Cost-effectively Identifying Causal Effects When Only Response Variable is Observable

Tian-Zuo Wang [1] **Xi-Zhu Wu** [1] **Sheng-Jun Huang** [2] **Zhi-Hua Zhou** [1]

Abstract

In many real tasks, we care about how to make decisions rather than mere predictions on an event, e.g. how to increase the revenue next month instead of merely knowing it will drop. The key is to identify the causal effects on the desired event. It is achievable with do-calculus if the causal structure is known; however, in many real tasks it is not easy to infer the whole causal structure with the observational data. Introducing external interventions is needed to achieve it. In this paper, we study the situation where only the response variable is observable under intervention. We propose a novel approach which is able to cost-effectively identify the causal effects, by an active strategy introducing limited interventions, and thus guide decision-making. Theoretical analysis and empirical studies validate the effectiveness of the proposed approach.

1. Introduction

Making accurate predictions on an event has been extensively studied, and achieved great success in various applications. In many real tasks, however, it is more important to know how to make decisions rather than mere predictions. For example, instead of knowing that there will be a drop in sales, sellers care more about how to increase sales; and doctors care about how to rehabilitate patients rather than merely predicting the disease. A reasonable solution for such problems is to *make causal effect identification*, i.e. identify the causal effect of each variable on the event. Then one can know how to make the optimal decision by an intervention on a specific variable with a specific value.

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Correspondence to: Sheng-Jun Huang <huangsj@nuaa.edu.cn>, Zhi-Hua Zhou <zhouzh@lamda.nju.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Pearl contributed a series of seminal works (Pearl, 2009) in identifying causal effects. Given the knowledge of the causal graph, do-calculus is an excellent tool to make it. However, the causal graph is usually not available in advance, and thus it is crucial to discover the causal structure.

There are many classical off-the-shelf causal discovery methods (Spirtes et al., 2000; Chickering, 2002), which typically obtain a *Markov equivalence class* based on the observational data. Unfortunately, there often remains unknown causal relations in the Markov equivalence class, leading to some unidentified causal effects. A primary solution is to discover the additional causal relations from the interventional data generated by actively intervening on some variables. Some related works are proposed to discover the whole structure with the target of reducing the total intervention times or cost (He & Geng, 2008; Hauser & Bühlmann, 2014; Kocaoglu et al., 2017). When all variables under interventions are observable, these methods work quite well.

In many real tasks, however, when performing interventions, it is expensive or even hard to observe all variables. For example, after receiving treatments, the patients are not willing to undergo a full medical examination once again. Usually, doctors merely get the feedback of their illness conditions. In many cases, only the event we want to change, i.e., the *response variable*, is easily accessible when we impose an intervention. In this paper, to deal with such problems, we propose a novel approach **ACI** for **A**ctive **C**ausal effect **I**dentification, i.e., identify causal effects when *only* response variable is observable under active intervention.

Due to the limitation of observations, it is no longer easy to identify the causal structure by methods of He & Geng (2008) and Hauser & Bühlmann (2014). Specifically, these methods identify causal relations by just *whether* the distribution of some variable changes under interventions on others. In their setting, such consideration is sufficient for the causal discovery process. But when only the post-interventional information of response variable is available, they may fail to discover the causal structure. For example, the structures in Fig. 1(a) and Fig. 1(b) cannot be distinguished because no matter which variables we intervene on, the distribution of Y will change. We make a progress on how to use interventional data. To the best of our knowledge,

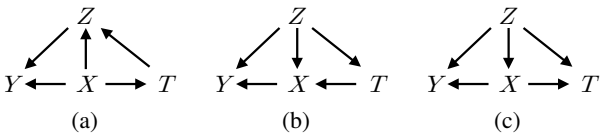


Figure 1. When only Y is observable, no matter whether structure (a) or (b) is true and which variable is intervened on, the distribution of Y will change under intervention. Hence it is not sufficient to differentiate them by observing whether Y changes. Besides, the causal effects of X on Y in structure (b) and (c) are equivalent.

we are the first one to exploit *how* the distribution changes, which can identify the indistinguishable case before.

Moreover, discovering the whole causal structure is no longer always attainable in our setting. Nevertheless, we notice that its discovery is not necessary for causal effect identification. Therefore, we identify *ancestor causal structure*, whose discovery is necessary and sufficient for identifying causal effect of each variable on the response variable. By exploiting the intrinsic distribution of the response variable, with actively applied interventions, the ancestor causal structure can be effectively identified with fewer interventions.

In summary, our contributions are twofold:

(1) Make things happen. Identify causal effects when only response variable is observable. We exploit the interventional distribution of the response variable and prove the identifiability of causal effects;

(2) Make things better. Reduce the number of interventions to make causal effect identification by an active intervention strategy. Empirically and theoretically, we show our strategy is effective and reasonable.

The reminder of paper is organized as follows. Section 2 reviews some related works. A brief preliminary is given in Section 3. Section 4 describes the proposed method ACI. Section 5 provides the theoretical guarantee for causal effect identifiability and presents an intervention cost analysis to make causal effect identification. Section 6 reports experimental results. Finally, we conclude our paper in Section 7.

