Recursively Learning Causal Structures Using Regression-Based Conditional Independence Test

Hao Zhang,[1] **Shuigeng Zhou,**[1*] **Chuanxu Yan,**[1] **Jihong Guan,**[2] **Xin Wang**[3]
[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
[3]University of Calgary, Canada, and Northwest University, China
[1]{haoz15, sgzhou,17110240047}@fudan.edu.cn; [2]jhguan@tongji.edu.cn; [3]xcwang@ucalgary.ca

Abstract

This paper addresses two important issues in causality inference. One is how to reduce redundant conditional independence (CI) tests, which heavily impact the efficiency and accuracy of existing constraint-based methods. Another is how to construct the true causal graph from a set of Markov equivalence classes returned by these methods.

For the first issue, we design a recursive decomposition approach where the original data (a set of variables) is first decomposed into three small subsets, each of which is then recursively decomposed into three smaller subsets until none of subsets can be decomposed further. Consequently, redundant CI tests can be reduced by inferring causality from these subsets. Advantage of this decomposition scheme lies in two aspects: 1) it requires only low-order CI tests, and 2) it does not violate d-separation. Thus, the complete causality can be reconstructed by merging all the partial results of the subsets.

For the second issue, we employ regression-based conditional independence test to check CIs in linear non-Gaussian additive noise cases, which can identify more causal directions by $x - E(x|Z) \perp z$ (or $y - E(y|Z) \perp z$). Therefore, causal direction learning is no longer limited by the number of returned V-structures and the consistent propagation.

Extensive experiments show that the proposed method can not only substantially reduce redundant CI tests but also effectively distinguish the equivalence classes, thus is superior to the state of the art constraint-based methods in causality inference.

Introduction

Inferring causal relationships between variables from observed data is a challenging task if no controlled experiment is available. From computational perspective, causal discovery is usually formulated as a graphical probabilistic model on the variables, such that directed edges between variables indicate causal relationships. When it is difficult to manipulate the samples in experiments, conditional independence (CI) tests (Fukumizu et al. 2007) are commonly employed in constraint-based methods to detect local causalities among the variables (Edwards 2012; Gao and Ji 2015), under the faithfulness assumption (Koller and Friedman 2009). To recover causal graphs, we often check CIs between variables. For example, let X, Y and Z denote sets of variables, if X and Y are independent given the controlling set Z (i.e., X and Y are d-separated by Z), denoted by $X \perp Y|Z$, then we can deduce that X and Y have no directed causality. In practice, the CI relationship $X \perp Y|Z$ allows us to separate $X - Y$ when constructing a probabilistic model for $P(X, Y, Z)$, which results in a parsimonious representation. Generally speaking, by using CI test, existing causal discovery methods like the PC algorithm (Spirtes, Glymour, and Scheines 2000) can determine a partially directed acyclic graph (PDAG) representing the equivalence classes.

In the constraint-based methods, a tough problem for causality discovery is the search for d-separators (Pearl 2009; Cai, Zhang, and Hao 2013), which becomes exponentially complicated with the number of variables (Bergsma 2004). Specifically, we face two challenges: one is that the number of candidate controlling sets grows exponentially with the number of variables, and exhaustive search for d-separators becomes prohibitively expensive; another challenge is that CI tests tend to be unreliable when the size of conditional set Z gets large, and easily fall into Type II errors (Zhang et al. 2011; Doran et al. 2014; Zhang et al. 2017; Strobl, Zhang, and Visweswaran 2017), i.e., the CI hypothesis is accepted even though it is not true.

To overcome the difficulties mentioned above, researchers resorted to recursive approaches (Geng, Wang, and Zhao 2005; Xie, Geng, and Zhao 2006; Xie and Geng 2008; Cai, Zhang, and Hao 2017; Liu et al. 2017). These methods aim to split a variable set recursively into two or more subsets, such that each subset corresponds to a subproblem that can be solved efficiently by using existing methods, finally the original problem is solved by merging all the results of the subproblems. For example, as shown in Fig. 1(a), given a set of variables V, generally we can reconstruct the corresponding causal graph by detecting the set CI_V of all CIs among V (here a CI indicates two variables that are conditional independent, and we do not care the controlling set). Alternatively, we can first decompose V into m small subsets $V_1, ..., V_m$ by discovering a set of CIs (say CI_0). The CIs of each subset can be further discovered separately, denoted by CI_1 , ..., CI_m, respectively. By combining the $m + 1$ sets of CIs, we can also recover the causal graph if $\bigcup_{i=0}^{m} CI_i = CI_V$, i.e., the splitting process does not violate d-separation. In contrast to learning causal graph by directly using constraint-

[*] Correspondence author.

based methods, such a split-and-merge strategy can avoid some redundant CI tests, is therefore faster and more accurate. However, designing such a decomposition scheme is a nontrivial task, and the problems of inefficiency and violating d-separation are still tough challenges to the existing recursive methods.

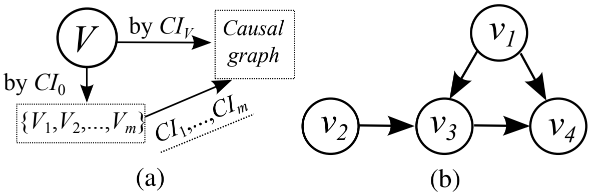


Figure 1: (a) An example applying the split-and-merge strategy on V=$\{V_1, V_2, ..., V_m\}$; (b) An example of variable set splitting that violates d-separation: $\{V_1$=$\{v_1, v_3, v_4\}, V_2$=$\{v_2, v_3, v_4\}\}$.

