a subset of X that contains x_i (i.e., $U \subseteq X$ and $x_i \in U$). Let M denote the sequence of M_j where M_j is a set of ancestors of x_j. For each $x_j \in U$, let y_j denote the residual obtained by the multiple linear regression of x_j on the common ancestors of U, where the set of common ancestors of U is $\bigcap_{x_j \in U} M_j$. We define $f(x_i, U, M)$ as a function that returns 1 when each $y_j \in \{y_j \mid x_j \in U \setminus x_i\}$ is independent of the residual obtained by the multiple linear regression of y_i on $\{y_j \mid j \neq i\}$; otherwise it returns 0. If $f(x_i, V, M) = 0$ for each $V \subset U$ and $f(x_i, U, M) = 1$, then each $x_j \in U$ is an ancestor of x_j.

We describe the procedure and the implementation of how RCD extracts the ancestors of each observed variable in Algorithm 1. The output of the algorithm is sequence $M = \{M_i\}$, where M_i is the set of identified ancestors of x_i. Argument α_C is the alpha level for the p-value of the Pearson's correlation. If the p-value of two variables is smaller than α_{C}, then we estimate that the variables are linearly correlated. Argument α_{I} is the alpha level for the p-value of the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al., 2008]. If the p-value of the HSIC of two variables is greater than α_{I}, then we estimate that the variables are mutually independent. Argument α_S is the alpha level to test whether a variable is generated from a non-Gaussian process using the Shapiro-Wilk test [Shapiro and Wilk, 1965]. Argument n is the maximum number of explanatory variables used in multiple linear regression for identifying causal directions; i.e., the maximum number of $(|U| - 1)$ in Lemma 3. In practice, this should be set to a small number when the number of samples is smaller than the number of variables. RCD does not perform multiple regression analysis of more than n explanatory variables.

RCD initializes M_i to be an empty set for each $x_i \in X$. RCD repeats the inference between the variables in each $U \subset X$ that has $(l+1)$ elements. Number l is initialized to 1. If there is no change in M, l is increased by 1. If there is a change in M, l is set to 1. When l exceeds n, the repetition ends. Variable *changed* has information about whether there is a change in M within an iteration.

In line 16 of Algorithm 1, RCD confirms that there is no identified ancestor of x_i in U by checking that $M_i \cap U = \emptyset$. This confirms that $f(x_i, V, M) = 0$ for each $V \subset U$ in Lemma 3. In lines 17–24, RCD checks whether $f(x_i, U, M) = 1$ in Lemma 3. When $f(x_i, U, M) = 1$ is satisfied, x_i is put into S. S is a set of candidates for a *sink* (a variable that is not a cause of the others) in U. It is necessary to test whether there is only one *sink* in U because two variables may be misinterpreted as causes of each other when the alpha level for the independence test (α_I) is too small.

We use least squares regression for removing the effect of common causes in line 12 of Algorithm 1, but we use a variant of multiple linear regression called multilinear HSIC regression (MLHSICR) to examine the causal directions between variables in U in line 20 of Algorithm 1 when $l \geq 2$. Coefficients obtained by multiple linear regression using the ordinary least squares method with linearly correlated explanatory variables often differ from true values due to estimation errors. Thus, the relationship between the explanatory variables and the residual may be misinterpreted to be dependent in the case that explanatory variables are affected by the same latent confounders. To avoid such failure, we use MLHSICR defined as follows:

Definition 2 Let variable x_i denote an explanatory variable, $\mathbf{x}$ denote a vector that collects explanatory variables x_i, and y denote a response variable. MLHSICR models the relationship $y = \boldsymbol{\lambda}^\top \mathbf{x}$ by the coefficient vector $\boldsymbol{\lambda}$ in the following equation:

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \sum_i \widehat{\mathrm{HSIC}}(x_i, y - \boldsymbol{\lambda}^\top \mathbf{x}) \qquad (5)$$

where $\widehat{\mathrm{HSIC}}(a, b)$ denotes the Hilbert-Schmidt independence criterion of a and b.

Mooij et al. [Mooij et al., 2009] have developed a method to estimate the nonlinear causal function between variables by minimizing the HSIC between the explanatory variables and the residual. RCD estimates $\boldsymbol{\lambda}$ by minimizing the sum of the HSICs in Equation 5 using the L-BFGS method [Liu and Nocedal, 1989], similar to Mooij et al. [Mooij et al., 2009]. L-BFGS is a quasi-Newton method, and RCD sets the coefficients obtained by the least squares method to the initial value of $\boldsymbol{\lambda}$.