2. Related Works

In Pearl's causality framework, there are lots of works towards identifying causal effects. Many related criteria have been established (Tian & Pearl, 2002; Shpitser & Pearl, 2006; Huang & Valtorta, 2006; Perkovic et al., 2015; Jaber et al., 2019; Lee et al., 2019). For example, Perkovic et al. (2015) provided necessary and sufficient graphical criteria for causal effect identification by covariate adjustment. And Jaber et al. (2019) proposed a complete result for causal effect identification under Markov equivalence class. They all provide solid results. But the causal graph is often unknown, or the partial causal graph we have, for instance an essential graph, is not enough to guarantee causal effect identification, so that we need discover some additional causal relations.

Identifying causal relations from data has been widely studied. Some are towards obtaining the essential graph (Spirtes et al., 2000; Chickering, 2002; Huang et al., 2018). Also, there are lots of works discovering causal structure based on additional assumptions (Shimizu et al., 2006; Hoyer et al., 2008; Zhang & Hyvärinen, 2009; Peters et al., 2011; 2014; Zhang et al., 2017; 2018; Cai et al., 2018). But in some conditions, the assumptions are not satisfied. To pursue a general framework, many works introduce interventions. One line is causal discovery based on observational data and existing interventional data (Cooper & Yoo, 1999; Hauser & Bühlmann, 2012; Triantafillou & Tsamardinos, 2015; Peters et al., 2016; Meinshausen et al., 2016; Wang et al., 2017; Kocaoglu et al., 2019). Another line enables active interventions (He & Geng, 2008; Hyttinen et al., 2013; Hauser & Bühlmann, 2014; Kocaoglu et al., 2017). These works aim to identify the whole causal structure and conduct strategies to reduce the intervention times. They mainly discover the causal relations by observing whether distribution of some variable changes under an intervention on others. It is sufficient for the setting that observes the post-interventional information of all the variables. However, when the full observations are not available, the methods above may fail.

3. Preliminary

Let $G = (\mathbf{V}, \mathbf{E})$ be a graph. $\mathbf{V}$ contains features $X_1, X_2, \cdots, X_p$ and the response variable Y. A *partially directed graph* contains both directed and undirected edges. After removing all arrowheads, we obtain the *skeleton*. V_i is a *parent/child/sibling* of V_j if $V_i \to V_j / V_i \leftarrow V_j / V_i - V_j$. If there is a directed path from V_i to V_j, then V_i (V_j) is an *ancestor* (*descendant*) of V_j (V_i). We denote the *parents/ancestors/siblings* set of V_i by $\mathrm{Pa}_i/\mathrm{Anc}_i/\mathrm{Sib}_i$. The set of undirected edges of V_i is denoted by E_{Sib_i}.

A (partial) causal graph is a *(partially) directed acyclic graph*, which is *(PDAG) DAG* for short. If two DAGs share the same conditional independence, they are *Markov equivalent*. The *Markov equivalence class (MEC)* is a set of DAGs in which each graph is Markov equivalent to others. An *essential graph* is a partially directed acyclic graph, and the edge is $V_i \to V_j$ if and only if in each DAG of MEC the edge is $V_i \to V_j$. A partially directed graph is a *chain graph* if there is no *partial cycle*, which is a partially directed path starting and ending in a same node (Lauritzen & Richardson, 2002). After deleting the directed edges in a chain graph, we divide it into a few *chain components* whose variables are connected in an undirected graph. *Meek rules* are criteria to orient some undirected edges in a partial causal graph. We

guide readers to Meek (1995a;b) for more details.

In this paper, capital and lower-case letters denote random variables and values respectively. In Pearl's *do-calculus* framework (Pearl, 2009), $do(X_i = x_i)$ represents intervening on variable X_i with value x_i. The causal effect of X_i on Y is denoted by $P(Y|do(X_i))$. We would claim X_i has a causal effect on Y if X_i is an ancestor of Y. In this circumstance, intervening on X will take a distribution change to the response variable Y, i.e. $P(Y|do(X_i = x_i)) \neq P(Y)$. Otherwise, we say X_i has *no causal effect* on Y.

Next, we give a brief introduction to back-door criterion. It is a tool for inferring the causal effect of X on Y with observational data given *the knowledge of causal graph*.

A set of variables $\mathbf{Z}$ is called a *back-door admissible set* for (X_i, Y) in a DAG G if no variable in $\mathbf{Z}$ is a descendant of X_i and $\mathbf{Z}$ blocks every path between X_i and Y that contains an arrow into X_i. By definition, Pa_i is one of the back-door admissible sets for (X_i, Y). With a back-door admissible set $\mathbf{Z}$ for (X_i, Y), we have

$$P\big(Y|do(X_i = x_i)\big) = \int_{\mathbf{Z}} P(\mathbf{Z})P(Y|X_i = x_i, \mathbf{Z})\,\mathrm{d}\mathbf{Z}. \tag{1}$$

This equation is across our paper. In general, we use the data of Y under intervention on X_i to infer which variable sets are back-door admissible sets for (X_i, Y), which implies that the direction of the edges between these variables and X_i are into X_i. Hence we can obtain some causal structure information from the post-interventional data of only Y.