In this paper, we present a new recursive framework (called **CAPA**, the abbreviation of ***CA****usality* ***PA****rtitioning*) to support effective and scalable causality discovery. There are two major contributions in our work: 1) We propose a novel decomposition scheme that does not violate d-separation and requires only low-order CI tests, which can therefore reduce the redundant CI tests as many as possible; 2) We employ regression-based conditional independence test to check CIs in linear non-Gaussian additive noise cases, i.e., to test $x \perp y|Z$ by using $x-E(x|Z) \perp y-E(y|Z)$. We show that more causal directions can be identified by $x-E(x|Z) \perp z$ (or $y-E(y|Z) \perp z$). Therefore, the task of direction learning is no longer limited by the number of returned V-structures and the corresponding consistent propagation.

Related Work

In the previous works, a common solution of decomposing is to find a decomposition V=$\{A, B, C\}$ where the CI of A and B given C holds, then V is split into two subsets V_1=$A \cup C$ and V_2=$B \cup C$. Such a decomposing process is recursively applied to each subset. Following this idea, existing works, including (Geng, Wang, and Zhao 2005), (Xie, Geng, and Zhao 2006), (Xie and Geng 2008) and (Liu et al. 2017) proposed different recursive decomposition algorithms for causality discovery.

The method proposed in (Xie and Geng 2008) first reconstructs an undirected independence graph (UIG) by removing the edge between every two variables $x, y \in V$ if $x \perp y|V\backslash_{x,y}$. Then, the original variable set V is split into two small subsets by finding a decomposition V=$\{A, B, C\}$ in the UIG, where A and B are d-separated by C. This procedure is applied recursively to each subset till none of subsets can be decomposed further. The subproblems are solved by using some specific constraint-based method, and finally the original problem is solved by merging all the results of the subproblems. This method is more efficient than those proposed in (Geng, Wang, and Zhao 2005) and (Xie, Geng, and Zhao 2006), because the method proposed in (Geng, Wang, and Zhao 2005) requires that each separator has a complete undirected graph, while the method in (Xie, Geng, and Zhao 2006) removes this condition, but it performs decomposition only based on the entire UIG of V, and cannot decompose undirected independence subgraphs. Recently, Liu et al. (Liu et al. 2017) proposed a novel recursive method based on UIGs. Different from the methods above, they innovatively combined the score and search based methods to solve the problem and achieved the state-of-the-art performance.

In practice, it is expensive to construct an accurate UIG, due to requiring high-order CI tests. To circumvent this problem, (Cai, Zhang, and Hao 2017) proposed a recursive method called SADA, which is able to find a decomposition V=$\{A, B, C\}$ by using only a high-order CI test and some lower-order CI tests in each iteration. SADA consists of two major steps: 1) Selects two variables $x, y \in V$ such that $x \perp y|V\backslash_{x,y}$, and finds a minimal d-separator C (Tian, Pearl, and Paz 1998) of x and y; 2) Let V_1=x, V_2=y and V=$V\backslash_{x,y,C}$, consider $\forall w \in V$, add w into V_2 if $\forall u \in V_1$ and $\exists \tilde{C} \subseteq C$ such that $u \perp w|\tilde{C}$; add w into V_1 if $\forall v \in V_2$ and $\exists \tilde{C} \subseteq C$ such that $v \perp w|\tilde{C}$; otherwise, add w into C. Finally, we have a decomposition V=$V_1 \cup V_2$ where V_1=$V_1 \cup C$ and V_2=$V_2 \cup C$. However, there are two drawbacks in SADA: 1) The d-separation is violated. An example is shown in Fig. 1(b), where the partitioning V=$\{V_1$=$\{v_1, v_3, v_4\}, V_2$=$\{v_2, v_3, v_4\}\}$ is returned by SADA. We can see that two non-adjacent variables v_2 and v_4 are d-separable neither in V_1 nor in V_2, since the only d-separator $v_1 \cup v_3$ is divided into two different subsets V_1 and V_2; 2) The separator C is generated randomly. The only way to downsize C is using enumeration (Cai, Zhang, and Hao 2017), which requires many CI tests.

Another tough issue with these methods is that they infer causal direction by checking V-structures and doing consistent propagation, so they cannot distinguish Markov equivalence classes (Chickering 2002). That is, they cannot distinguish $x \rightarrow y \rightarrow z$, $x \leftarrow y \leftarrow z$ and $x \leftarrow y \rightarrow z$. In addition, unreliable local structures returned by these methods negatively impact the performance of consistent propagation.

Recently, regression and residual-based methods were proposed for CI testing, which can distinguish equivalence classes. (Grosse-Wentrup et al. 2016) proved that if there exists a function f such that $x-f(Z) \perp (y, Z)$, then $x \perp y|Z$. (Zhang et al. 2017) showed that if there exists two functions f and g such that $x-f(Z) \perp (y - g(Z), Z)$, then $x \perp y|Z$. These methods find the function f (or g) by regressing x (or y) on Z, so are able to relax a CI test to a set of marginal independence tests. Actually, these methods can determine causal directions as $x-E(x|Z) \perp (\cdot, Z) \Rightarrow Z$ causes x in many cases (Zhang and Hyvärinen 2009). However, due to the high computational complexity of measuring dependence between a variable and a joint distribution, these methods are not suitable for high dimensional cases.

