

Efficient Causal Structure Learning from Multiple Interventional Datasets with Unknown Targets

Yunxia Wang¹, Fuyuan Cao^{1,*}, Kui Yu², Jiye Liang¹

¹School of Computer and Information Technology, Shanxi University, Taiyuan 030006, P.R. China

²School of Computer and Information, Hefei University of Technology, Hefei, 230009, P.R. China
wangyx_cloud@163.com, cfy@sxu.edu.cn, yukui@hfut.edu.cn, ljj@sxu.edu.cn

Abstract

We consider the problem of reducing the false discovery rate in multiple high-dimensional interventional datasets under unknown targets. Traditional algorithms merged directly multiple causal graphs learned, which ignores the contradictions of different datasets, leading to lots of inconsistent directions of edges. For reducing the contradictory information, we propose a new algorithm, which first learns an interventional Markov equivalence class (I-MEC) before merging multiple graphs. It utilizes the full power of the constraints available in interventional data and combines ideas from local learning, intervention, and search-and-score techniques in a principled and effective way in different intervention experiments. Specifically, local learning on multiple datasets is used to build a causal skeleton. Perfect intervention destroys some possible triangles, leading to the identification of more possible V-structures. And then a theoretically correct I-MEC is learned. Search and scoring techniques based on the learned I-MEC further identify the remaining unoriented edges. Both theoretical analysis and experiments on benchmark Bayesian networks with the number of variables from 20 to 724 validate that the effectiveness of our algorithm in reducing the false discovery rate in high-dimensional interventional data.

Introduction

Causal relationship discovery plays an irreplaceable role in various fields, including biology, epidemiology, medicine and economics (Pearl and Mackenzie 2018; Schölkopf et al. 2021). Therefore, learning a causal model described the relations among variables is important, which is in the form of a directed acyclic graph (DAG) (Yu et al. 2019; Xun et al. 2020). Many algorithms are proposed from observational and/or experimental data. Since different causal DAG models can generate the same observational distribution, a DAG is in general only identifiable up to its Markov equivalence class (MEC) (Chickering 2002) from observational data (Hauser and Bühlmann 2012; Shohei et al. 2006; Tsamardinos, Brown, and Aliferis 2006). The availability of interventional (experimental) data opens up new opportunities to reduce the size of the equivalence class down, possibly to recover the true causal graph (Ghassami et al. 2017; Peters, Bhlmann, and Meinshausen 2016; Meinshausen et al.

2016; M, B, and Turner R 2018). That is, one can distinguish the corresponding causes and effects by the idea of intervention, which is also the unique advantage of causality (Huang et al. 2020; Zhang et al. 2017).

Usually, one only knows that a dataset is interventional, but does not know which variables are intervened in this dataset (Bareinboim and Pearl 2016; Yu et al. 2020). For example, in molecular biology, the effects of various added chemicals to the cell are not set to one specific value and are not also precisely known. Also gene knockout technologies are known to have off-target effects, i.e., the CRISPR-Cas gene-editing technology performs cleavage at unknown genome sites other than their intended target (Antonia, Wendell, and Lei 2016; Wu et al. 2015). Facing different intervention experiments, if the intervention targets are forcibly known or not accounting for these additional targets while learning a causal structure, it may lead to incorrect conclusions in the learned causal DAG. In addition, compared with the known intervention targets setting, the unknown one requires a separate treatment since it is certainly less informative. Therefore, learning a causal inference algorithm effectively that can make full of interventional data under the unknown intervention targets is the purpose of the present paper. Here, we focus on reducing the number of inconsistencies produced by constraint-based methods.

There are algorithms are proposed for causal structure learning from multiple experimental datasets with unknown intervention targets. He and Geng (He and Geng 2016) adopted constraint-based algorithms (Spirtes, Glymour, and Scheines 2000) to learn each causal DAG from each interventional dataset, and then merged the graphs directly to get the final structure. However, due to the influence of sample selection bias and data noise, the idea ignores the contradictory information of multiple structures learned, resulting in large indeterminacies in the estimated graph. In addition, it learns each causal DAG from each dataset, which also leads to a higher time complexity. Squires et al. (Squires, Wang, and Uhler 2020) proposed the UT-IGSP algorithm to estimate causal DAG models from a mix of observational and interventional data, when the intervention targets are partially or completely unknown. Brouillard et al. (Brouillard et al. 2020) proposed a general continuous-constrained method for causal discovery which can leverage various types of interventional data as well as expressive neural architectures.

*Corresponding author: Fuyuan Cao. Email: cfy@sxu.edu.cn.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The two algorithms are proposed based on structural equation model, which are easy to accumulate errors in the inference process when learning the causal relations among hundreds of variables.

Recently, Mooij et al. (Mooij, Magliacane, and Claassen 2020) proposed a novel Joint Causal Inference (JCI) framework, which can be used to adapt an existing observational causal inference algorithm into a method for causal structure learning from interventional data with unknown targets. However, facing discrete data with unknown intervention targets, context variables are not particularly suitable to be added to the JCI framework. Although JCI can find different context variables corresponding to different datasets, the unknown intervention targets under perfect intervention will reduce the effect of context variables on system variables, resulting in that the effect of causal inference cannot be effectively improved. To take an extreme example, assuming that context variables affect all intervened variables, and then the performance of JCI does not increase effectively.

Thus, to battle the challenging issues, we propose a new algorithm, called EMIDGS (Multiple Interventional Datasets for Efficient Global causal Structure learning). It combines local learning, intervention characteristics and search-and-score techniques in a unified way, and considers the interventional information of multiple datasets simultaneously. Specifically, first, EMIDGS learns a causal skeleton by using a local learning algorithm proposed in this paper, called EMIDPC, which learns parents and children (PC) of a given variable from multiple interventional datasets simultaneously. Then EMIDGS orients the edges of the skeleton using two steps: (1) finding an Interventional Markov Equivalence Class (I-MEC) (Yang, Katcoff, and Uhler 2018) by two lemmas; (2) based on the learned I-MEC, applying a score function to greedily search each causal DAG from each dataset in the remaining search space, and then merging the graphs to obtain the final structure.

Theoretical analysis proves that the skeleton identification phase is sound and the EMIDPC algorithm is robust. The first step of the orientation phase provides theoretical guarantee, obtaining the correct directions of edges without selection bias and data noise. For the second step of the orientation phase, search-and-score techniques based on I-MEC from each dataset greatly reduces the search space, and the combination of multiple graphs further identifies the unoriented edges, improving the indeterminacies of the estimated structure. Therefore, our proposed algorithm reduces the contradictions of multiple causal structures, and experimental results also verify the effectiveness.

The paper is organized as follows. Section 2 gives preliminaries work. Section 3 proposes our new algorithm. Section 4 presents the theoretical analysis and complexity analysis. Section 5 describes and discusses the experiments and Section 6 concludes the paper and presents future work.