4. The Proposed Approach

In this paper, Y denotes the response variable and $X_1, X_2, \cdots, X_p$ denote p variables. *Faithfulness* and *no latent variables* are assumed. We focus on discovering the causal structure for causal effect identification by observational data of all the variables as well as active experiments, in which we take singleton hard[1] interventions on variable from $X_1, X_2, \cdots, X_p$, but *only* the response variable can be collected. By the post-interventional data of Y, our approach orients some undirected edges. We repeat the process until identifying the causal effect of each variable on Y.

Now, we first give a definition about *minimal parental back-door admissible set (MPS)*, followed by an important assumption *interventional-faithfulness* across our paper for the causal discovery process with interventional data.

Definition 1 (Minimal Parental Back-door Admissible Set (MPS))**.** $\mathbf{M}$ is called a minimal parental back-door admissible set for (X_i, Y) in a DAG G if (1). all variables in $\mathbf{M}$ are parents of X_i, (2). $\mathbf{M}$ is a back-door admissible set for (X_i, Y), (3). no variable in $\mathbf{M}$ is conditional independent of Y given X_i and the other variables in $\mathbf{M}$.

Assumption 1 (Interventional-faithfulness)**.** *For two Markov equivalent DAGs with the same observational distribution, if $X_i \in Anc_Y$ and minimal parental back-door admissible sets for (X_i, Y) are different in the two DAGs, then $P\big(Y|do(X_i = x)\big)$ are different in the two DAGs.*

Similar to faithfulness assumption, which avoids the situation that there is an edge $X \to Y$ in the graph while X and Y happen to be independent in the observational distribution, the interventional-faithfulness assumption is to avoid the situation that the causal effects of X on Y in two causal graphs with *totally* different back-door admissible sets *happen to* be equal. For example, the causal effect of X on Y is $P(Y|X = x_0)$ in Fig. 1(a), while it is $\int_z P(Y|z, X = x_0)P(z)\,\mathrm{d}z$ in Fig. 1(c). In general, they are not equivalent. Assumption 1 assumes they cannot happen to be equal. It is noteworthy that different back-door admissible sets may induce the same causal effect. In Fig. 1(b), it is $\int_{z,t} P(z,t)P(Y|z,t,X = x_0)\,\mathrm{d}z\,\mathrm{d}t = \int_z P(z)P(Y|z, X = x_0)\,\mathrm{d}z$, which equals to that in Fig. 1(c). In such a situation, we call the back-door admissible sets are not *totally* different. In the assumption, MPS is introduced to help distinguish whether two back-door admissible sets are totally different, i.e., whether the causal effects in two causal graphs are not equivalent in general. We provide an analysis about the reasonableness of interventional-faithfulness assumption in Appendix B.

Our approach contains *graph decomposition*, *structure inference*, and *intervention variable selection*. Graph decomposition is to simplify the graph to be oriented. Structure inference is to infer the causal structure with the interventional data of response variable. And intervention variable selection aims to select the variable to be manipulated.

4.1. Graph Decomposition

The observational data contains some causal information between the variables. We thus obtain the essential graph by observational data. There exist some classical causal discovery approaches to make it, such as PC algorithm (Spirtes et al., 2000) and GES algorithm (Chickering, 2002). Here both of them are applicable. Based on the essential graph, we design an adaptive strategy[2] to do active experiments to make causal effect identification with fewer number of interventions. Considering the high cost of intervention, we first introduce the definition of *ancestor causal structure* and prove its discovery is the necessary and sufficient condi-

[1] Hard intervention means we intervene in a specific value. The alternative is soft intervention, which affects the causal mechanism generating the variable (Eberhardt, 2007; He & Geng, 2016).

[2] An adaptive strategy is an intervention strategy in which every intervention variable depends on the result of interventions before.

tion for causal effect identification. Then we present some theoretical guarantees for the graph decomposition part.

Definition 2 (Ancestor Edges and Ancestor Causal Structure)**.** In a causal graph G, ancestor edges are all the edges including at least one variable in Anc_Y. We denote them by E_A. Ancestor causal structure is the subgraph $G[E_A]$ induced by E_A.

Theorem 1. *The causal effect of each variable on the response variable Y is identifiable if and only if all the ancestor edges (ancestor causal structure) are identified.*

The proof is given in Appendix D.1. It implies if we know some undirected edges are not ancestor edges in the process of causal discovery, there is no need for us to orient them because they do not contribute to causal effect identification.

Proposition 1. *If some undirected edges in one chain component are oriented, after applying Meek rules, the undirected edges in other chain components remain.*

This follows Theorem 4 and 5 of He & Geng (2008).