In (Ramsey 2014), the authors suggested to use a simpler form $x-E(x|Z) \perp y-E(y|Z)$ to test $x \perp y|Z$ under the faithfulness and additive noise assumptions. In (Zhang et al. 2017), the authors conjectured that $x-f(Z) \perp y-g(Z)$ can lead to $x \perp y|Z$ under the faithfulness condition, where f and g are arbitrary nonlinear functions, x, y and Z are generated by following the nonlinear additive noise model (ANM) (Zhang and Hyvärinen 2009; Peters, Janzing, and Schölkopf 2011). (Flaxman, Neill, and Smola 2016) showed that given the additive noise, faithfulness and Markov assumptions (Pearl 2009), whenever Z causes x or y, it follows that $x \perp y | Z$ if and only if $x - E(x|Z) \perp y - E(y|Z)$. Note that a strong precondition that Z causes x or y is assumed here. If these conditions are given, then it is easy to derive the corresponding causalities. (Zhang, Zhou, and Guan 2018) proved that regression-based conditional independence test (ReCIT) in linear Gaussian and non-Gaussian cases, $x - E(x|Z) \perp y - E(y|Z) \Rightarrow x \perp y | Z$ under the faithfulness and Markov assumptions. Although they showed that $x - E(x|Z) \perp y - E(y|Z)$ can lead to $x - E(x|Z) \perp z$ or $y - E(y|Z) \perp z$ ($\forall z \in Z$) in many cases, it is not true in all situations. For example, consider a causal graph of $a \to b \to c \to d \to e$ and $a \to e$, we can see that $c \perp e | (b, d)$ but $c - E(c|b, d) \not\perp b$ (d) and $e - E(e|b, d) \not\perp b$ (d). Therefore, it is still a challenge to draw causal directions by ReCITs, especially in high-dimensional cases.

The CAPA Method

The Framework

We first give a formal definition of the new decomposition used in our method (called *causal partitioning*, **CAPA** in short), and then present the framework of the CAPA method.

Definition 1 *Let $G=(V, E)$ denote a DAG on a variable set V, we say a group of variable sets $S = \{V_1, V_2, ..., V_m\}$ constitute a* ***causal partitioning*** *over V iff 1) $\bigcup_{i=1}^{m} V_i = V$; 2) $\forall u, v \in V$, if $\exists V_i, V_j \subset S$ such that $u \in V_i$, $v \notin V_i$, $u \notin V_j$ and $v \in V_j$, then u and v are non-adjacent in G; 3) if $\forall u, v \in V$ are non-adjacent in G, then $\exists V_i, V_j \subset S$ such that $u \in V_i$, $v \notin V_i$, $u \notin V_j$ and $v \in V_j$, or $\exists V_k \subset S$ such that u and v are d-separable in V_k.*

The three conditions in Definition 1 imply that any adjacent variable cannot be separated in causal partitioning, and any non-adjacent variable is either separated in the process of causal partitioning or d-separable in at least one subset V_k. This means that d-separation is not violated in the causal partitioning process. Therefore, the causality discovery problem on V can be transformed into m smaller causality discovery subproblems over the m variable sets $V_1, ..., V_m$, respectively. The details of finding causal partitionings will be discussed in the next subsection, here we first present the framework of the CAPA method, which is presented as Alg. 1.

The inputs of CAPA include the original variable set V and an user specified constraint-based algorithm A_g (say the PC algorithm) for discovering causality from the resulting subsets. The major sub-procedure in Alg. 1 is to find a causal partitioning (Line 3). We can see that if a subset cannot be further partitioned, the structure of this subset will be reconstructed by the algorithm A_g, otherwise it will be further partitioned into three smaller subsets. In what follows, we give the details of finding causal partitioning.

Finding Causal Partitioning

The search of causal partitionings is crucial to the CAPA method. To identify potential causal partitionings, our algorithm resorts to CI tests between input variables. The process of finding causal partitioning includes the following 3 steps, and the pseudo-code is outlined in Alg. 2.

Algorithm 1 CAPA

```
 1: Input: The original variable set $V$, algorithm $A_g$.
 2: Output: The causal graph $G$.
 3: Find a causal partitioning $\{V_1, V_2, V_3\}$ on $V$.
 4: if $\max\{|V_1|, |V_2|, |V_3|\} = |V|$ then
 5:     Return $G$ by running algorithm $A_g$ on $V$.
 6: else
 7:     $G_1$=CAPA($V_1, A_g, \delta$),
 8:     $G_2$=CAPA($V_2, A_g, \delta$),
 9:     $G_3$=CAPA($V_3, A_g, \delta$).
10:     Return $G$ by merging $G_1$, $G_2$ and $G_3$.
11: end if
```

Step 1. We construct the 0-order CI table M (an adjacent matrix) of the input set $V = \{v_1, ..., v_n\}$ where $M_{i,j} = 1$ indicates v_i and v_j are marginal independent, i.e., $v_i \perp v_j$; and $M_{i,j} = 0$ means v_i and v_j are dependent, i.e., $v_i \not\perp v_j$. The entries of M are calculated by marginal independence tests (Line 3).

Step 2. We partition V into three non-overlapping subsets $\{A, B, C = V \backslash_{A,B}\}$ according to $\forall v_i \in A, \forall v_j \in B, M_{ij} = 1$, and simultaneously minimize the size of C (Line 4). Intuitively, C blocks all the links between A and B, but note that C is not a d-separator regarding A and B. Then, we remove V's maximum subset independent of C, and let $V_1 = A \cup C$, $V_2 = B \cup C$ and $V_3 = V$ (Lines 5-8).

Step 3. If the partitioning operation in Step 2 fails, we construct a higher-order CI table, and the procedure goes back to Step 2. Follow this way, we finally obtain the causal partitioning $V = \{V_1, V_2, V_3\}$ (Lines 9-15).