Preliminaries Work

Causal DAG Model

Let $G = (V, E)$ represent a directed acyclic graph (DAG) and P be the joint probability distribution over a ran-

dom vector $X_V = \{X_1, \dots, X_n\}$. Each node $j \in V = \{1, \dots, n\}$ is associated with a random variable X_j and each edge $(i, j) \in E$ represents a direct causal relation from X_i to X_j , i.e. $X_i \rightarrow X_j$ in a causal DAG denotes that X_i is a direct cause of X_j . For simplicity, we do not distinguish between V and X_V . The distribution P is Markov to the graph G , which makes the joint probability P can be decomposed into the product of conditional probabilities as

$$P(V) = P(X_1, \dots, X_n) = \prod_{X_j \in V} P(X_j | pa(X_j)) \quad (1)$$

where $pa(X_j)$ denotes the set of parents of X_j . A DAG G and a joint distribution P are faithful to each other, which enables us to recover a DAG G from a distribution P . In addition, we use $X_i \perp\!\!\!\perp X_j | S$ and $X_i \not\perp\!\!\!\perp X_j | S$ to represent that given S , X_i is conditionally independent of and dependent on X_j , respectively.

Post-intervention DAG

Let $D = \{D_1, \dots, D_m\}$ be the m experimental datasets. For $\forall i \in \{1, \dots, m\}$, let $R_i \subseteq V$ be the set of variables manipulated in the i -th experiment and $do(R_i)$ denote the intervention on the set of variables R_i (Pearl 2009). After the intervention on R_i in the i -th experiment, the post-intervention DAG of G is $G_i = (V, E_i)$ where $E_i = \{(a, b) | (a, b) \in E, b \notin R_i\}$. The joint distribution of the post-intervention DAG G_i with respect to R_i can be written as

$$P_i(V | do(R_i)) = \prod_{X_j \in V \setminus R_i} P(X_j | pa(X_j)) \prod_{X_j \in R_i} P_i(X_j) \quad (2)$$

where $P(X_j | pa(X_j))$ is the same as the conditional probability of X_j in Eq. (1) and $P_i(X_j)$ is the post-intervention conditional probability of X_j after X_j is manipulated.

Let $R = \bigcup_{i=1}^m R_i$. If $\exists X_j \in R$ and $\exists R_i \in R$ such that $X_j \notin R_i$, then R is conservative, called conservative rule (Pearl 2009). The conservative rule states that given m intervention experiments, if for any manipulated variable X_j , one can always find an experiment in which X_j is not manipulated. With the definition of this conservative rule, we can theoretically analyze the possibility of learning a causal structure under unknown intervention targets.

Learning Causal Structure from Multiple Interventional Datasets with Unknown Targets

In this section, we introduce EMIDGS proposed in this paper. First, EMIDGS assumes that there are no unmeasured confounders, and faithfulness is also assumed in the paper as Assumption 1. In addition, EMIDGS assumes that intervention is perfect and intervention targets are unknown. Perfect intervention means that the causal relations between the manipulated variable and its direct causes are completely eliminated. Lastly, EMIDGS assumes R is conservative as Assumption 2.

Assumption 1 *The joint probability P_i is faithful to the DAG G_i for any $i \in \{1, \dots, m\}$.*

Algorithm 1: The EMIDGS algorithm.

```

1: Input:  $D = \{D_1, \dots, D_m\}$ :  $m$  interventional datasets;
    $V = \{X_1, \dots, X_n\}$ :  $n$  variables.
2: Output:  $G_f$ .
   /*Phase 1: Building a skeleton  $G_s$  */
3:  $G = G_{out} = (|V|, |V|)$ ;
4: for  $X_j \in V$  do
5:    $[pc(X_j), sep, kpc, kindep] = EMIDPC(D, X_j, V)$ ;
6:   for  $Y \in pc(X_j)$  do
7:      $G(X_j, Y) = 1$ ;
8:   end for
9: end for
10:  $G_s = (V, E_s) = G$ ;
    Phase 2: Orienting edges
11: for  $A \in V$  do
12:   for  $X, Y \in pc(A)$  do
13:     for  $i = 1$  to  $m$  do
14:        $S = sep_X(Y) \cup sep_Y(X)$ 
15:       if  $X \not\perp\!\!\!\perp Y \mid S, X \perp\!\!\!\perp Y \mid \{S \cup A\}$  in  $D_i$  then
16:          $G(X, A) = -1; G(Y, A) = -1$ ;
17:          $G(A, X) = 0; G(A, Y) = 0$ ;
18:          $tem = kindep_A(i) \cap pc(A)$ ;
19:          $G(A, tem) = -1; G(tem, A) = 0$ ;
20:       end if
21:     end for
22:   end for
23:   updating  $G$  by Meek rules;
24: end for
25:  $G_0 = (V, E_0) = G; G_{out} = G_0$ ;
26: for  $i = 1$  to  $m$  do
27:   Based  $G_0$ , learning a graph  $G_f(i) = (V, E_i)$  from
     the  $i$ -th dataset  $D_i$  by performing a scoring method
     on the remaining search space.
28: end for
29: Combine  $G_f(i)$  to a graph  $G_{out}$ .
30: return  $G_{out}$ 

```

Assumption 2 R is conservative in m interventional datasets.

EMIDGS (Algorithm 1) first reconstructs the skeleton of a causal DAG by learning parents and children of each node from multiple interventional datasets, which is achieved by a subsection: EMIDPC (Multiple Interventional Datasets for Efficient PC discovery, illustrated in Algorithm 2) in Section 3.1. After obtaining the skeleton, EMIDGS orients edges by two lemmas as Lemmas 1 and 2 and outputs results in Section 3.2.

Building a Skeleton

For each variable X_j in V , EMIDGS looks for its parents and children from m datasets through EMIDPC, and then connects them by applying OR rules. After testing all variables in V , EMIDGS gets a causal skeleton G_s as lines 3-10 in Algorithm 1.

EMIDPC (Algorithm 2) is proposed in this paper to find parents and children (PC) of a given variable T included in V from m interventional datasets with unknown targets. It is

implemented in two steps. Suppose that $canpc(T)$ keeps the candidate parents and children of T , storing all variables dependent on T conditioned on an empty set from all datasets, and $pc(T)$ denotes the set of true parents and children of T . Step 1 gets $canpc(T)$ from m interventional datasets. Step 2 removes false positives from $canpc(T)$ to get $pc(T)$ and outputs results. In addition, suppose that $kpc_T(i)$ stores the variables that depend on T conditioned on an empty set in the i -th dataset. And similarly, suppose $kindep_T(i)$ stores variables that are independent of T in the i -th dataset, where possible parents or children of T are included due to the characteristics of the intervention.

Step 1: Find candidate PC of T (lines 5-16). EMIDPC judges the dependence of $X_j \in V \setminus T$ and T conditioned on an empty set in each interventional dataset. If $X_j \not\perp\!\!\!\perp T$ holds in the i -th dataset, EMIDPC adds X_j to $canpc(T)$ and $kpc_T(i)$, otherwise it adds X_j to $kindep_T(i)$. The next variable is considered until X_j has been tested in all datasets. So when a variable X_j is independent of T in m datasets, it means that the variable must not be a parent or a child of T . And the conclusion that the true PC set must be included in $canpc(T)$ is believed. Next, EMIDPC needs to remove false positives from $canpc(T)$ as much as possible.