Proposition 2. *In a chain component C of chain graph G, if the response variable $Y \notin C$ and there does not exist a directed path from v to Y for any variable $v \in C$ in G, then there is no directed path from C to Y in the causal graph.*

The proof is in Appendix D.1. In our method, graph decomposition contains two parts. For a chain graph G, which is an essential graph or the graph inferred after last intervention, we divide it into several chain components, and then intervene and identify causal relations separately. The reason is the causal relation inference in each chain component is independent as implied in Proposition 1. Moreover, we ignore the chain components without directed paths to Y in G because the undirected edges here cannot be ancestor edges as indicated in Proposition 2, identifying them is unnecessary for causal effect identification according to Theorem 1. Next, we will discuss structure inference and intervention variable selection in the level of *chain component*.

4.2. Structure Inference

Here, we focus on how to identify undirected edges in chain component C of chain graph G with the post-interventional data of Y. We assume X_i is the selected intervention variable. The criterion to select it is presented in the next part. We first introduce a definition.

Definition 3 (Possible Causal Structure (PCS) of X_i)**.** Let G be a chain graph, a possible causal structure of X_i is an acyclic graph such that E_{Sib_i} is oriented from G, as well as it shares the common v-structures with G.[3] Sometimes we omit "of X_i" and call PCS for short.

[3] *Possible causal structure of X_i* is not necessarily a directed graph because there may be undirected edges out of E_{Sib_i}.

Algorithm 1 Get the minimal parental back-door admissible set for (X_i, Y)

input: Intervention variable X_i, causal structure or PCS G
Initialize $\mathbf{M} = \{\text{Pa}_i \text{ in } G\}$
if there are undirected edges in G **then**
    Update G by orienting all undirected edges on the premise that generates no new v-structures or cycles
end if
for v **in** $\mathbf{M}$ **do**
    if $v \perp Y | \{X_i, \mathbf{M} \backslash \{v\}\}$ in G **then**
        $\mathbf{M} = \mathbf{M} \backslash \{v\}$
    end if
end for
output: $\mathbf{M}$.

All undirected edges of X_i in E_{Sib_i} have two possible directions. We list each PCS G_j of X_i, where G_j is obtained from an orientation on E_{Sib_i}, and j is the index for a PCS. Our main idea here is to find the PCS in which the estimated causal effect is consistent to interventional data under real intervention. Then identify some edges in E_{Sib_i}.

It is noteworthy that different PCSs possibly induce the same causal effect of X_i on Y. Hence we divide all PCSs into several classes. The causal effects in the structures from one class are the same. By the interventional data of Y, we can select the class in which the causal effect estimation is consistent to real intervention. Then we update the common edges of X_i in this class to G. The other undirected edges still remain. To identify which possible causal structures are bracketed, we consider Definition 1. We reveal in Proposition 3 that the causal effects of X_i on Y are the same for G_j and G_k if and only if the MPSs for (X_i, Y) in these PCSs of X_i are the same. By the definition, we propose Algorithm 1 to obtain MPS $\mathbf{M}_j$ for (X_i, Y) in each PCS G_j.

Proposition 3. *Let G_j and G_k be two possible causal structures of X_i. $P_{G_j}\big(Y|do(X_i = x_i)\big) = P_{G_k}\big(Y|do(X_i = x_i)\big)$ holds if and only if $\mathbf{M}_j = \mathbf{M}_k$, where $P_{G_j}\big(Y|do(X_i = x_i)\big)$ and $\mathbf{M_j}$ denote the causal effect of $X_i = x_i$ on Y and the MPS for (X_i, Y) in G_j, respectively.*

The necessity follows from back-door criterion and d-separation directly. The adequacy is from interventional-faithfulness assumption. Let $\hat{P}_{G_j}\big(Y|do(X_i = x_i)\big)$ and $\hat{P}\big(Y|do(X_i = x_i)\big)$ denote the estimated causal effect in PCS G_j and that under real intervention respectively. $Disc(P, Q)$ is the distribution discrepancy between P and Q. If there are m intervention samples $\big((do(X_i = x_{i_1}), Y_1\big), \cdots, \big(do(X_i =$

Algorithm 2 Update the Graph

input: Intervention variable X_i, chain graph G;
Obtain $\mathbf{M}_j$ for every possible causal structure G_j by Algorithm 1
Find $G_\star$ by (2) and get its minimal parental back-door admissible set $\mathbf{M}_*$
Initialize $\mathcal{G} = \emptyset$
for each G_j **do**
  if $\mathbf{M}_j$ equals to $\mathbf{M}_\star$ **then**
    $\mathcal{G} = \mathcal{G} \bigcup \{G_j\}$
  end if
end for
if the size of $\mathcal{G}$ is 1 **then**
  Orient G according to edges of X_i in $\mathcal{G}$
else
  Orient G according to the common edges of X_i in $\mathcal{G}$
end if
output: Updated graph G.

$x_{i_s}), Y_s\big), \cdots, \big(do(X_i = x_{i_m}), Y_m\big)\big)$, we take the PCS by

$$G_\star = \arg\min_{G_j} \sum_{s=1}^{m} \textit{Disc}\Big(\hat{P}_{G_j}\big(Y|do(X_i = x_{i_s})\big), \hat{P}\big(Y|do(X_i = x_{i_s})\big)\Big) \tag{2}$$

and get the corresponding MPS $\mathbf{M}_\star$. All PCSs with MPS $\mathbf{M}_\star$ are selected and their common edges of X_i are updated to the chain graph G. The process is shown in Algorithm 2.