Algorithm 2 Finding Causal Partitioning

```
 1: Input:
    $V$: The input variable set;
    $\sigma$: The threshold to limit the maximum order of CI table;
    $M$: The CI table w.r.t. $V$, initializes $M = zeros(|V|, |V|)$.
 2: Output:
    The causal partitioning $V = (V_1, V_2, V_3)$.
 3: $\forall v_i, v_j \in V$, set $M_{ij} = 1$ in case $v_i \perp v_j$.
 4: Divide $V$ into three non-overlapping parts $V = \{A, B, C = V \backslash_{A,B}\}$ by solving the optimization problem:
        $\min |C|$
        $s.t. \begin{cases} \forall v_i \in A, \forall v_j \in B, M_{ij} = 1 \\ |A| > 0, |B| > 0 \end{cases}$
 5: for $\forall v_i \in V_1 \cup V_2$ do
 6:     Remove $v_i$ from $V$ if $v_i$ and $\forall v_j \in C$ satisfy $M_{ij} = 1$;
 7: end for
 8: Let $V_1 = A \cup C$ , $V_2 = B \cup C$ and $V_3 = V$.
 9: if $\max\{|V_1|, |V_2|, |V_3|\} = |V|$ and M's order $k \le \sigma$ then
10:     for $\forall v_i, v_j \in V$ ($M_{i,j} = 0$) do
11:         Set $M_{i,j} = 1$ in case $\exists Z \subseteq V \backslash_{v_i, v_j}$ ($|Z| = k + 1$) such that $v_i \perp v_j | Z$.
12:     end for
13:     Goto line 4.
14: end if
15: Return $V = \{V_1, V_2, V_3\}$.
```

In practice, the maximum order of each CI table can be limited to a smaller number, likes 1 or 2, which is generally enough to partition V into subsets of small enough size, and can prevent CI tests from falling to Type II error. Furthermore, we have the following theorem, which ensures that the process of Steps 1-3 can find an appropriate partitioning.

Theorem 1 *The partitioning process in Alg. 2 returns a valid causal partitioning that d-separation is not violated.*

Proof: The input set V is first split into three non-overlapping subsets $\{A, B, C{=}V_{\backslash A,B}\}$ according to the adjacent matrix, A is therefore non-adjacent to B. Let $V_1{=}A \cup C$, $V_2{=}B \cup C$, then we can see that for any two variables in A (or B) must be d-separable in V_1 (or V_2). On the other hand, we remove V's maximum subset independent of C, and let $V_3 = V$. Then V_3 contains all the neighbors of C, thus for any two variables in C are d-separable in V_3 (Tian, Pearl, and Paz 1998). Therefore, u and v will not be separated in V_1, V_2 and V_3 if u is adjacent to v, and $\forall u, v \in V$ are d-separable in V_1 or V_2 or V_3 if u is non-adjacent to v. Therefore, the partitioning returned by Alg. 2 satisfies the definition of causal partitioning in Def. 1 and does not violate d-separation.

Example 1. Fig. 2(a) is an example to illustrate the process of Alg. 2. The input set $V{=}\{v_1, ..., v_6\}$ is partitioned by 1-order CI table M (as 0-order CI table is not enough to partition V, we therefore need to construct the 1-order CI table), we have $V{=}\{A{=}\{v_1, v_2\}, B{=}\{v_4, v_6\}, C{=}\{v_3, v_5\}\}$. As $M_{31}{=}1$ and $M_{36}{=}1$, we further obtain $V_1{=}\{v_1, v_2, v_3, v_5\}$, $V_2{=}\{v_3, v_4, v_5, v_6\}$ and $V_3{=}\{v_2, v_3, v_4, v_5\}$. We can see that $\forall v_i, v_j \in V$, if v_i and v_j are non-adjacent, then they are split into different subsets, or they are d-separable in at least one subset V_1 or V_2 or V_3, i.e., the d-separation holds.

Example 2. On the other hand, as Alg. 2 is a subroutine of CAPA (Alg. 1), here we give an example in Fig. 2(b) to illustrate the whole process of finding causal partitionings by CAPA. The original variable set V is partitioned into three subsets $\{V_1, V_2, V_3\}$ based on the k_1-order ($k_1 \geq 0$) CI table regarding V. Suppose V_1 and V_2 cannot be further partitioned, we terminate the recursive partitioning process on V_1 and V_2. And V_3 will be further partitioned into V_4, V_5 and V_6 based on the k_2-order ($k_2 \geq k_1$) CI table regarding V_3. Such a process continues till all subsets meet the termination condition.

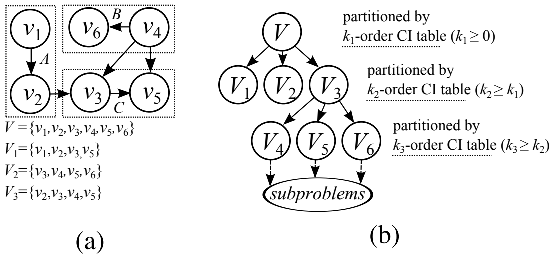

v_1
v_2
v_3
v_4
v_5
v_6
A
B
C
$V=\{v_1,v_2,v_3,v_4,v_5,v_6\}$
$V_1=\{v_1,v_2,v_3,v_5\}$
$V_2=\{v_3,v_4,v_5,v_6\}$
$V_3=\{v_2,v_3,v_4,v_5\}$
(a)
V
V_1
V_2
V_3
V_4
V_5
V_6
subproblems
partitioned by k_1-order CI table ($k_1 \geq 0$)
partitioned by k_2-order CI table ($k_2 \geq k_1$)
partitioned by k_3-order CI table ($k_3 \geq k_2$)
(b)

Figure 2: (a) An example of finding causal partitioning by using Alg.2; (b) An example of finding causal partitionings by CAPA.

We can see that the partitioning process in CAPA constitutes a hierarchy, like a ternary tree. Child subsets are resulted from their parent subset by using the same or a higher-order CI table. Therefore, this partitioning scheme can reduce the number of redundant CI tests as many as possible, and also generates more reliable results than the existing recursive methods such as SADA, as these methods use high-order CI tests in each iteration.