Step 2: Find PC of T (lines 17-45). EMIDPC removes false positives from $canpc(T)$ by the standard forward-backward strategy (SFBS) as shown in lines 17-33 of Algorithm 2. First, it sets $cpc(T)$ an empty set and selects the feature $X \in canpc(T)$, which has the highest association with T , and then adds X into $cpc(T)$ and removes X from $canpc(T)$. Next, EMIDPC determines whether the variable X just added to $cpc(T)$ is a false positive or not. False positives are those non descendants excluding parents and those descendants excluding children. However, the challenge is that EMIDPC does not know which variables are manipulated in each dataset. So the situations about removing false positives are discussed as follows.

For non descendants excluding parents, because R is conservative, we can always find at least one dataset, denoted as the i -th dataset (D_i), in which T is not intervened and then the parents of T can be found. If parents of T are also not intervened in this dataset, the non descendants excluding parents can be removed. On the contrary, if one parent of T in the i -th dataset is intervened, we cannot find the corresponding non descendants and further analysis is discussed. If T is intervened in a certain dataset, such as the j -th dataset, T is independent of a T 's non-descendant X and X is not included in $pa(T)$, and then whether X is a T 's non-descendant in D_j cannot be determined. For those descendants excluding children, whether T is intervened or not does not influence the removal of its descendants excluding children, but the removal is affected by whether T 's children are intervened. If a child of T is intervened in the i -th dataset, we cannot find the descendants excluding children in $kpc_T(i)$. If a child of T , denoted Y , is not intervened in the j -th data set, two situations are discussed. One is that if the children of Y , that is, the descendants excluding children of T , are intervened in the j -th dataset, these nodes are not in $kpc_T(j)$. The other is that if the children of Y are not intervened in the dataset, we can remove them through the separation set $\{Y\}$.

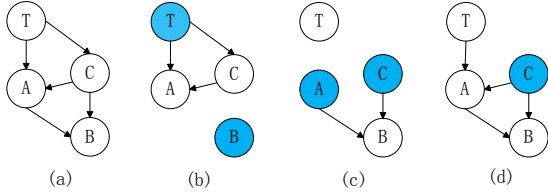


Figure 1: Four causal DAGs, where (a) is the underlying DAG; and (b), (c) and (d) are three post-intervention DAGs after the manipulation on the some variables in (a).

Corollary 1 Referring to Algorithm 2, assuming $X_j \in cpc(T)$ and $\exists S \subseteq \{cpc(T) \setminus X_j\}$, if $\exists k \in \{1, \dots, m\}$ such that $X_j \perp\!\!\!\perp T|S$ and $\{X_j \cup S\} \subseteq kpc_T(k)$ in D_k , then we know that $X_j \notin pc(T)$.

Using Corollary 1 to remove false positives in $canpc(T)$ and get $cpc(T)$. However, one disadvantage of the above is that $cpc(T)$ learned is a superset of $pc(T)$. That is to say, some false positives are still in $cpc(T)$. For example, as shown in Figure 1, B is still not removed from $cpc(T)$ in three interventional datasets corresponding to Figure 1 (b), (c) and (d). Specifically, we learn that $kpc_T(1) = \{A, C\}$, $kpc_T(2) = \emptyset$, $kpc_T(3) = \{A, B\}$, and then $canpc(T) = \{A, C, B\}$. First, assume that $A \in canpc(T)$ has the highest association with T , which is added to $cpc(T)$, that is $cpc(T) = \{A\}$. Next, assume that $B \in canpc(T)$ is the next variable with the highest association with T , which is also added to $cpc(T)$ and $cpc(T) = \{A, B\}$. Since we do not find a set $Z \subseteq cpc(T) \setminus B$ such that $B \perp\!\!\!\perp T|\{Z\}$ holds in the dataset (i.e., D_3), where $\{B \cup Z\} \subseteq kpc_T(3)$, $cpc(T)$ remains unchanged. In other words, in D_3 , $T \not\perp\!\!\!\perp B|\phi$ and $T \not\perp\!\!\!\perp B|A$, so B cannot be removed from $cpc(T)$. Lastly, the variable C is added to $cpc(T)$ and $cpc(T) = \{A, B, C\}$. Since there is not a set $Z \subseteq cpc(T)$ and $C \cup Z \subseteq kpc_T(i)$, $i \in \{1, 2, 3\}$ such that $C \perp\!\!\!\perp T|\{Z\}$ holds, $cpc(T)$ still remains unchanged and $cpc(T) = \{A, B, C\}$.

Therefore, as shown in lines 34-44 of Algorithm 2, EMIDPC removes the false positives from $cpc(T)$ learned directly by changing the length of the conditioning set S from 1 to $|cpc(T)|$ to obtain the theoretically correct parents and children of T . We define the length of S as $temp$, and initialize $temp$ to 1. When the length of S is less than $|cpc(T)|$, the false positives will be removed by Corollary 1. Until all $X_j \in cpc(T)$ are judged under the same length of S , EMIDPC considers the length of S with $temp + 1$. This process ends at a condition where $temp$ is greater than the length of $cpc(T)$. Finally, as lines 45-46, EMIDPC sets $pc(T) = cpc(T)$, and outputs $pc(T)$, kpc_T , $kindep_T$, sep_T .

Orienting Edges

The phase of edge orientation is divided into two steps as lines 11-29 in Algorithm 1. In step 1, EMIDGS orients edges to learn an I-MEC by the invariance of V-structures and the property of perfect intervention as Lemma 1 and Lemma 2. In step 2, based on the I-MEC learned in step 1, EMIDGS performs a score-and-search strategy in finite research space