Any distance measure of distribution can be used here. In our experiments, we just take the difference of expectations as the measure for the convenience of calculation. In this condition, (2) is $G_\star = \arg\min_{G_j} \sum_{s=1}^{m} \big|\hat{\mathbb{E}}_{G_j}\big(Y|do(X_i = x_{i_s})\big) - Y_s\big|$. The calculation of causal effects is prone to suffer the curse of dimensionality. We refer to Monte Carlo Method to estimate it avoiding high-dimensional estimation. The details are given in Appendix C.

4.3. Intervention Variable Selection

In our criterion to select the intervention variable, we consider the two following aspects.

(1) Each intervention must reveal some information about ancestor edges considering the high cost of intervention. The selected variable should thus guarantee identifying some undirected ancestor edges by our method to infer the structure with the interventional data of Y no matter what the potential causal structure is;

Algorithm 3 ACI (Active Causal Effect Identification)

input: Chain graph G;
repeat
  $\mathcal{C}'$ = {all the chain components of G}
  $\mathcal{C} = \{A | A \in \mathcal{C}', \exists v \in A$, s.t. $v \to \cdots \to Y\}$
  // `Delete the chain components without directed paths to` Y
  if $\mathcal{C}$ is not empty **then**
    Choose $C \in \mathcal{C}$ and select an intervention variable X according to our selection criterion in Section 3.3
    Orient some edges according to Algorithm 2 and obtain an updated graph G
    G = Meek(G) // `Applying Meek rules`
  end if
until $\mathcal{C} = \emptyset$
output: The ancestor causal structure $G[E_A]$.

(2) Identify more undirected edges by this intervention on the basis of achieving the first goal.

The main challenge here is to achieve the first goal. We present our strategy to determine the intervention variable selection set in Fig. 2. We denote the variable set in chain component C by $\text{Anc}_Y[\ddot{C}]$ in which each variable has a directed path to Y *outside* C in chain graph G, and its size is $|\text{Anc}_Y[\ddot{C}]|$. The selected set is marked in the blue box. The idea behind it is to intervene on a variable which convinces us that it has at least one undirected ancestor edge. At the same time, the different directions of this edge will lead to distinct MPSs such that different causal effects, which guarantees we can identify it by the data of Y under intervention. All variables in this selection set meet the first goal. A theoretical guarantee is given in the next section.

We notice that it is possible to identify all undirected edges of the intervened variable in one experiment by exploiting the post-interventional distribution of Y in our approach. Therefore, to pursue the second goal, we greedily select the variable with the maximum undirected edges from the intervention variable selection set to intervene.

Combining the above points, we keep intervening and orienting some undirected edges until identifying all ancestor edges. The complete algorithm ACI is as in Algorithm 3. For the variables in the ancestor causal structure, their causal effects on Y are determined by back-door criterion, while for others, it is proved that they have no causal effects on Y.

5. Theoretical Guarantee

In this section, we first provide the theoretical guarantee for causal effect identifiability. Then, analyses about the intervention cost to make causal effect identification follow.

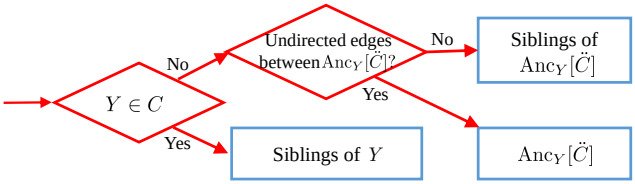


Figure 2. The decision tree to determine the selection set.

5.1. Causal Effect Identifiability

In this section, we prove each intervention can orient at least one undirected ancestor edge in Theorem 2. Then, combining with Theorem 1, the identifiability of ancestor causal structure and the causal effect of each variable on response variable are directly concluded in Corollary 1. A supporting lemma is given at first, with the proof in Appendix D.2.

Lemma 1. *Under Assumption 1 and our intervention variable selection strategy, if X is intervened in chain component C, $Y \notin C$, undirected edges exist between $\mathrm{Anc}_Y[\ddot{C}]$, then all the undirected edges between X and the "next" variable located in the shortest undirected path $\mathcal{P}$ from X to any variable $Z \in \mathrm{Anc}_Y[\ddot{C}]$ can be identified by the intervention, the "next" variable is the one adjacent to X in $\mathcal{P}$.*

Theorem 2. *Under Assumption 1 and our intervention variable selection strategy, each intervention can orient at least one undirected ancestor edge.*

Proof. For a chain graph G to be oriented, there are three possible conditions. 1. $Y \in C$; 2. $Y \notin C$ and no undirected edges between $\mathrm{Anc}_Y[\ddot{C}]$; 3. others. Denote the intervention variable by X_i. In condition 1, the undirected edge $X_i - Y$ can be identified by whether Y changes under intervention. In condition 3, since the intervention variable has at least one directed path to Y, its undirected edges are ancestor edges. Lemma 1 indicates that some of them can be identified. In condition 2, if there is only one undirected path from X_i to $\mathrm{Anc}_Y[\ddot{C}]$, the edge between them can be identified as condition 1. Otherwise, the edge between X_i and its sibling which belongs to $\mathrm{Anc}_Y[\ddot{C}]$ is identifiable as condition 3. □