3.3 Finding parents of each variable

When x_j is an ancestor but not a parent of x_i, the effect of x_j on x_i is mediated through $M_i \setminus \{x_j\}$. Therefore, $x_j \perp\!\!\!\perp x_i \mid M_i \setminus \{x_j\}$. Zhang et al. [Zhang et al., 2017] proposed a method to test the conditional independence using unconditional independence testing in Theorem 1 (proved by them):

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 $x_i - f(Z) \perp\!\!\!\perp x_j - g(Z)$, and $x_i - f(Z) \perp\!\!\!\perp Z$ or $x_j - g(Z) \perp\!\!\!\perp Z$.

In our case, $x_j \perp\!\!\!\perp x_i \mid (M_i \setminus \{x_j\}) \Leftrightarrow x_j - f(M_i \setminus \{x_j\}) \perp\!\!\!\perp x_i - g(M_i \setminus \{x_j\})$, where f and g are multiple linear regression functions of x_j on $M_i \setminus \{x_j\}$ and x_i on $M_i \setminus \{x_j\}$, respectively. Since $(M_i \setminus \{x_j\}) \cap M_j = M_i \cap M_j$, we can assume that $x_j \perp\!\!\!\perp x_i \mid (M_i \setminus \{x_j\}) \Leftrightarrow x_j - h(M_i \cap M_j) \perp\!\!\!\perp x_i - g(M_i \setminus \{x_j\})$ where h is a multiple linear regression function of x_j on $(M_i \cap M_j)$.

Based on Theorem 1, RCD uses Lemma 4 to distinguish the parents from the ancestors. Lemma 4 is proved without using Theorem 1, and the proof is presented in Appendix A.4 in [Maeda and Shimizu, 2020].

Lemma 4 Assume that $x_j \in M_i$; that is, x_j is an ancestor of x_i. Let z_i denote the residual obtained by the multiple regression of x_i on $M_i \setminus \{x_j\}$. Let w_j denote the residual obtained by the multiple regression of x_j on $(M_i \cap M_j)$. If z_i and w_j are linearly correlated, then x_j is a parent of x_i; otherwise, x_j is not a parent of x_i.

3.4 Identifying pairs of variables that have the same latent confounders

RCD infers that two variables are affected by the same latent confounders when those two variables are linearly correlated even after removing the effects of all the parents. RCD identifies the pairs of variables affected by the same latent confounders by using Lemma 5. The proof of Lemma 5 is available in Appendix A.5 in [Maeda and Shimizu, 2020].

Lemma 5 Let M_i and M_j respectively denote the sets of ancestors of x_i and x_j, and P_i and P_j respectively denote the sets of parents of x_i and x_j. Assume that $x_i \notin M_j$ and $x_j \notin M_i$. Let y_i denote the residual obtained by the multiple regression of x_i on P_i, and y_j denote the residual obtained by the multiple regression of x_j on P_j. If y_i and y_j are linearly correlated, then x_i and x_j have the same latent confounders.

Algorithm 1: Extract ancestors of each variable

Input: X: the set of observed variables, α_{C}: the alpha level for Pearson's correlation, α_{I}: the alpha level for independence test, α_{S}: the alpha level for Shapiro-Wilk test, n: the maximum number of explanatory variables

Output: M: the sequence $\{M_i\}$ where M_i is a set of ancestors of x_i.

```
function extractAncestors(X, α_C, α_I, α_S, n)
    initialization
        foreach i do
            M_i ← ∅
        l ← 1
    while l ≤ n do
        changed ← FALSE
        foreach U ⊆ X; (|U| = l + 1) do
            H_U ← ⋃_{x_j ∈ U} M_j
            S ← ∅
            foreach x_j ∈ U do
                y_j ← the residual obtained by regression of x_j on H_U
                t_j ← the p-value of Shapiro-Wilk test of y_j
            if ∀t_k < α_S then
                foreach x_i ∈ U do
                    if M_i ∩ U = ∅ then
                        foreach x_j ∈ U \ {x_i} do
                            c_ij ← the p-value of linear correlation between y_i and y_j
                        if ∀c_ij < α_C then
                            s_i^U ← the residual obtained by regression of y_i on {y_j | x_j ∈ U \ {x_i}}
                            foreach x_j ∈ U \ {x_i} do
                                h_ij ← the p-value of the HSIC between s_i^U and y_j
                            if ∀h_ij > α_I then
                                S ← S ∪ {x_i}
                if |S| = 1 then
                    foreach x_i ∈ S do
                        M_i ← M_i ∪ (U \ {x_i})
                    changed ← TRUE
        if changed = TRUE then
            l ← 1
        else
            l ← l + 1
    return M
```