Algorithm 2: The EMIDPC algorithm.

```

1: Input:  $D = \{D_1, \dots, D_m\}$ ;  $m$  interventional datasets;
    $T$ : the target variable;  $V = \{X_1, \dots, X_n\}$ ;  $n$  variables.
2: Output:  $pc(T)$ ,  $sep_T$ ,  $kpc_T$ ,  $kindep_T$ .
3:  $canpc(T) = \phi$ ;  $pc(T) = \phi$ ;
4:  $kpc_T = cell(1, m)$ ;  $kindep_T = cell(1, m)$ ;
5: for  $X_j \in \{V \setminus T\}$  do
6:   for  $i = 1$  to  $m$  do
7:     if  $X_j \not\perp\!\!\!\perp T$  in  $D_i$  then
8:        $canpc(T) = canpc(T) \cup X_j$ ;
9:        $kpc_T(i) = kpc_T(i) \cup X_j$ ;
10:       $Dep(X_j, i) = Dep(T, X_j)$ ;
11:     else
12:        $kindep_T(i) = kindep_T(i) \cup X_j$ ;
13:     end if
14:   end for
15:    $dep(X_j) = max(Dep)$ ;  $sep_T(X_j) = \phi$ ;
16: end for
17:  $cpc(T) = \phi$ ;
18: repeat
19:    $Y = argmax dep(X, T|\phi)$ ,  $X \in canpc(T)$ ;
20:    $cpc(T) = cpc(T) \cup Y$ ;  $canpc(T) = canpc(T) \setminus Y$ ;
21:   for  $X \in cpc(T)$  do
22:     for  $i = 1$  to  $m$  do
23:       if  $X \perp\!\!\!\perp T | S$ ,  $S \subseteq cpc(T) \setminus X$ ,  $S \subseteq$ 
          $kpc_T(i)$ ,  $X \in kpc_T(i)$  in  $D_i$  then
24:          $cpc(T) = cpc(T) \setminus X$ ;
25:          $sep_T(X) = S$ ;
26:         if  $\exists h \in \{1, \dots, m\}$  such that  $X \in kpc_T(h)$ 
           then
27:            $kpc_T(h) = kpc_T(h) \setminus X$ ;
28:         end if
29:         break;
30:       end if
31:     end for
32:   end for
33: until  $canpc(T)$  is empty
34:  $temp = 0$ ;
35: repeat
36:   for  $X \in cpc(T)$  do
37:     for  $i = 1$  to  $m$  do
38:       if  $\exists S \subseteq cpc(T) \setminus X$  and  $|S| = temp$ ,  $X_j \perp\!\!\!\perp T | S$ 
         in  $D_i$  and  $X_j \in kpc_T(i)$  and  $S \subseteq kpc_T(i)$  then
39:          $cpc(T) = cpc(T) \setminus X$ ;  $sep_T(X) = S$ ; break;
40:       end if
41:     end for
42:   end for
43:    $temp = temp + 1$ ;
44: until  $temp > |cpc(T)|$ 
45:  $pc(T) = cpc(T)$ ;
46: return  $pc(T)$ ,  $sep_T$ ,  $kpc_T$ ,  $kindep_T$ 

```

to learn m DAGs from m interventional datasets and then the m causal DAGs are combined to get the final structure.

Lemma 1 *The invariance of V-structures. Suppose R is conservative and there are three variables $X, Y, A \in V$ and m intervention datasets without data noise and selec-*

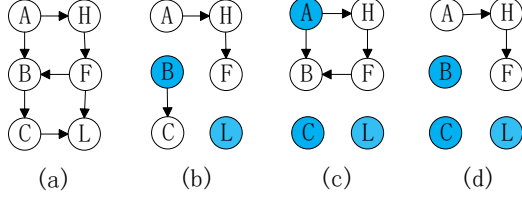


Figure 2: Four causal DAGs, where (a) is a underlying DAG; and (b), (c) and (d) are three post-intervention DAGs after the manipulation on the some variables in (a).

tion bias. If $\exists S \subseteq V \setminus \{X, Y, A\}$ such that $X \perp\!\!\!\perp Y | S$, $X \not\perp\!\!\!\perp Y | \{S \cup A\}$, and the variables in $\{X, Y\} \cup S$ are dependent on the variable A in the i -th dataset, then X, Y, A forms a V-structure with the collider A , that is $X \rightarrow A \leftarrow Y$.

Proof: Since R is conservative, if $\exists A \in V$ such that for $\exists R_i \in R$, $A \in R_i$ holds, then A is not manipulated in the i -th dataset (i.e., D_i). Thus the V-structure containing the collider A may be found in D_i if it exists. That is to say, if $\exists S \subseteq V \setminus \{X, Y, A\}$ such that $X \perp\!\!\!\perp Y | S$, $X \not\perp\!\!\!\perp Y | \{S \cup A\}$ in the i -th dataset, where the variables in $\{X, Y\} \cup S$ are dependent on A , then $X \rightarrow A \leftarrow Y$ is identified. In addition, if the all datasets have no data noise and selection bias, the V-structures learned from many manipulated datasets are same. On the contrary, if R is not conservative, there may be a collider A , which is intervened in all datasets. In other words, $\exists A \in V$ such that for $\forall R_i \in R$, $A \in R_i$ ($i \in \{1, \dots, m\}$) holds. And then the V-structure including the collider A cannot be found. For example, for the variable B in Figure 2 (a), since $A \perp\!\!\!\perp F | \{H\}$ and $A \not\perp\!\!\!\perp F | \{H, B\}$ hold in D_2 responding to Figure 2 (c), where $\{A, F\} \cup \{H\}$ are dependent on B , we can orient $A - B - F$ to $A \rightarrow B \leftarrow F$. However, for the variable L in Figure 2 (a), we find that L is intervened in all intervention datasets and then the V-structure including L as a collider cannot be determined.

Lemma 2 The unique property of perfect intervention. For $X, Y, A \in V$ and m interventional datasets without data noise and selection bias, assuming a triple $\langle X, Y, A \rangle$ is determined as a V-structure in the i -th dataset D_i and A is a collider. If $\exists A_1 \in pc(A)$ such that $A \perp\!\!\!\perp A_1$ in D_i , then A_1 is a child of A .

Proof: Assume that a triple $\langle X, Y, A \rangle$ is determined as a V-structure in the i -th dataset D_i and A is a collider, then the conclusion is concluded that A is not manipulated in D_i . If $\exists A_1 \in pc(A)$ such that $A \perp\!\!\!\perp A_1$ in D_i , which shows that A_1 is manipulated in D_i , then the conclusion that A_1 is a child of A is believed. As shown in Figure 2, we can discover the V-structure $A \rightarrow B \leftarrow F$ holds in D_2 responding to Figure 2 (c). So we judge that whether the variables in $pc(B)$ are independent to B or not in D_2 . We learn that $pc(B) = \{A, F, C\}$ and $C \in pc(B)$. So $\exists C \in pc(B)$ such that $C \perp\!\!\!\perp B$ holds in D_2 , that is, the collider B is not manipulated in D_2 but C is manipulated in D_2 . Then the conclusion is reached that C is a child of B .

Step1: Orient edges by Lemmas 1 and 2 (lines 11-24). EMIDGS orients the edges containing in the learned skele-

ton by two properties described in Lemmas 1 and 2, respectively. Some edges can be oriented by the invariance of V-structures as described in Lemma 1. EMIDGS first selects a variable $A \in V$ as the current variable. Then for every two variables belonging to $pc(A)$ (e.g., X, Y), EMIDGS judges the independence of X and Y conditioned on a set S and judges the dependence of X and Y conditioned on $\{S \cup A\}$ in each dataset. We choose $S = \{sep_X(Y) \cup sep_Y(X)\}$ as a set which makes X and Y independent. If the original structure connected with the V-structure is not destroyed, then $sep_X(Y)$ and $sep_Y(X)$ are identical theoretically. If only V-structures exist, then the conditioning set should have been an empty set, but X and Y are independent given $S = \{sep_X(Y) \cup sep_Y(X)\}$. In order to reduce the time complexity of EMIDGS, we still choose $\{sep_X(Y) \cup sep_Y(X)\}$ as a conditioning set. Therefore, if $X \perp\!\!\!\perp Y | S$ and $X \not\perp\!\!\!\perp Y | \{S \cup A\}$ occur in D_i ($i \in \{1, \dots, m\}$), then the relationship between X, Y and A is determined as $X \rightarrow A \leftarrow Y$.