Corollary 1. *The ancestor causal structure and the causal effect of each variable on Y are identifiable by up to p interventions when there are $p+1$ variables $X_1, \cdots, X_p, Y$.*

If there are some ancestor edges unidentified, we select the intervention variable and infer the structure with the interventional data until discovering them all. Hence Corollary 1 holds. Moreover, Lemma 1 implies if the manipulated variable has m different "next" variables in all the shortest undirected paths to other variables in $\mathrm{Anc}_Y[\ddot{C}]$, at least m edges can be identified. This observation guides us to intervene on the variable with maximum siblings, since it may have more different "next" variables and thus identify more undirected edges. That is exactly what we do in intervention variable selection criterion to pursue the second goal.

5.2. Intervention Cost Analysis

In this part, we analyze the number of interventions to make causal effect identification in running our algorithm, which is the cost we hope to reduce. According to Corollary 1, p is an upper bound for the intervention cost. For example, if the causal structure is $X_1 \to \cdots \to X_p \to Y$, the cost is p. Although it is high in the worst case, our method can reduce the cost in common situations. We analyze two common skeletons. One is random line skeleton. The other is random complete skeleton. The results are in the following two propositions, with detailed proofs in Appendix D.3.

Proposition 4. *For a line skeleton with $p+1 \geq 4$ variables $X_1, \cdots, X_p, Y$, if all the causal relations and positions of variables $X_1, \cdots, X_p, Y$ are totally random, then the expected number of interventions to make causal effect identification is $\frac{19}{8} - \frac{39}{8p+8} + \frac{6}{p+1}(\frac{1}{2})^p < 3$.*

Proposition 5. *For a complete skeleton with $p+1 \geq 4$ variables $X_1, \cdots, X_p, Y$, if all the causal relations and positions of variables $X_1, \cdots, X_p, Y$ are totally random, then the expected number of interventions to make causal effect identification is less than $\frac{5}{6}(p+1) - \frac{11p-10}{6p+6} + \ln\frac{p}{2}$.*

Proof sketch: In a complete graph, there is an exact causal order for $p+1$ variables, in which Y is located in each position with the same possibility. We assume the order is $X_1, X_2, \cdots, X_i, Y, X_{i+1}, \cdots, X_p$. Each variable X_j after Y costs one intervention to identify its edge with Y. For the subgraph induced by Y and all its ancestors $X_1, \cdots, X_i, Y$, the expected number of experiments $F(i)$ is $F(i) = \frac{2}{3}p + \frac{1}{6} + \frac{2}{3}\sum_{i=3}^{p}\frac{1}{i}, i \geq 3$. Hence the expected number $C(p)$ to make causal effect identification is $C(p) = \frac{1}{p+1}\sum_{i=0}^{p}(p - i + F(i))$, which concludes the upper bound. □

In Appendix D.3, we show more analyses. In a complete graph where Y is not ancestor of any other variables, ancestor causal structure is the causal graph. By Eberhardt (2007), the expected number of interventions to make causal effect identification is $\frac{2}{3}(p+1) - \frac{1}{3}$ by taking singleton hard interventions and identifying the causal relations by just *whether* the distribution of other variable changes when *whole* variables are observable, while it is $\frac{2}{3}(p+1) - \frac{2}{3} + \ln\frac{p}{2}$ by our approach observing *only* Y. Both their ratios to the variable number are $\frac{2}{3}$ when $p \to \infty$. Hence our approach achieves the same efficiency by exploiting *how* the distribution of Y changes. In a general complete causal graph, the ratio by our approach converges to $\frac{5}{6}$, while it is $\frac{1}{3}$ for the full observations setting. The gap is because when we intervene on a variable not belonging to Anc_Y, the distribution of Y remains. The information revealed by such intervention is very limited compared with fully observable scenario.

6. Experiments

In this section, we apply our approach on both synthetic datasets and real-world data. The code is developed based on R package "pcalg" (Kalisch et al., 2012).

6.1. Simulations

In this part, our approach is compared to two active intervention strategies "Entropy[4]" proposed by He & Geng (2008) and "OPTSINGLE" proposed by Hauser & Bühlmann (2014), as well as RESIT (Peters et al., 2014) and LiNGAM (Shimizu et al., 2006)-two causal discovery approaches based on observational data. Two degenerated versions of ACI are also compared. The only difference between them and ACI is in the part of intervention variable selection. In the chain component with directed paths to Y,

(1) *Legal-random*: determines a set as Fig. 2 and randomly selects the intervention variable in it;

(2) *Max-siblings*: selects the variable with maximum siblings as the intervention variable.

Although only the interventional data of Y is observed in our setting, in order to reflect the *efficiency* rather than mere *effectiveness* of our method, we allow some setting relaxations in the compared methods. We list the detailed input information of all the compared approaches in Table 1.

Table 1. The input information of all approaches.