4 Performance evaluation

We evaluated the performance of RCD relative to the existing methods in terms of how accurately it finds the pairs of variables that are affected by the same latent confounders and how accurately it infers the causal directions of the pairs of variables that are not affected by the same latent confounder. In regard to the latent confounders, we compared RCD with FCI [Spirtes et al., 1999], RFCI [Colombo et al., 2012], and GFCI [Ogarrio et al., 2016]. In addition to these three methods, we compared RCD with PC [Spirtes and Glymour, 1991], GES [Chickering, 2002], DirectLiNGAM [Shimizu et al., 2011], and RESIT [Peters et al., 2014] to evaluate the accuracy of causal directions. In the following sections, DirectLiNGAM is called LiNGAM for simplicity.

4.1 Performance on simulated structures

We performed 100 experiments to evaluate RCD relative to the existing methods. We prepared 300 sets of samples for each experiment. The data of each experiment were generated as follows: The data generation process was modeled the same as Equation 1. The number of observed variables x_i was set to 20 and the number of latent confounders f_k was set to 4. Let X and Y denote the stochastic variables, and assume that $Y \sim N(0.0, 0.5)$ and $X = Y^3$. We used the random samples of X for e_i and f_k because X is non-Gaussian. The number of causal arrows between the observed variables is 40, and the start point and the end point of each causal arrow were randomly selected. We randomly drew two causal arrows from each latent confounder to the observed variables. Let Z denote a stochastic variable that comes from a uniform distribution on $[-1.0, -0.5]$ and $[0.5, 1.0]$. We used the random samples of Z for b_{ij} and λ_{ik}.

We evaluated (1) how accurately each method infers the pairs of variables that are affected by the same latent confounders (called the evaluation of latent confounders), and (2) how accurately each method infers causality between the observed variables that are not affected by the same latent confounder (called the evaluation of causality). The evaluation of latent confounders corresponds to the evaluation of bi-directed arrows in a causal graph, and the evaluation of causality corresponds to the evaluation of directed arrows. We used precision, recall, and F-measure as evaluation measures. In regard to the evaluation of latent confounders, true positive (TP) is the number of true bi-directed arrows that are correctly inferred. In regard to causality, TP is the number of true directed arrows that a method correctly infers in terms of their positions and directions. Precision is TP divided by the number of estimations, and recall is TP divided by

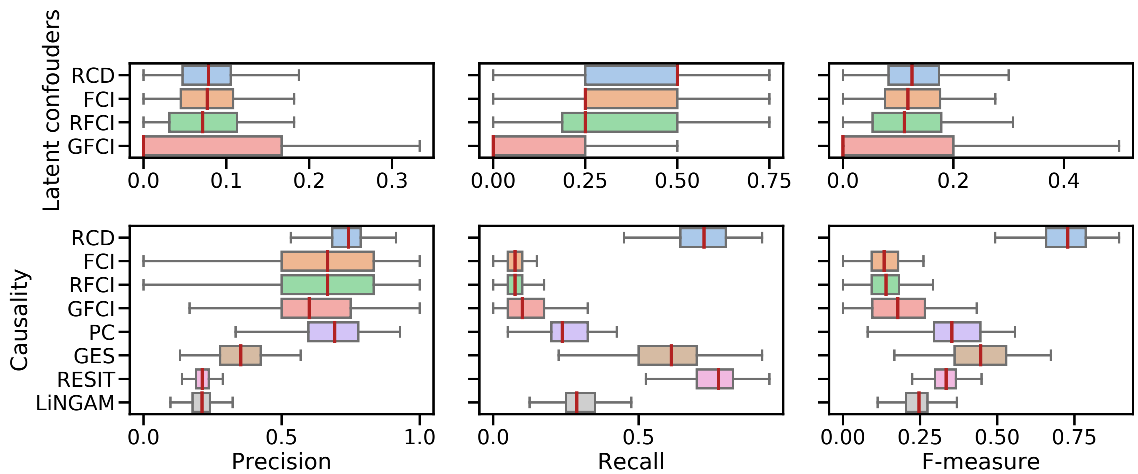

Figure 3: Performance evaluation on causal graphs using simulated data: The vertical red lines indicate the median values of the results. The evaluation of the latent confounders corresponds to the evaluation of bi-directed arrows. The evaluation of causality corresponds to the evaluation of directed arrows.

the number of all true arrows. F-measure is defined as F-measure = 2 · precision · recall/(precision + recall).