Also, EMIDGS orients some edges by Lemma 2, which is the unique property of perfect intervention. Specifically, when EMIDGS determines that a triple $\langle X, A, Y \rangle$ is a V-structure by Lemma 1, it records the current dataset D_i . In D_i , EMIDGS learns that $tem = \{kindep_A(i) \cap pc(A)\}$, and concludes the conclusion that the variables in tem are the children of the collider A . In order to make full of the interventional property of a causal model, EMIDGS finds V-structures in all datasets and records the datasets with the V-structure, and then orients edges by Lemma 2.

Finally, Meek rules are applied as shown in line 23. Since variables in $ch(X)$ and X cannot form V-structures, if some parents of X point to X , and we can be sure that the remaining variables connected to X are the children of X . So Meek rules work in the process.

Corollary 2 Referring to Algorithm 1, if $X \perp\!\!\!\perp Y | S$ and $X \not\perp\!\!\!\perp Y | \{S \cup A\}$, $\exists S \subseteq V \setminus \{X, Y, A\}$ and $\{X, Y, A\} \cup S \subseteq kpc_A(i)$ in D_i ($i \in \{1, \dots, m\}$), then X, Y, A forms a V-structure with the collider A , orienting $\langle X, A, Y \rangle$ as $X \rightarrow A \leftarrow Y$.

Corollary 3 Referring to Algorithm 1, assuming $\langle X, A, Y \rangle$ is identified as a V-structure in the i -th dataset (D_i) and A is a collider, then the variables in the set $tem = kindep_A(i) \cap pc(A)$ are children of A .

Using Corollary 2 and Corollary 3 to determine the causal relationships among variables in V , EMIDGS learns an I-MEC, which includes more causal information.

Step2: Score-and-search, and merge (lines 25-29). Based on the I-MEC learned (i.e., G_0) by lines 11-24 in Algorithm 1, EMIDGS uses a scoring method to learn and search the graph with the highest score in the remaining graph space as the best causal DAG. Next, after getting m causal DAGs, EMIDGS combines these DAGs to the final graph G_f . Specifically, first, EMIDGS sets the final graph $G_{out} = G_0$, providing a higher priority to oriented edges of I-MEC. Then based on the I-MEC (i.e., G_0), the remaining unoriented edges are further tested. In other words, For an unoriented edge e_j in I-MEC (i.e. G_0), if the direction of e_j is determined in a certain graph $G_f(i)$, then the edge

e_j is oriented in G_{out} . If the direction of e_j is contradictory in k ($1 \leq k \leq m$) graphs, then the direction of e_j is not determined in G_{out} . According to the above conditions, EMIDGS judges all unoriented edges in G_0 and learns the merged graph G_{out} . Finally, EMIDGS outputs the final causal DAG G_{out} .

Theoretical Analysis and Complexity Analysis

In this section, we theoretically analyze and prove the correctness of our algorithm in Section 4.1. Also we analyze the computational complexity of our algorithm in Section 4.2.

Theoretical Analysis

We focus on discovering the causal relations of all variables. In this paper, we will have the three following propositions:

Proposition 1 *When R is conservative, the EMIDPC algorithm can learn parents and children of $T \in V$ correctly and efficiently.*

Proof: Lines 5-16 in Algorithm 2 (EMIDPC) show that if $X_j \not\perp T$ holds in any a datasets, EMIDPC considers X_j as a candidate PC variable for being added to $canpc(T)$. When assuming R is not conservative, true parents or children of T may be lost in the process. So when R is conservative in EMIDPC, lines 5-16 show that no false positives are removed from $canpc(T)$. That is to say, the true $pc(T)$ is included in $canpc(T)$.

Lines 17-33 in Algorithm 2 show that X_j just added in $cpc(T)$ is judged whether a false positive or not. If existing a set $S \subseteq cpc(T) \setminus X_j$, $S \subseteq kpc_T(i)$, and $X_j \in kpc(i)$ in D_i such that X_j and T are independent conditioned on S , then X_j is considered as a false positive. This process can filter out most of false positive nodes. In other words, $cpc(T)$ learned in lines 17-33 is a superset of the true $pc(T)$. And because the length of the conditioning set is as small as possible, it can effectively reduce the time complexity.

Lines 34-44 in Algorithm 2 show the size of the conditioning set S is from 1 to $|cpc(T)|$, which covers all of possible conditioning subsets, and ensures all false positives are removed theoretically.

Proposition 2 *When R is conservative, the EMIDGS algorithm can discover a theoretically correct I-MEC effectively, which includes more causal relationships.*

Proof: Lines 3-9 in Algorithm 1 (EMIDGS) show that a causal skeleton is reconstructed by applying OR rules to connect the edges between each node (variables) belonging to V and its PC nodes. So the learned skeleton is theoretically correct.

Lines 11-24 in Algorithm 1 orient edges of the learned skeleton by Lemma 1 and Lemma 2. Lines 13-17 of EMIDGS orient $X - A - Y$ to $X \rightarrow A \leftarrow Y$ by Lemma 1. If R is not conservative, that is, if the collider A is manipulated in all datasets, the edges pointing to A are lost and the V-structures including A as a collider are lost. Lines 18-19 orient some edges by Lemma 2, which destroys some possible triangles and identifies more V-structures. If R is not conservative, the correctness of Lemma 2 needs to be considered. Therefore, when R is conservative, EMIDGS obtains a

theoretically correct I-MEC by Lemmas 1 and 2 effectively, which includes more causal relationships.

Proposition 3 *When R is conservative, the combination of multiple graphs, that is the second step of the orientation phase of EMIDGS, can effectively reduce the search space and improve the efficiency of the EMIDGS algorithm.*

Proof: Based on the learned I-MEC in lines 11-24, lines 25-28 in Algorithm 1 (EMIDGS) learn m graphs $G_f(i)$ ($\forall i \in \{1, \dots, m\}$) with the highest score from m datasets on the remaining search space, which effectively reduces the search space. Multiple graphs in the search space are formed by the underlying I-MEC and the remaining unoriented edges with different orientations. Line 29 combines the m graphs to get the final graph G_{out} . For an undirected edge in I-MEC, if the direction of the edge is contradictory in multiple learned graphs, the direction of the edge is still not determined, which further improves the accuracy of the EMIDGS algorithm.

Computational Complexity

In the lines 5-16 of the EMIDPC algorithm (Algorithm 2), the complexity of checking variables in V in m datasets is $O(m|V|)$. At lines 17-33 of Algorithm 2, EMIDPC examines the subsets of $cpc(T)$ which is learned by adding the newly features from an empty set. And at lines 34-44, EMIDPC examines the subsets with the size of $1:|cpc(T)|$, where $cpc(T)$ is obtained from lines 17-33. Assuming that the size of $cpc(T)$ is p , the complexity of lines 17-44 is $O(m * |p| * (C(p, 1) + \dots + C(p, p)))$. In the best case, the size of the conditioning set is 1 and the complexity of EMIDPC is $O(|cpc(T)|^2 * m)$. And in the worst case, we need to search all subsets in $cpc(T)$, and the complexity of the algorithm is $O(|cpc(T)|^2 * 2^{|cpc(T)|} * m)$. Thus, the total complexity of EMIDPC in the worst case is $O(m|V| + |cpc(T)|^2 * 2^{|cpc(T)|} * m)$, and reduced to $O(2^{|cpc(T)|} |cpc(T)|^2 m)$.

