Efficient Causal Structure Learning from Multiple Interventional Datasets with Unknown Targets

Yunxia Wang,[1] Fuyuan Cao,[1,*] Kui Yu,[2] Jiye Liang[1]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, P.R. China
[2]School 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, ljy@sxu.edu.cn

Abstract

We consider the problem of reducing the false discovery rate in multiple high-dimensional interventional datasets under unknown intervention targets when the time efficiency is acceptable. Traditional algorithms merged directly multiple causal DAGs learned, which ignores the contradictions of different datasets, leading to lots of contradictory directions of edges. For reducing the contradictory information, we propose a new algorithm, which first learns an intervention Markov equivalence class (I-MEC) before merging multiple causal DAGs. 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 and obtains more V-structures, getting a theoretically correct I-MEC. Search and scoring techniques based on the learned I-MEC further identify the remaining unoriented edges. 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.

1 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, we can distinguish the corresponding causes and results 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 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 relatively poor accuracy. 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 UN-IGST 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. The two algorithm-

*Corresponding author: Fuyuan Cao. Email: cfy@sxu.edu.cn.

s 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 (Efficient Multiple Interventional Datasets for 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 Intervention Markov Equivalence Class (I-MEC) (Yang, Katcoff, and Uhler 2018) by two lemmas (See Lemma 1 and 2); (2) based on the learned I-MEC, applying a score function to greedily search each causal DAG from each interventional 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. The second step of the orientation phase greatly reduces the search space, leading to a great improvement of EMIDGS in the efficiency. Therefore, our proposed algorithm reduces the contradictions of multiple causal structures by first learning an intervention Markov equivalence class (I-MEC) before merging multiple causal DAGs. And experimental results also verify the effectiveness of our algorithm.

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.

2 Preliminaries work

2.1 Causal DAG model

Let $G = (V, E)$ represent a directed acyclic graph (DAG) and P be the joint probability distribution over a random vector $X_V = \{X_1, \cdots, X_n\}$. Each node $j \in V = \{1, \cdots, n\}$ 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. $X_i \to X_j$ 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

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

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.

2.2 Post-intervention DAG

Let $D = \{D_1, \cdots, D_m\}$ be the m experimental datasets, $\forall i \in \{1, \cdots, m\}$, $R_i \subseteq V$ be the set of variables manipulated in the i-th experiment and $do(R_i)$ denotes 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 $E_i = \{(a, b) | (a, b) \in E, b \notin Ri\}$. 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 causal structure under unknown intervention targets.

3 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 $i \in \{1, \cdots, m\}$.*

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 (Efficient Multiple Interventional Datasets for 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.

Algorithm 1: The *EMIDGS* algorithm.

1: **Input:** $D = \{D_1, \cdots, D_m\}$:m interventional datasets; $V = \{X_1, \cdots, X_n\}$:n variables.
2: **Output:** G_f .
Step1: get the causal skeleton G_s
3: **for** $X_j \in V$ **do**
4: $[pc, sep, kpc, kindep]$=*EMIDPC*(D, X_j, V);
5: **for** $Y \in pc$ **do**
6: $G(X_j, Y) = 1$;
7: **end for**
8: **end for**
9: $G_s = G = (V, E_s)$;
Step 2: orient edges
10: **for** $A \in V$ **do**
11: **for** $X, Y \in pc(A)$ **do**
12: **for** $i = 1$ to m **do**
13: **for** $S \subseteq sep_X(Y) \cup sep_Y(X)$ **do**
14: **if** $X \not\perp\!\!\!\perp Y \mid S, X \perp\!\!\!\perp Y | \{S \cup A\}$ in D_i **then**
15: $G(X, A) = -1; G(Y, A) = -1$;
16: $G(A, X) = 0; G(A, Y) = 0$;
17: $tem = kindep_A(i) \cap pc(A)$;
18: $G(A, tem) = -1; G(tem, A) = 0$;
19: **end if**
20: **end for**
21: **end for**
22: **end for**
23: updating G by Meek rules;
24: **end for**
25: $G_0 = G = (V, E_0)$;
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_f.
30: return G_f

3.1 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 3-9 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 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 child of T. And the conclusion that the true PC set must be included in $canpc(T)$ is believed. The next thing we need to do is to remove false positives from $canpc(T)$ as much as possible.

Step 2: Find PC of T (lines 17-45). We remove false positives from $canpc(T)$ by the standard forward-backward strategy (SFBS) as shown in lines 17-33. Firstly, we set $cpc(T)$ an empty set and select the feature $X \in canpc(T)$, with the highest association with T, and then add X into $cpc(T)$ and remove X from $canpc(T)$. Next, we determine 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 we do not know which variables are manipulated in each dataset. So we discuss the situations about removing false positives as follows.

For non descendants excluding parents, because R is conservative, we can always find at least one dataset, denoted as 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 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 datset, T is independent of a T'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 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 the child of T is intervened in the i-th dataset, we cannot find the descendants excluding children in $kpc_T(i)$. If the 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 $\{Y\}$.