The arguments of RCD, that is, α_{C} (alpha level for Pearson's correlation), α_{I} (alpha level for independence), α_{S} (alpha level for the Shapiro-Wilk test), and n (maximum number of explanatory variables for multiple linear regression) were set as $\alpha_C = 0.01, \alpha_I = 0.01, \alpha_S = 0.01$, and $n = 2$.

In regard to the types of edges, FCI, RFCI, and GFCI produce partial ancestral graphs (PAGs) that include six types of edges: $\rightarrow$ (directed), $\leftrightarrow$ (bi-directed), $\circ\!\!\rightarrow$ (partially directed), $\circ\!\!-\!\!\circ$ (nondirected), and $\circ\!\!-$ (partially undirected). In the evaluation, we only used the directed and bi-directed edges. PC, GES, LiNGAM, and RESIT produce causal graphs only with the directed edges; thus, we did not evaluate those methods in terms of latent confounders.

The box plots in Figure 3 display the results. The vertical red lines indicate the median values. Note that some median values are the same as the upper or lower quartiles. For example, the median and the upper quartile of the recalls of RCD in the results of latent confounders are the same. It means that the results between the median and the upper quartile are the same. In regard to the evaluation of latent confounders, the precision, recall, and F-measure values are almost the same for RCD, FCI, RFCI, and GFCI, but the medians of precision, recall, and F-measure values of RCD are the highest among them. In regard to causality, RCD scores the highest medians of the precision and F-measure values among all the methods, and the median of recall for RCD is the second highest next to RESIT.

The results suggest that RCD does not greatly improve the performance metrics compared to the existing methods. However, there is no other method that has the highest or the second highest performance for each metric. FCI, RFCI, and GFCI perform as well as RCD in terms of finding the pairs of variables that are affected by the same latent confounders, but they do not perform well in terms of the recall of causality. In addition, no other method performs well in terms of both precision and recall of causality. RCD can successfully find the pairs of variables that are affected by the same latent confounders and identify the causal direction between variables that are not affected by the same latent confounder.

4.2 Performance on real-world structures

Causal structures in the real-world are often very complex. Therefore, RCD likely produces a causal graph where each pair of observed variables is connected with a bi-directed arrow. The result of identifying latent confounders is affected by the threshold of the p-value for the independence test, α_I. If α_I is too large or too small, then all the variable pairs are likely concluded to have the same latent confounders. Therefore, we need to find the most appropriate value of α_I. We increased k from 1 to 25 and set α_I as $\alpha_I = 0.1^k$ and repeated the process. We adopted a result that has the smallest number of pairs of variables with the same latent confounders.

We analyzed the General Social Survey data set, taken from a sociological data repository.[1] The data have been used for the evaluation of DirectLiNGAM in

[1] http://www.norc.org/GSS+Website/

Table 1: The results of the application to sociological data.

Method	Bidirected arrows (Latent confounders)			Directed arrows (Causality)		
	# of estimation	# of successes	Precision	# of estimation	# of successes	Precision
RCD	4	4	1.0	5	4	0.8
FCI	3	3	1.0	3	1	0.3
RFCI	3	3	1.0	3	1	0.3
GFCI	0	0	0.0	0	0	0.0
PC	-	-	-	2	1	0.5
GES	-	-	-	2	1	0.5
RESIT	-	-	-	12	4	0.3
LiNGAM	-	-	-	5	4	0.8

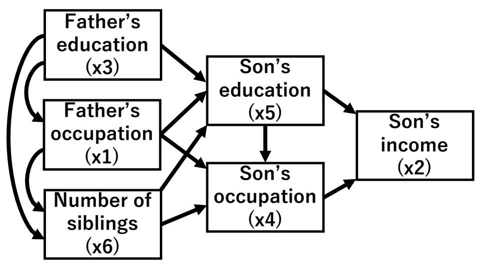

Figure 4: Variables and causal relations in the General Social Survey data set used for the evaluation.