In Algorithm 1, EMIDGS assumes that the time taken to discover parents and children of a given variable $X_j \in V$ is tpc . At lines 3 to 9, the complexity of reconstructing a skeleton is $O(|V|tpc)$. At lines 11 to 24, the complexity of orienting the edges recursively by employing Lemma 1 and Lemma 2 is $O(2^{|pc|} |V|m)$, where pc is the largest set of parents and children over all variables in V . In conclusion, the total complexity of EMIDGS in the worst case is $O(|V| * tpc + 2^{|pc|} |V|m)$, reduced to $O(2m|V|2^{|pc|} |pc|^2)$.

Experiments

In this section, we evaluate the performance of the proposed EMIDGS algorithm with the existing algorithms under different conditions. To our best knowledge, the only approach for finding a global causal structure from multiple high-dimensional interventional datasets is a graph-merging method proposed by He&Geng (He and Geng 2016), so we compare it with our algorithm. In addition, in order to enrich the experiments, we add two additional baseline algorithms. One is called baseline-MMHC, which learns a causal DAG from each interventional dataset by MMHC and then merges

Net	Nodes	Arcs	Parameters	Average MB	Type
child	20	25	230	3.00	Medium
insurance	27	52	984	5.19	Medium
alarm	37	46	509	3.51	Medium
win95pts	76	112	574	5.92	Large
munin	186	273	15622	3.81	Very Large
link	724	1125	14211	4.80	Very Large

Table 1: Description of benchmark BNs.

them. And another is called MIMB-GS. It is constructed to obtain a global causal DAG by extending MIMB (Yu, Liu, and Li 2019), which identifies a Markov blanket (MB) of a given variable from multiple datasets with unknown intervention targets and describes under what conditions one can identify the causes of a given variable. The core idea is as follows. The skeleton is reconstructed by the learned MB of each variable, and the causes of each variable can be identified by MIMB.

We select six Bayesian networks (BNs) with the number of nodes ranges from 20 to 724 as shown in Table ?? to conduct multiple types of experiments. In the first experiment, we randomly select q variables to manipulate where we set $q \in \{1, 2\}$, namely $mT = 2$, and make sure the m sets of intervention targets are conservative. Moreover, in the setting $mT = 2$, we conduct four simulations to generate four types of interventional datasets in each benchmark bayesian network. The first one is that we run 5 simulations to generate 5 corresponding post-intervention DAGs and probability distributions, getting five interventional datasets as a group, namely $mD = 5$. The remaining three settings are 10 interventional datasets as a group, 15 datasets as a group and 20 datasets as a group, and use $mD = 10, 15$ and 20 to represent, respectively. Similarly, in the remaining experiments about the setting of the number of intervention targets, we set that $mT = 4, 6, 8$ and 10, respectively. At the same time, in each experiment, we perform four types of simulations, namely $mD = 5, 10, 15$ and 20. For each simulation, we run each of them for 10 times, and each simulation generates mD datasets, that is, each simulation gets $10 * mD$ datasets. Each dataset contains 5000 samples. In all experiments, G^2 tests are used for the conditional independence tests. All experiments are conducted on a computer with an AMD Core A8-6410 2.00 GHz with 4GB RAM.

We evaluate and compare the performance of our proposed EMIDGS algorithm with other algorithms using the following metrics: SHD-normalized, reverseEdge-normalized, missEdge-normalized, extraEdge-normalized, F1, Precision, Recall and time, where the first seven indicators represent accuracy, and the latter indicator represents time efficiency. Among the seven indicators, the smaller the first four indicators, the higher the accuracy of the corresponding algorithm. On the contrary, the larger the latter three indicators, the higher the accuracy of the corresponding algorithm. For the first seven indicators representing algorithm accuracy, SHD-normalized and F1 are two indicators of aggregation. To present the experimental results clearly, for SHD-normalized, we in-

roduce reverseEdge-normalized, missEdge-normalized and extraEdge-normalized so that we can see what types of errors broken out each method makes. For F1, we introduce Precision and Recall to display the results clearly. The experimental results are shown in Figures 3 and 4. In Figures 3 and 4, six DAGs, i.e., “child”, “alarm”, “insurance”, “win95pts”, “munin”, “link” as the labels of horizontal axis, i.e., ‘C’, ‘A’, ‘In’, ‘W’, ‘M’, ‘L’.

In addition, we conduct experiments with different confidence levels of $\alpha = 0.01$ to $\alpha = 0.1$ in each causal network with fixed values of the other two parameters, i.e. $mT = 10$ and $mD = 5$. And we test $mT \in \{2, 4, 6, 8, 10\}$ under the conditions: $\alpha = 0.01$, $mD = 5$, and $mD \in \{5, 10, 15, 20\}$ under the conditions: $\alpha = 0.01$, $mT = 10$. And the following metrics are used: SHD-normalized, reverseEdge-normalized, missEdge-normalized, extraEdge-normalized and nTest. The experimental results are shown in Figures 5, 6, 7, 8, 9 and 10.

Lastly, we evaluate and compare the EMIDPC algorithm with the MIPC algorithm using the following metrics: F1, Precision, Recall and nTest. And the experimental results are shown in Figure 11.

(1) EMIDGS vs. baseline-MMHC and MIMD-GS

Figure 3 shows that the comparison of EMIDGS, MIMB-GS, and baseline-MMHC under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$. We observe that EMIDGS is significantly better than the other two algorithms in accuracy except “child”. Especially in large-sized networks, EMIDGS has a significant improvement. That is to say, the lower SHD-normalized and the higher F1 show that the effectiveness of EMIDGS in reducing the false discovery rate in high-dimensional data. In contrast, the baseline-MMHC algorithm directly integrates multiple graphs, accumulating more contradictory information especially in large-sized networks as “win95pts”, “munin” and “link”. For the MIMD-GS algorithm, the MIPC algorithm adopted by MIMB-GS is relatively inferior in accuracy, compared with EMIDPC proposed by this paper. In other words, the causal skeleton constructed by MIMB-GS is relatively inferior than EMIDGS.

In addition, Figure 3 shows that the time efficiency of the EMIDGS algorithm is relatively acceptable in the first five networks, but the time efficiency of EMIDGS is worse than that of the baseline-MMHC algorithm on “link”. In general, our algorithm does effectively reduce the contradictory information of multiple graphs when the time efficiency is acceptable, especially in large-sized networks.

(2) EMIDGS vs. the HeGeng algorithm

Figure 4 shows that the comparison of EMIDGS and the HeGeng algorithm under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$. Since the outputs of the “win95pts”, “munin”, “link” cannot be generated in 72 hours by the HeGeng algorithm, we only compare it with our algorithm in three medium-sized networks as shown in Figure 4. Figure 4 shows that EMIDGS has a greater improvement in efficiency and accuracy, compared with the HeGeng algorithm.