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

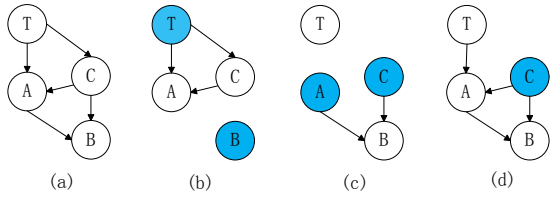

Figure 1: Four causal DAGs, where (a) is the underlying DAG; and (b)(c)(d) are three post-intervention DAGs after the manipulation on the some variables in (a).

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 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. 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, 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, we consider the length of S with $temp + 1$. This process ends at a condition where $temp$ is greater than the length of $cpc(T)$. Finally, we set $pc(T) = cpc(T)$, and output $pc(T)$, kpc_T, $kindep_T$, sep_T.

3.2 Orienting edges

The phase of edge orientation is divided into two steps. In step 1, EMIDGS orients edges 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 obtained in Step 1, EMIDGS performs a score-and-search strategy in finite research space to learn m structures from m interventional datasets and then the m graphs are combined to get the final structure. Due to page limit, the proofs of Lemmas 1 and 2 are placed in the Appendix A.

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

Proof: *See Appendix A.*

Lemma 2 *The unique property of perfect intervention. For m interventional datasets without data errors and selection*

Algorithm 2: The *EMIDPC* algorithm.

1: **Input:** $D = \{D_1, \cdots, D_m\}$: m interventional datasets; T: the target variable; $V = \{X_1, \cdots, X_n\}$: n variables.
2: **Output:** $pc(T)$,sep_T, kpc_T, $kindep_T$.
3: $canpc(T) = \phi$; $pc(T) = \phi$;
4: $kpc_T = cell(1, m)$; $kindep_T = cell(1, m)$;
Step1: get $canpc(T)$, kpc_T, $kindep_T$
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**
Step 2: get $pc(T)$
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** $X \perp\!\!\!\perp T \mid S, S \subseteq cpc(T) \setminus X, S \subseteq kpc_T(i), X \in kpc_T(i)$ in D_i **then**
24: $cpc(T) = cpc(T) \setminus X$;
25: $sep_T(X) = S$;
26: **if** $\exists h \in \{1, \cdots, m\}$ 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 $|S| = temp$, $X_j \perp\!\!\!\perp T \mid S$ *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 $pc(T), sep_T, kpc_T, kindep_T$

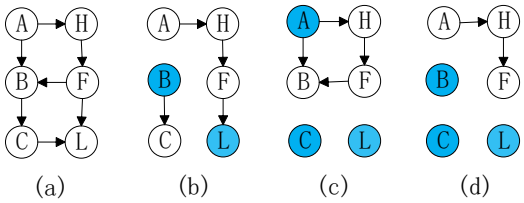

Figure 2: Four causal DAGs, where (a) is a underlying DAG; and (b)(c)(d) are three post-intervention DAGs after the manipulation on the some variables in (a).

bias, and $X, Y, A \in V$, assuming a triple $< X, Y, A >$ 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: *See Appendix A.*

Step1: Orient edges by Lemmas 1 and 2 (lines 10-24). We orient the edges containing in the learned skeleton by two properties described in Lemmas 1 and 2, respectively. Some edges can be oriented by the unique of V-structure as described in Lemma 1. EMIDGS first selects a variable $A \in V$ as the current variable. Then for every two variables, i.e. X, Y, belonging to $pc(A)$, EMIDGS judges the independence of X and Y conditioned on a set Z and judges the dependence of X and Y conditioned on $\{Z \cup A\}$ in each dataset. We choose Z as a subset of $\{sep_X(Y) \cup sep_Y(X)\}$. If the original structure connected with V-structure is not destroyed, then the set Z is $sep_X(Y)$ or $sep_Y(X)$. If only V-structures exist, then the set Z is a subset of $\{sep_X(Y) \cup sep_Y(X)\}$, also including an empty set. Therefore, if $X \perp\!\!\!\perp Y|Z$ and $X \not\perp\!\!\!\perp Y|\{Z \cup A\}$ occur in D_i, 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. The reason is as follows. If the V-structure $\langle X, A, Y \rangle$ including the collider A is determined, then A is not intervened in the dataset D_i. When the variables in $pc(A)$ are independent of A, then the variables are the children of 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 25. 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 $\{X, Y, A\} \cup S \subseteq kpc_A(i)$, 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 a V-structure in 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, getting an I-MEC, which includes more causal relationships.

Step2: Score-and-search, and merge (lines 25-29). Based on the I-MEC learned (i.e., G_0) by Step 1, EMIDGS uses a scoring method to learn and search the graph with the highest score in the remaining graph space as the best graph. This operation can greatly reduce the graph search space and improve the time efficiency. In addition, using Lemma 1 and Lemma 2, the edges oriented in Step 1 are all the correct directions of edges in theory, so the accuracy of the algorithm will be improved.