Distinguishing Markov Equivalence Classes

As aforementioned in Alg. 1, a specified causal learning algorithm A_g will be used to solve a subset if it cannot be further partitioned by Alg. 2. In this subsection, we study how to distinguish Markov equivalence classes via ReCIT. Here, we first review the process of discovering V-structures, which is the critical step for CI-based methods to determine causal directions. Consider a local structure $x - z - y$, if x is independent of y given a set of variables Z, $x \perp y|Z$, then we can infer that $x - z - y$ is $x{\rightarrow}z{\leftarrow}y$ in case $z \notin Z$ according to the mechanism of d-separation. But, if $z \in Z$, we cannot draw any conclusion about the causal directions of $x - z - y$ (before consistent propagation). Thus, the problem turns to how to orient directions in the case of $z \in Z$. We have the following theorem:

Theorem 2 *Given a variable set V generated by linear non-Gaussian additive noise model. For any subset of V and its corresponding subgraph satisfying the faithfulness condition, and containing two random variables x and y as well as a set of other variables Z, if x (or y) is adjacent to z ($z \in Z$) and $x{-}E(x|Z){\perp}z$ (or $y{-}E(y|Z){\perp}z$) holds, then z causes x (or z causes y).*

Proof: Without loss of generality, we assume $x{-}E(x|Z){\perp}z$. Let ε denote the exogenous disturbance of z. If $x{-}E(x|Z){\not\perp}\varepsilon$, then $x{-}E(x|Z){\not\perp}z$ according to Darmois-Skitovitch theorem (Darmois 1953; Skitovich 1953). We therefore have $x{-}E(x|Z){\perp}\varepsilon$, which means x and ε can be d-separated by Z under the faithfulness condition. If z is a child of x, then $x{\rightarrow}z{\leftarrow}\varepsilon$ forms a V-structure where z is a collider, then there must be $x - E(x|Z) \not\perp \varepsilon$ or faithfulness is violated. This is a contradiction. Therefore, z can only be the parent of x. Similarly, we can prove the case w.r.t. y.

According to the process of causal partitioning, the (local) structure similar to $x{\rightarrow}z{\rightarrow}y$, $x{\leftarrow}z{\leftarrow}y$ and $x{\leftarrow}z{\rightarrow}y$ will be preserved in at least one subset. Therefore, the dependence between $x{-}E(x|Z)$ and z can be used for determining directions (Line 3, Alg. 1).

Theoretical Analysis

In this section, we study the properties of CAPA, including correctness, completeness and complexity.

Correctness and completeness. We have the following theorem:

Theorem 3 *Given a variable set $V{=}\{v_1, ..., v_n\}$ following a causal graph G. If all CI tests performed in CAPA are reliable, then CAPA returns the actual graph G.*

Proof: Let G' denote the causal graph reconstructed by CAPA. The correctness and completeness are equivalent to the propositions:

1. Completeness: $\forall(v_1{\rightarrow}v_2) \in G \Rightarrow (v_1{\rightarrow}v_2) \in G'$;
2. Correctness: $\forall(v_1{\rightarrow}v_2) \in G' \Rightarrow (v_1{\rightarrow}v_2) \in G$.

As discussed in Theorem 1, the causal partitioning returned by Alg. 2 in each iteration is theoretically valid. Assume that V is first partitioned into (V_1, V_2, V_3), and the three subsets cannot be further partitioned. According to the Condition 2 in Def. 1, we know that any adjacent variables cannot be separated during partitioning, the completeness is therefore guaranteed. On the other hand, because of the Condition 3 in Def. 1, any non-adjacent variable is either separated during partitioning or d-separable in at least one subset, i.e., the correctness is satisfied.

Therefore, the question turns to that if V_1 (or V_2, V_3) can be further partitioned, can CAPA still meet the completeness and correctness? We present the following proposition: *If two groups of variable sets* S_1=$\{V_1,...,V_t,...,V_m\}$ *and* S_2=$\{V_{t_1},...,V_{t_k}\}$ *are two causal partitionings over* V *and* V_t *respectively, then* $S_2 \cup S_1 \backslash_{V_t}$ *forms a causal partitioning over* V. The proof of this proposition is straightforward according to Definition 1. It implies that no matter how many times a set is partitioned by CAPA, if the returned causal partitioning in each iteration is valid, then all the resulting partitionings constitute an whole valid partitioning. Thus, as aforementioned, CAPA meets the completeness and correctness.

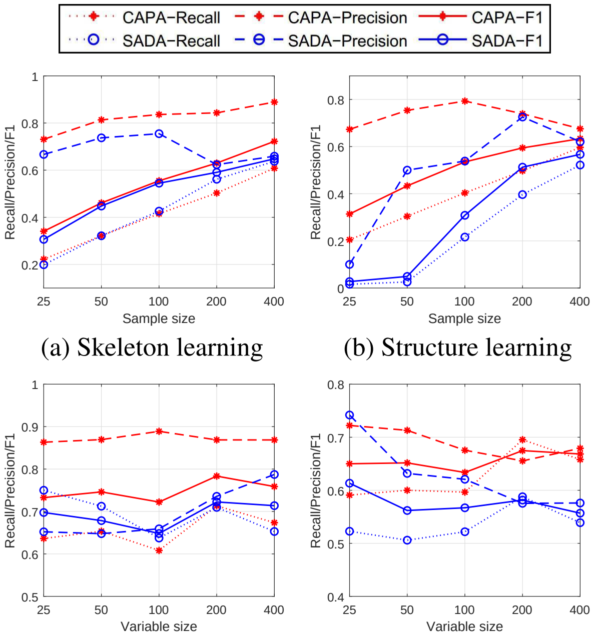
Time complexity. We focus on the number of CI tests used in CAPA since the other operations are computationally negligible compared to CI tests. Suppose that the original variable set V=$\{v_1, ..., v_n\}$ is recursively partitioned into m subsets $\{V_1, ..., V_m\}$ where $|V_m| \leq n$ for all m. Suppose that we use the PC algorithm as the basic algorithm. Then the time complexity of solving subproblems is $O(mk_{max}^2 2^{k_{max}-2})$, where k_{max} = $\max(|V_1|, ..., |V_m|)$. On the other hand, we need to calculate a CI table in each iteration. In the worst case, we have to calculate a σ-order CI table w.r.t. the original variable set V. Therefore, the upper bound of time complexity of CAPA is $O(mk_{max}^2 2^{k_{max}-2} + n^{\sigma+2})$. In practice, the step of dividing a set into three non-overlapping subsets (Line 4 in Alg. 2) may consume considerable time if the causal structure is very large or complex. We have three strategies to accelerate CAPA: 1) using a σ-order CI table in the first time instead of using ones from 0 to σ-order, 2) terminating the causal partitioning process when the current subset is sufficient small, and 3) employing a faster CI testing method such as partial correlation with Fisher transformation (Cai, Zhang, and Hao 2017) to check CI. If CI holds, we further use ReCIT to orient causal directions.