(3) Different values of three parameters

The results of the different values of the three parameters of EMIDGS on six networks as shown in Figures 5, 6,

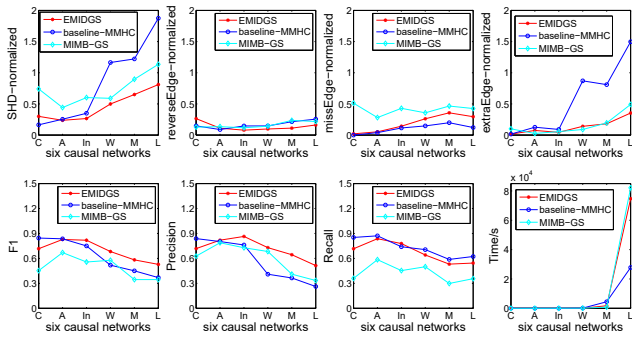


Figure 3: Comparison of EMIDGS, MIMB-GS, and baseline-MMHC under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$.

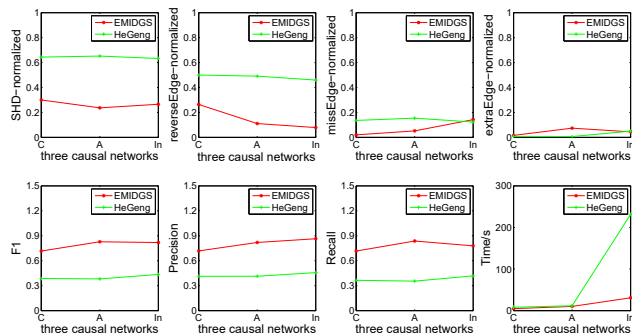


Figure 4: Comparison of EMIDGS and the HeGeng algorithm under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$.

7, 8, 9 and 10. Figures 5, 6, 7 and 10 show that EMIDGS achieves relatively better performance in accuracy and efficiency when $\alpha = 0.01$ than that obtained with other values of α on “child”, “insurance”, “alarm” and “link”. And in the remaining networks, different values of have a little effect on the change in accuracy as Figures 8 and 9 (i.e., “win95pts” and “munin”). In addition, we also observe that the gap of the accuracy among multiple situations is relatively large in medium-sized networks. However the gap is small in large-sized networks. Considering accuracy and time, $\alpha = 0.01$ is more appropriate.

For mT , referring Figures 6, 8, 9 and 10, we observe that there is no significant difference in the accuracy of the experimental results obtained with different values of mT . However, Figures 5 and 7 show that EMIDGS achieves relatively better performance in accuracy when $mT = 2$ than that obtained with other values of mT on “child” and “alarm”. In addition, referring Figures 5-10, we observe that the time efficiency of EMIDGS is diminishing, with mT increasing and $mT \in \{2, 4, 6, 8, 10\}$, except “link”. That is to say, the efficiency of the results obtained when $mT = 10$ is relatively better than that others in many networks. Considering accuracy and time, $mT = 10$ is more appropriate.

Also, we analyze the performance of different number-

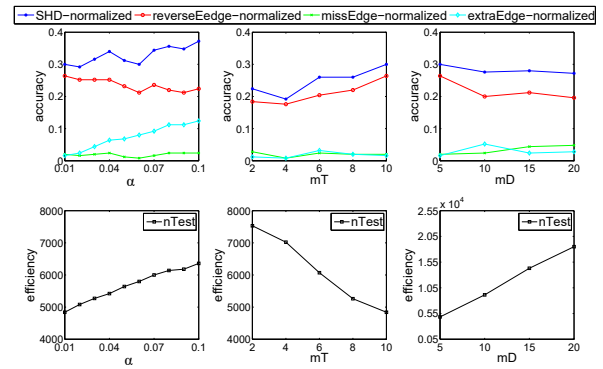


Figure 5: The experimental results of EMIDGS under different values of three parameters on “child”.

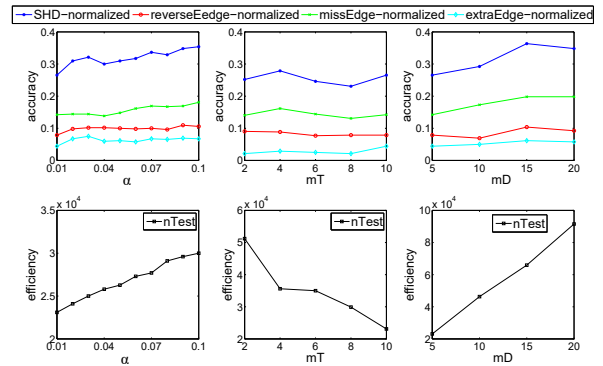


Figure 6: The experimental results of EMIDGS under different values of three parameters on “insurance”.

s of datasets in each group. Figures 6, 7, 8 and 10 show that EMIDGS has the best performance in terms of accuracy and efficiency when $mD = 5$. In Figures 5 and 9, although the accuracy of experimental results when $mD = 20$ is slightly higher than that of $mD = 5$, the difference is not obvious. And the efficiency of the result corresponding to $mD = 5$ is high. Moreover, referring Figures 5-10, the time consumed by EMIDGS increases as the value of mD increases ($mD \in \{5, 10, 15, 20\}$). In the “link” network, the time assumed by EMIDGS when $mD = 20$ approximately reaches 10^7 . Considering accuracy and time, $mD = 5$ is more appropriate. However, $mD = 20$ is appropriate if one is more concerned with accuracy.

(4) EMIDPC vs. MIPC

We evaluate the performance of the proposed EMIDPC algorithm with the MIPC algorithm proposed by Yu et al. (Yu, Liu, and Li 2019), using the following metrics: Precision, Recall, F1 and nTest, where the first three indicators represent accuracy, and the latter indicator represents time efficiency. Figure 11 shows that the experimental results of EMIDPC and MIPC under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$. The comparisons show that EMIDPC has higher performance in accuracy than MIPC, especially in large-sized networks. However, the time efficiency

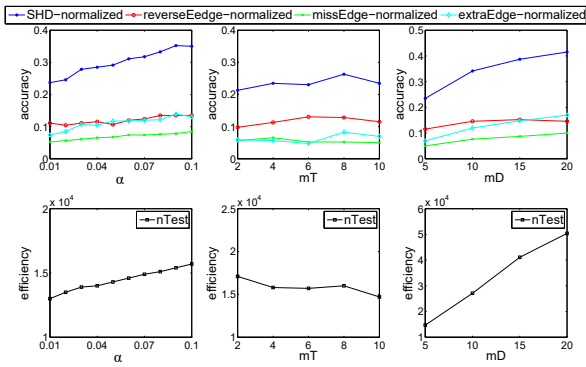


Figure 7: The experimental results of EMIDGS under different values of three parameters on “alarm”.

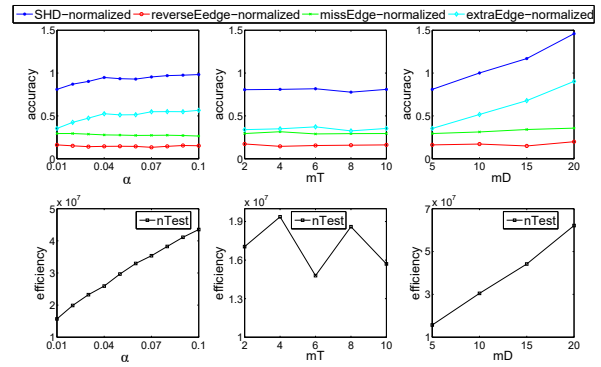


Figure 10: The experimental results of EMIDGS under different values of three parameters on “link”.