Approach	Interventional data
RESIT	None of interventional data
LiNGAM	None of interventional data
ACI	only Y observed
Max-siblings	only Y observed
Legal-random	only Y observed
Entropy	adapted to only Y observed
OPTSINGLE	adapted to only Y observed
Entropy-full	full variables observed
OPTSINGLE-full	full variables observed

Because ancestor causal structure discovery is sufficient and necessary for causal effect identification, we evaluate on the number of *newly* identified ancestor edges, which have not been identified in the essential graph. More identified ancestor edges usually imply more identified causal effects.

At first, we give an example to illustrate the process of identifying ancestor causal structure by various methods in Fig. 3. The non-linear data generating process is shown in Appendix E. When we intervene, the intervention value is set to the mean value of the intervention variable in observational data. The red node is the response variable. And the ancestor causal structure is the part with solid lines in Fig. 3(f), where the solid lines denote ancestor edges. The essential graph Ess and the chain components are shown in Fig. 3(a). There are two chain components in total. The experiments are conducted based on Ess. By RESIT and LiNGAM, undirected edges in Ess are inferred. Fig. 3(b) and Fig. 3(c) show the results. The blue and red edges are the correctly and wrongly identified edges, respectively. They are both classical methods. But in such complex setting, we can see it is hard to discover the causal graph based on only observational data by these approaches. Fig. 3(d) depicts the experiment processes by Entropy-full and OPTSINGLE-full. When they focus on chain component 1, the variable selected at first by both of them is labeled by number 1. By utilizing the post-interventional data of *full* variables, the blue edges are identified. So is the second intervention. Fig. 3(e) is by our approach. The ancestor causal structure is identified by only once intervention and only the data of response variable. Although Entropy-full and OPTSINGLE-full have the information of full variables, they still possibly need more interventions to achieve the same goal with ours.

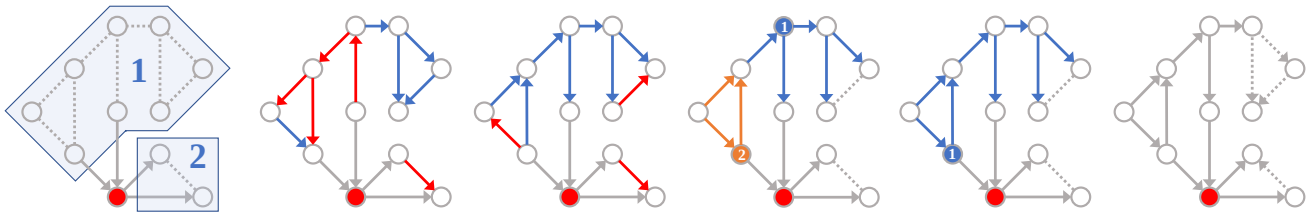
Then, we conduct a simulation to evaluate the effectiveness and efficiency of our method. We generate 100 linear structural equation models with the number of variables $p = 30$ and noise $\epsilon \sim \mathcal{N}(0, \mathbf{1}_p)$. The variable with the maximum degree is set to be the response variable. For each model, we generate 2500 samples as observational data. Our experiments begin from the essential graph. In each intervention, we have 1000 samples with the intervention value set to 2.

We evaluate the number of newly identified ancestor edges under different intervention time restrictions. For compared methods, if some unidentified edges remain after using up the intervention times, we orient them randomly. The results are in Fig. 4. RESIT and LiNGAM are exceeded by other approaches. The reason is the linear-Gaussian setting is not suitable for them, which indicates in general, intervention is necessary for our mission. Also, the comparison results with Entropy and OPTSINGLE imply the efficiency of our approach and considering *how* the distribution of the response variable changes is important given only response variable is observable. Besides, the superiority to the two degenerated approaches implies the effectiveness of our active strategy.

6.2. Application to Real Data

In this part, we apply our approach on a dataset used in causal discovery with both observational and interventional data (Sachs et al., 2005). It consists of 7466 measurements of the abundance of phosphoproteins and phospholipids recorded under different experimental conditions in primary

[4]Two strategies are proposed by them and achieve similar results. We thus only show that of the strategy "maximum entropy".



(a) Essential graph (b) RESIT (c) LiNGAM (d) Entropy-full & OPTSINGLE-full (e) ACI (f) Causal graph

Figure 3. (a) depicts the two chain components in essential graph. (b)-(e) imply the identified causal relations by different approaches, where blue edges or yellow edges denote the edges correctly identified in different stages, while red edges denote the edges wrongly identified. The last figure shows the ground-truth causal graph, where the solid lines denote the ancestor edges.

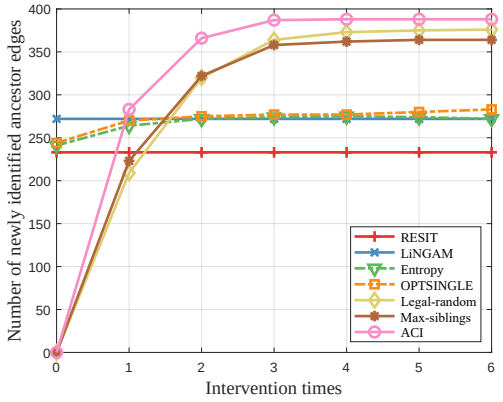


Figure 4. The number of newly identified ancestor edges.

human immune system cells. After processing, 5846 measurements remain. Refer to Appendix E for more information about the dataset and the causal graph.