Performance Evaluation

We first compare CAPA with one of the latest recursive learning methods SADA (Cai, Zhang, and Hao 2017) by extensive simulated experiments for evaluating their abilities of finding causal partitioning and learning causal directions. To further illustrate the advantage of CAPA in causal structure learning, we also compare CAPA with four other existing causal learning methods, including LiNGAM (Shimizu et al. 2006), DLiNGAM (Shimizu et al. 2011), Sparse-ICA LiNGAM (Zhang et al. 2009) and SADA-LiNGAM (Cai, Zhang, and Hao 2017), over various real-world causal structures. Note that all these four methods can distinguish Markov equivalence classes, in which SADA-LiNGAM stands for the state of the art in high-dimensional cases.

Performance on simulated structures. In this group of experiments, we evaluate our method on datasets generated by simulated causal network structures, under the linear non-Gaussian model. Because there are not large-scale causal inference problems with ground truth, simulated data on synthetic and real-world structures are used in most causal structure learning methods (Kalisch and Bühlmann 2007). The structures and data generating processes are similar to those presented in (Cai, Zhang, and Hao 2017). Concretely, we first randomly generate a set of root nodes, then iteratively generate descendants in two steps: 1) Randomly select a subset of nodes from the generated nodes; 2) Using the selected nodes as parent nodes to generate a descendant with the average in-degree being 1.5. We then generate the data according to the corresponding structures with a linear function as v_i=$\sum_{v_j \in Pa_{v_i}} w_{ij} v_j + \varepsilon_i$, where Pa_{v_i} denotes the parents of v_i and ε_i is the non-Gaussian noise term. When generating these linear functions, we let $\sum_{Pa_{v_i}} w_{ij}$=1 and the variance Var_{ε_i}=1 for every variable v_i. To save time, we terminate the causal partitioning process when the regarding subset is smaller than $|V|/10$ and limit the maximum size of controlling set at 3.

We first compare CAPA with SADA over the above simulated models with different sample sizes {25, 50, 100, 200, 400} and 100 nodes. Both CAPA and SADA use PC algorithm (Spirtes, Glymour, and Scheines 2000) as their basic algorithm for causality discovery from the partitioned subsets. The results are shown in Fig. 3(a) and (b).



(a) Skeleton learning (b) Structure learning

(c) Skeleton learning (d) Structure learning

Figure 3: Performance comparison between CAPA and SADA on simulated causal networks. (a) and (b) are the results of causal skeleton and structure learning over different sample sizes {25, 50, 100, 200, 400} with 100 nodes; (c) and (d) are the results of different dimensional networks {25, 50, 100, 200, 400} with 400 samples.

From Fig. 3(a), we can see that CAPA performs better than SADA in terms of *precision* and $F1$ on causal skeleton learning for different sample sizes. The reason is that d-separation is violated in SADA, thus many non-adjacent variables cannot be separated by the base solver. However, the *recall* score of SADA is slightly better than that of CAPA. This is the result of two counteracting factors: 1) d-separation is preserved in CAPA, thus there are some subsets that can no longer be partitioned by CAPA, but can be still partitioned by SADA. That is, the CI tests regarding some adjacent variables in these 'larger' subsets are easier to fall into Type II error. 2) In SADA, the variable set is divided by using normal order CI tests, while in CAPA, the variable set is partitioned the σ-order CI table. Moreover, CAPA uses much fewer CI tests than SADA. Therefore, as the size of variable set increases, CI testing between two variables in SADA becomes easier and easier to fall into Type II error.

We also evaluate our method on causal structure learning (with direction orientation), and the results are shown in Fig. 3(b). CAPA performs much better than SADA in terms of *Recall*, *Precision* and $F1$. This is because CAPA can learn more causal directions by $x-E(x|Z) \perp z$ (or $y-E(y|Z) \perp z$) according to Theorem 2, while SADA orients directions based on only V-structure learning and consistent propagation.

We then compare CAPA with SADA over different dimensional networks {25, 50, 100, 200, 400} with 400 samples. The results are presented in Fig. 3(c) and (d), which show that the performance of CAPA and SADA for different variable sizes is relatively stable on both skeleton and structure learning. We can conclude that the two methods are able to solve relatively higher dimensional problems over these simulated networks. However, real-world causal structures are more complex, and the dimensionality will impact the performance of the two methods, which will be further discussed later.