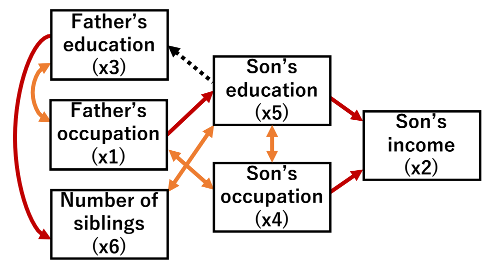

Figure 5: Causal graph produced by RCD: The dashed arrow, $x_3 \leftarrow x_5$ is incorrect inference, but the other arrows are reasonable based on Figure 4

Shimizu et al. [Shimizu et al., 2011]. The sample size is 1380. The variables and the possible directions are shown in Figure 4. The directions were determined based on the domain knowledge in Duncan et al. [Duncan et al., 1972] and temporal orders.

We evaluated the directed arrows (causality) in the causal graphs produced by RCD and the existing methods, based on the directed arrows in Figure 4. In addition, we evaluated the bi-directed arrows in causal graphs produced by the methods as accurate inference if they exist in Figure 4 as directed arrows.

The results are listed in Table 1. In regard to bi-directed arrows (latent confounders), the number of successful inferences by RCD is the highest, and the precisions of RCD, FCI, and RFCI are all 1.0. In regard to the directed arrows (causality), the numbers of the successful arrows of RCD, RESIT, and LiNGAM are the highest. The precisions of RCD and LiNGAM are also the highest. The causal graph produced by RCD is shown in Figure 5. The dashed arrow $x_3 \leftarrow x_5$ is the incorrect inference, but the others are correct.

RCD performs the best among the existing methods in terms of both identifying the pairs of variables that are affected by the same latent confounders and identifying the causal direction of the pairs of variables that are not affected by the same latent confounder.

5 Conclusion

We developed a method called repetitive causal discovery (RCD) that produces a causal graph where a directed arrow indicates the causal direction between the observed variables, and a bi-directed arrow indicates a pair of variables have the same confounder. RCD produces a causal graph by (1) finding the ancestors of each variable, (2) distinguishing the parents from the indirect causes, and (3) identifying the pairs of variables that have the same latent confounders. We confirmed that RCD effectively analyzes data confounded by unobserved variables through validations using simulated and real-world data.

In this paper, we did not discuss the utilization of prior knowledge. However, it is possible to make use of prior knowledge of causal relations in practical applications of RCD. In this study, information about the ancestors of each variable was initialized to be an empty set. If we have prior knowledge about causal relations, the information about the ancestors of each variable that RCD retains can be set according to the prior knowledge.

There is still room for improvement in the RCD method. The optimal settings of the arguments of RCD and the extension of RCD for nonlinear causal relations will be investigated in future studies.

Acknowledgements

Shohei Shimizu was partially supported by ONRG NICOP N62909-17-1-2034 and Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science (JSPS) #16K00045.

References

[Chickering, 2002] Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

[Colombo et al., 2012] Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1):294–321.

[Duncan et al., 1972] Duncan, O. D., Featherman, D. L., and Duncan, B. (1972). *Socioeconomic background and achievement.* New York, NY, USA.

[Gretton et al., 2008] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. Curran Associates, Inc.

[Hoyer et al., 2009] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc.

[Hoyer et al., 2008] Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362 – 378. Special Section on Probabilistic Rough Sets and Special Section on PGM'06.

[Liu and Nocedal, 1989] Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

[Maeda and Shimizu, 2020] Maeda, T. N. and Shimizu, S. (2020). Causal discovery of linear non-Gaussian acyclic models in the presence of latent confounders.

[Mooij et al., 2009] 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*, ICML '09, pages 745–752, New York, NY, USA. ACM.

[Ogarrio et al., 2016] Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379.

[Pearl, 1993] Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

[Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning and inference.* Cambridge University Press.

[Peters et al., 2014] Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

[Shapiro and Wilk, 1965] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[Shimizu et al., 2006] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

[Shimizu et al., 2011] 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. *Journal of Machine Learning Research*, 12(Apr):1225–1248.

[Spirtes and Glymour, 1991] Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

[Spirtes et al., 1999] Spirtes, P., Meek, C., and Richardson, T. (1999). Causal discovery in the presence of latent variables and selection bias. In Cooper, G. F. and Glymour, C. N., editors, *Computation, Causality, and Discovery*, pages 211–252. AAAI Press.

[Yamada and Sugiyama, 2010] Yamada, M. and Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[Zhang et al., 2017] Zhang, H., Zhou, S., Zhang, K., and Guan, J. (2017). Causal discovery using regression-based conditional independence tests. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.