Table 2. The regression coefficients.

Variable	Raf	Mek	PLCG	PIP2	PIP3
Coefficient	-0.024	0.052	0.008	0.023	-0.012
	Akt	PKA	PKC	p38	JNK
	1.016	0.077	0.003	-0.023	-0.004

Table 3. The number of newly identified ancestor edges.

Approach	GIES	IGSP	Entropy-full	OPTSINGLE-full	ACI
#ancestor edges	1	2	2	2	3

We set "Erk", which is the variable at the end of causal order, as the response variable. First, we run a linear regression between "Erk" and other variables. The coefficient of each regressor is shown in Table 2. The coefficient of "Akt" is far beyond others in the predictive model. But in the causal graph, we know "Akt" has no causal effect to "Erk", while "Mek" has a significant influence to "Erk", which can be identified by our approach based on causality. It indicates the necessity of considering causality in decision-making.

Next, our approach is compared to Entropy-full (He & Geng, 2008), GIES (Hauser & Bühlmann, 2012), OPTSINGLE-full (Hauser & Bühlmann, 2014) and IGSP (Wang et al., 2017). Both GIES and IGSP are towards causal discovery based on observational and interventional data. And we evaluate the number of newly identified ancestor edges. For fairness, GES algorithm is used in all cases to get an estimated essential graph. To simulate the active process, we then select the intervened variable according to our strategy and take corresponding interventional data of Y in. If the required interventional data is not in the dataset, we refer to the ground truth edges of the intervened variable to orient the PDAG. These "copying" edges will be also oriented to the results of other approaches and not be counted into our evaluation. For the compared approaches, we allow them to have the data of full variables under the intervention our approach takes. The results are shown in Table 3. It indicates ACI can find causal relations more efficiently in real tasks.

7. Conclusion

In this paper, we propose a method for causal effect identification when only response variable is observable under intervention. Our approach begins from a Markov equivalence class. We design a strategy to intervene and infer the causal relations by the post-interventional data of response variable until identifying the causal effect of each variable on the response variable. We verify the effectiveness and efficiency of our method theoretically and empirically.

Acknowledgements

This research was supported by the National Key R&D Program of China (2018YFB1004300), NSFC (61921006), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

Authors want to thank Kun Zhang, Yang-Bo He, Peng Zhao, Jia-Lve Chen, Zhi-Hao Tan and Tian Qin for insightful discussions, and thank reviewers for helpful comments.

References

Andersson, S. A., Madigan, D., Perlman, M. D., et al. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Cai, R.-C., Qiao, J., Zhang, K., Zhang, Z.-J., and Hao, Z.-F. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, pp. 2666–2674, 2018.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3: 507–554, 2002.

Cooper, G. F. and Yoo, C. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 116–125, 1999.

Eberhardt, F. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, pp. 93, 2007.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Hauser, A. and Bühlmann, P. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4): 926–939, 2014.

He, Y.-B. and Geng, Z. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

He, Y.-B. and Geng, Z. Causal network learning from multiple interventions of unknown manipulated targets. *CoRR*, abs/1610.08611, 2016.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pp. 689–696, 2008.

Huang, B.-W., Zhang, K., Lin, Y.-Z., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM International Conference on Knowledge Discovery & Data Mining*, pp. 1551–1560, 2018.

Huang, Y.-M. and Valtorta, M. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*, 2006.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2981–2989, 2019.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Kocaoglu, M., Dimakis, A., and Vishwanath, S. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1875–1884, 2017.

Kocaoglu, M., Jaber, A., Shanmugam, K., and Bareinboim, E. Characterization and learning of causal graphs with latent variables from soft interventions. In *Advances in Neural Information Processing Systems*, pp. 14346–14356, 2019.

Lauritzen, S. L. and Richardson, T. S. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64 (3):321–348, 2002.

Lee, S., Correa, J. D., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, pp. 144, 2019.

Meek, C. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11st Annual Conference on Uncertainty in Artificial Intelligence*, pp. 403–410, 1995a.

Meek, C. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the 11st Annual Conference on Uncertainty in Artificial Intelligence*, pp. 411–418, 1995b.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27): 7361–7368, 2016.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Perkovic, E., Textor, J., Kalisch, M., and Maathuis, M. H. A complete generalized adjustment criterion. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pp. 682–691, 2015.

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

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

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

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pp. 1219–1226, 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT Press, 2000.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence and 14th Conference on Innovative Applications of Artificial Intelligence*, pp. 567–573, 2002.

Triantafillou, S. and Tsamardinos, I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

Wang, Y.-H., Solus, L., Yang, K., and Uhler, C. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pp. 5824–5833, 2017.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Zhang, K., Huang, B.-W., Zhang, J.-J., Glymour, C., and Schölkopf, B. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1347–1353, 2017.

Zhang, K., Schölkopf, B., Spirtes, P., and Glymour, C. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2018.