Performance of distinguishing equivalence classes. To further illustrate the advantage of CAPA in inferring causal directions, here we apply CAPA to a causal graph presented in (Shimizu et al. 2006), which was generated by following linear non-Gaussian structure equation model w.r.t. a DAG consisting of six variables as shown in Fig. 4(a). The graphs reconstructed by CAPA and SADA are shown in Fig. 4(b) and (c), respectively. We can see that all the causal directions discovered by CAPA are correct, because CAPA can infer $(v_1, v_3) \rightarrow v_4$ according to $v_4 - E(v_4|v_1, v_3) \perp (v_1, v_3)$. Similarly, $(v_1, v_2) \rightarrow (v_3, v_5)$ and $(v_2, v_3) \rightarrow v_6$ can also be obtained by CAPA. There is only one edge $v_1 - v_2$ that is not oriented. On the other hand, as shown in Fig. 4(c), though the skeleton is correct, the corresponding directions are not inferred by any propagation. Because there is no V-structure in this graph, theoretically SADA cannot find any causal direction.

Performance on real-world structures

In this subsection, we compare CAPA with four other existing causal structure learning methods, including LiNGAM (Shimizu et al. 2006), DLiNGAM (Shimizu et al. 2011), Sparse-ICA LiNGAM (Zhang et al. 2009) and SADA-LiNGAM (Cai, Zhang, and Hao 2017). As all these methods can break Markov equivalence classes, we therefore can further evaluate the performance of our method in causal structure learning. The implementations of LiNGAM, DLiNGAM and SADA-LiNGAM strictly follow the corresponding original papers (Shimizu et al. 2006; 2011; Cai, Zhang, and Hao 2017). The implementation of Sparse-ICA LiNGAM is based on the sparse-ICA of (Zhang et al. 2009) and the pruning algorithm of (Shimizu et al. 2006). Among these existing methods, SADA-LiNGAM is the most effective approach for learning causal structures of high dimensional cases, where LiNGAM is selected as the base solver by default. All methods are evaluated on eight real-world causal network structures [1] that cover a variety of applications, including insurance evaluation (*Insurance*), medicine (*Alarm and Pathfinder*), agricultural industry (*Barley*), weather forecasting (*Hailfinder*), system troubleshooting (*Win95pts and Andes*) and the pedigree of breeding pigs (*Pigs*). Table 1 gives the structural statistics of these causal networks.

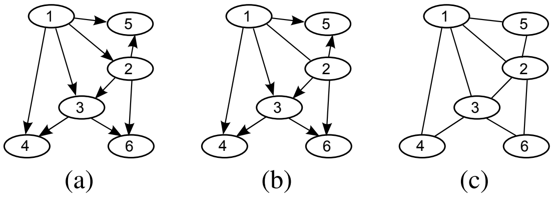


Figure 4: Performance comparison in causal direction inference. (a) The ground truth causal model; (b) The DAG reconstructed by CAPA; (c) The PDAG reconstructed by SADA.

.

Note that the performance of the four counterparts is highly influenced by the ratio of the sample size to the number of nodes (Cai, Zhang, and Hao 2017), and the baseline approach LiNGAM is usually unreliable when the number of samples is less than $2|V|$. We therefore compare CAPA against the four existing methods by fixing the sample size to $2|V|$ in the following experiments.

Table 1: Statistics of the eight causal network structures.

Dataset	Nodes#	Avg. degree	Max degree
Insurance	27	3.95	9
Alarm	37	2.49	6
Barley	48	3.50	8
Hailfinder	56	2.36	17
Win95pts	76	1.84	9
Pathfinder	109	3.58	106
Andes	223	3.03	12
Pigs	441	2.68	41

The results are shown in Table 2, where for compressing the space in the table, SADA-LiNGAM, LiNGAM, DLiNGAM and Sparse-ICA LiNGAM are simply denoted as SL, LG, DLG and SICA, respectively. It can be seen that CAPA achieves significantly better *precision* and *F1* score on all structures. Only in the case of *Insurance* LiNGAM works

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

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

Dataset	*Recall*					*Precision*				
	CAPA	SL	LG	DLG	SICA	CAPA	SL	LG	DLG	SICA
Insurance	0.28	0.24	**0.47**	0.39	0.23	**0.78**	0.45	0.11	0.09	0.06
Alarm	**0.39**	0.38	0.38	0.27	0.39	**0.81**	0.44	0.24	0.16	0.22
Barley	**0.37**	0.26	0.33	0.23	0.36	**0.73**	0.43	0.21	0.15	0.22
Hailfinder	**0.55**	0.50	0.25	0.20	0.31	**0.80**	0.50	0.22	0.17	0.28
Win95pts	**0.51**	0.49	0.28	0.20	0.36	**0.82**	0.45	0.37	0.25	0.35
Pathfinder	0.59	**0.81**	0.35	0.35	0.32	**0.68**	0.09	0.22	0.22	0.20
Andes	**0.72**	0.55	0.21	0.12	0.26	**0.61**	0.18	0.46	0.26	0.51
Pigs	**0.88**	0.53	0.15	N.A.	N.A.	**0.75**	0.22	0.59	N.A.	N.A.