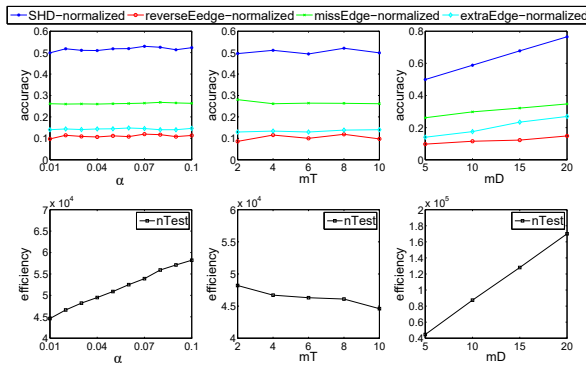


Figure 8: The experimental results of EMIDGS under different values of three parameters on “win95pts”.

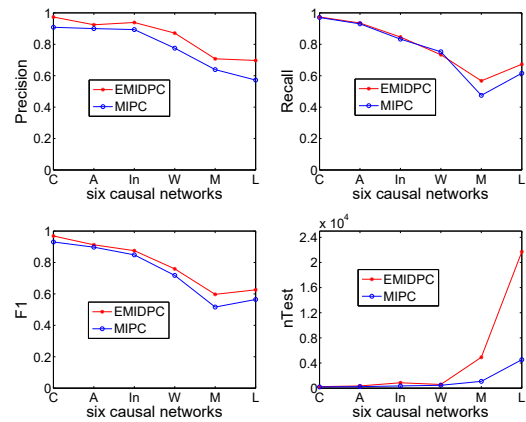


Figure 11: Comparison of MIPC and our proposed EMIDPC under the conditions: $\alpha = 0.01$, $mT = 10$ and $mD = 5$.

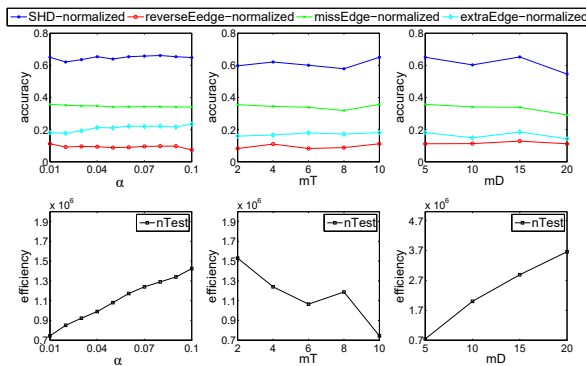


Figure 9: The experimental results of EMIDGS under different values of three parameters on “munin”.

of EMIDPC is relatively lower than MIPC.

Conclusion and Future Work

In this paper, we proposed EMIDGS, a novel algorithm for learning causal structures from multiple high-dimensional manipulated datasets with unknown targets. EMIDGS reconstructs a skeleton of a causal DAG by the local learning

algorithm, i.e., EMIDPC proposed in this paper. And it orients as many edges as possible by two Lemmas proposed in this paper, getting an I-MEC, which includes more causal information. We proved that the learned I-MEC is theoretically correct and the combination of multiple graphs reduces the search space, leading to an improvement in efficiency. In addition, experimental results validated the effectiveness of our algorithm in reducing the false discovery rate of multiple high-dimensional interventional datasets. In the future, we focus on learning the local causal structure from soft interventions with unknown targets.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020AAA0106100) and the National Natural Science Foundation of China (61976128, 61876206).

References

Antonia, A.; Wendell, A.; and Lei, S. 2016. Beyond editing: repurposing CRISPR/Cas9 for precision genome regulation

- and interrogation. *Nature Reviews Molecular Cell Biology*, 17: 5–15.
- Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.
- Brouillard, P.; Lachapelle, S.; Lacoste, A.; Lacoste-Julien, S.; and Drouin, A. 2020. Differentiable Causal Discovery from Interventional Data. *34th Conference on Neural Information Processing Systems*.
- Chickering, D. M. 2002. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 445–498.
- Ghassami, A.; Salehkaleybar, S.; Kiyavash, N.; and Zhang, K. 2017. Learning Causal Structures Using Regression Invariance. *In Advances in Neural Information Processing Systems*, 3011–3021.
- Hauser, A.; and Bühlmann, P. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 2409–2464.
- He, Y.-B.; and Geng, Z. 2016. Causal network learning from multiple interventions of unknown manipulated targets. *arXiv preprint arXiv:1610.08611*.
- Huang, B.; Zhang, K.; Zhang, J.; Ramsey, J.; Sanchez-Romero, R.; Glymour, C.; and Schölkopf, B. 2020. Causal Discovery from Heterogeneous/Nonstationary Data. *Journal of Machine Learning Research* 21, 21: 1–53.
- M, R.-C.; B, S.; and Turner R, P. J. 2018. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1): 1309–1342.
- Meinshausen, N.; Hauser, A.; Mooij, J. M.; Peters, J.; Versteeg, P.; and Bühlmann, P. 2016. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27): 7361–7368.
- Mooij, J. M.; Magliacane, S.; and Claassen, T. 2020. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99): 1–108.
- Pearl, J. 2009. *Causality*. Cambridge University Press.
- Pearl, J.; and Mackenzie, D. 2018. The book of why: the new science of cause and effect. *Basic Books*.
- Peters, J.; Bhlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012.
- Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *Proceedings of the IEEE-Advances in Machine Learning and Deep Neural Networks*, 109(5): 612–634.
- Shohei, S.; Patrik, O. H.; Aapo, H.; and Antti, J. K. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(10): 2003–2030.
- Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. The MIT Press.
- Squires, C.; Wang, Y.; and Uhler, C. 2020. Permutation-Based Causal Structure Learning with Unknown Intervention Targets. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 124: 1039–1048.
- Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1): 31–78.
- Wu, X.; Wang, J.; Huang, H.; and Lin, R.-J. 2015. Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nature Biotechnology*, 33(2): 175–178.
- Xun, Z.; Chen, D.; Bryon, A.; Pradeep, R.; and Xing, E. P. 2020. Learning Sparse Nonparametric DAGs. *International Conference on Artificial Intelligence and Statistics*, 3414–3425.
- Yang, K. D.; Katcoff, A.; and Uhler, C. 2018. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. *Proceedings of the 35th International Conference on Machine Learning*, 80: 5537–5546.
- Yu, K.; Liu, L.; and Li, J. 2019. Learning markov blankets from multiple interventional data sets. *IEEE transactions on neural networks and learning systems*, 31(6): 2005–2019.
- Yu, K.; Liu, L.; Li, J.; Ding, W.; and Le, T. D. 2020. Multi-Source Causal Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9): 2240–2256.
- Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG Structure Learning with Graph Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 7154–7163.
- Zhang, K.; Huang, B.; Zhang, J.; Glymour, C.; and Schölkopf, B. 2017. Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1347–1353.