	F1 score					*Elapsed time (s)*				
	CAPA	SL	LG	DLG	SICA	CAPA	SL	LG	DLG	SICA
Insurance	**0.42**	0.30	0.18	0.14	0.10	1.33	6.40	**0.33**	1.68	1.12
Alarm	**0.49**	0.41	0.29	0.20	0.28	3.36	2.65	**0.40**	6.04	3.00
Barley	**0.53**	0.31	0.26	0.18	0.27	14.9	19.3	**0.45**	10.8	7.10
Hailfinder	**0.65**	0.50	0.23	0.18	0.29	16.9	21.3	**0.13**	22.1	12.0
Win95pts	**0.63**	0.47	0.32	0.22	0.35	22.3	17.8	**1.22**	66.2	34.5
Pathfinder	**0.64**	0.16	0.27	0.27	0.25	440	$4 \cdot 10^5$	**2.67**	490	212
Andes	**0.66**	0.26	0.28	0.16	0.34	$2 \cdot 10^3$	$6 \cdot 10^4$	**34.5**	$4 \cdot 10^3$	$2 \cdot 10^3$
Pigs	**0.81**	0.34	0.24	N.A.	N.A.	$1 \cdot 10^4$	$1 \cdot 10^6$	**641**	N.A.	N.A.

better than CAPA and in the case of *Pathfinder* SADA-LiNGAM works better than CAPA, all in terms of *Recall* score. In most cases, especially in larger causal networks (with $|V| > 100$), CAPA works much better than SADA-LiNGAM. The other three methods, LiNGAM, DLiNGAM and Sparse-ICA LiNGAM are not competitive in all these cases in terms of learning accuracy. As DLiNGAM and Sparse-ICA LiNGAM are of high time-complexity, here we do not present their results on the *Pigs* network.

In summary, we have the following observations:

1. The performance (*Recall*, *Precision* and *F1* score) of CAPA turns better with the increase of the sample size, rather than the ratio of the sample size to the number of nodes ($2|V|$), while the performance of the four existing methods work more stable with a fixed ratio of the sample size to the number of nodes ($2|V|$). We can also see that on larger networks, *Pathfinder*, *Andes* and *Pigs*, the *F1* score of CAPA is from 2 to 3 times higher than that of the four existing methods. Therefore, CAPA is more effective in causal discovery in high-dimensional cases in terms of inference accuracy when limited samples are given.

2. As the dimensionality of causal networks increases, the ratio of *Recall* to *Precision* of CAPA remains relatively stable, therefore the *F1* score of CAPA maintains at an acceptable level. On the contrary, we can see that on small causal networks, *Precision* of SADA-LiNGAM is slightly higher than *Recall*, while in the cases of larger causal networks, like *Pathfinder* and *Andes*, *Precision* of SADA-LiNGAM is much lower than *Recall*. Note that *Recall* is the fraction of actual causality found by the algorithm, and *Precision* is the actual fraction of inferred causality with respect to the true graph. We can say that SADA-LiNGAM cannot remove many incorrect causal relationships in these networks. On the other hand, CAPA does particularly well in all these networks in terms of *precision*.

3. By comparing the time costs of CAPA and SADA-LiNGAM, we notice that the time cost of CAPA increases with the size of variables, while the efficiency of SADA-LiNGAM is not stable. The reason is that in the step of finding causal partitioning, CAPA determines a smaller C set by solving an optimization problem (Line 4 in Alg. 2), while SADA-LiNGAM chooses a C set randomly. Generally, the smaller C, the higher accuracy and the less running time (Cai, Zhang, and Hao 2017). So we can conclude that CAPA is more applicable to causal discovery in high-dimensional cases than these existing methods.

Conclusion

In this paper, we propose a recursively causal structure learning method called CAPA to support effective and efficient causality discovery over large variable sets. We first design a new variables partitioning scheme and show that this partitioning scheme does not violate d-separation criterion. That is to say, the complete causality of the original variable set can be recovered from the resulting partitions. Then, we develop an effective and efficient algorithm that is recursively applied to searching for such partitionings from the input variable set, by using only low-order CI tests. Moreover, regression-based conditional independence test (ReCIT) is used for checking CIs, and we prove that more causal directions can be detected by ReCIT, thus CAPA can return a more accurate causal graph instead of a set of Markov equivalence classes. Our theoretical analysis proves the correctness and completeness of the proposed method, and extensive experiments on various real-world causal networks verify the advantage of CAPA over five benchmarks, including SADA, SADA-LiNGAM, LiNGAM, DLiNGAM and Sparse-ICA LiNGAM.

Acknowledgement: This work was supported by NSFC under gant No. U1636205. Jihong Guan and Xin Wang were supported by NSFC under grants No. 61772367 and No. 61772420, respectively.

References

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

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

Cai, R.; Zhang, Z.; and Hao, Z. 2017. SADA: A general framework to support robust causation discovery with theoretical guarantee. *CoRR* abs/1707.01283.

Chickering, D. M. 2002. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research* 2:445–498.

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

Doran, G.; Muandet, K.; Zhang, K.; and Schölkopf, B. 2014. A permutation-based kernel conditional independence test.

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

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.

Gao, T., and Ji, Q. 2015. Local causal discovery of direct causes and effects. In *Advances in Neural Information Processing Systems*, 2512–2520.

Geng, Z.; Wang, C.; and Zhao, Q. 2005. Decomposition of search for v-structures in DAGs. *Journal of Multivariate Analysis* 96(2):282–294.

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.

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.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Liu, H.; Zhou, S.; Lam, W.; and Guan, J. 2017. A new hybrid method for learning bayesian networks: Separation and reunion. *Knowledge-Based Systems* 121:185 – 197.

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

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

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

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

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.

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*.

Tian, J.; Pearl, J.; and Paz, A. 1998. Finding minimal d-separators.

Xie, X., and Geng, Z. 2008. A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research* 9(3):459–483.

Xie, X.; Geng, Z.; and Zhao, Q. 2006. Decomposition of structural learning about directed acyclic graphs. *Artificial Intelligence* 170(4-5):422–439.

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.; Peng, H.; Chan, L.; and Hyvärinen, A. 2009. Ica with sparse connections: Revisited. In *International Conference on Independent Component Analysis and Signal Separation*, 195–202. Springer.

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, H.; Zhou, S.; Zhang, K.; and Guan, J. 2017. Causal discovery using regression-based conditional independence tests. In *AAAI Conference on Artificial Intelligence*.

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*.