After getting m causal DAGs, EMIDGS combines these DAGs to the final graph G_f. That is, based on the I-MEC (i.e., G_0), we analyze the remaining unoriented edges e_j (i.e., $e_j \in E_s \setminus E_0$), in which E_0 represents an oriented edge set of the I-MEC G_0. And we set the final graph $G_f = G_0$. If the direction of e_j is determined in a certain graph $G_f(i)$, then the directed edge e_j is added to G_f. If the direction of e_j is contradictory in k ($1 \leq k \leq m$) graphs, then e_j is not added to G_f. According to the above conditions, we judge all $e_j \in E_s \setminus E_0$ and finally get the merged graph G_f. Finally, EMIDGS outputs the final causal DAG G_f.

4 Theoretical analysis and complexity analysis

In this section, we theoretically analyze and prove the correctness of our algorithm in Section 4.1 and due to page limit, the proofs of Propositions 1, 2 and 3 are placed in the Appendix B. Also we analyze the computational complexity of our algorithm in Section 4.2.

4.1 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 correctly and efficiently.*

Proof: *See Appendix B.*

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

Proof: *See Appendix B.*

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: *See Appendix B.*

4.2 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 to 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 $O(m * |p| * (C(p,1) + ... + C(p,p)))$. 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 8, the complexity of reconstructing a skeleton is $O(|V|tpc)$. At lines 10 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 $O(2m|V|2^{|pc|}|pc|^2)$.

5 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 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, the paper adds two additional baseline algorithms. One is called baseline-MMHC, which learns a causal DAG from each interventional dataset by MMHC and then merges 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.

Table 1: Description of benchmark BNs.

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

We select six Bayesian networks (BNs) with the number of nodes ranges from 20 to 724 as shown in Table 1 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 intervention targets sets 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 $mD = 10, 15, 20$ 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. Each dataset contains 5000 samples.

We conduct experiments with different confidence levels of α = 0.01 to α = 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: $\alpha = 0.01$, $mT = 10$. In all experiments, G^2 tests are used for the conditional independence tests. In addition, we evaluate and compare the performance of the 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. All experiments are conducted on a computer with an AMD Core A8-6410 2.00 GHz with 4GB RAM. The experimental results are shown in Figures 3,4,5 and 6. 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'.

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

Figure 3 shows 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 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 MIPC algorithm adopted by MIMB-GS is relatively inferior, compared with EMIDPC proposed by this paper, that is the causal skeleton constructed by MIMB-GS is relatively inferior. Make it easier to understand, we will give the experimental results of EMIDPC and MIPC in Appendix C. The baseline-MMHC algorithm directly integrates multiple graphs, accumulating more contradictory information especially in large-sized networks. 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.

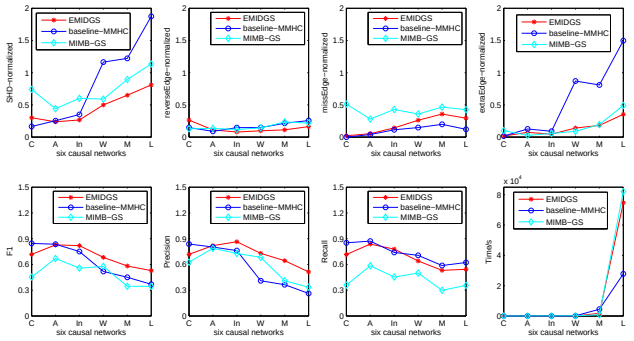

Figure 3: Comparison of performance of EMIDGS, MIMB-GS, and baseline-MMHC algorithms.

(2) EMIDGS vs. the HeGeng algorithm

Since the outputs of the "win95pts", "munin", "link" cannot be generated in 72 hours by the HeGeng algorithm, we only compare the three medium-sized networks with our algorithm as shown in Figure 4. Figure 4 shows that EMIDGS has a greater improvement in efficiency and accuracy, compared with the HeGeng algorithm.

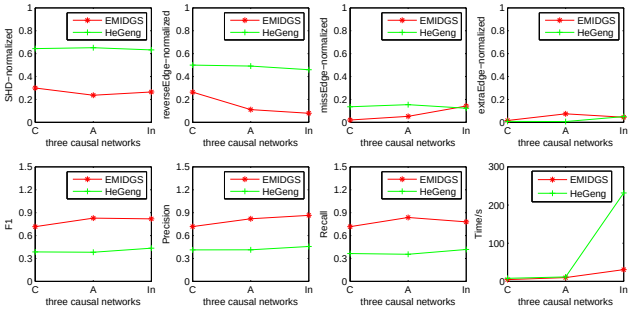

Figure 4: Comparison of performance of EMIDGS and the HeGeng algorithms.

(3) Different values of three parameters

Due to page limit, we present the results of the different values of the three parameters of EMIDGS on a medium-sized network ("insurance") and a large-sized network ("munin") as shown in Figures 5 and 6, and the rests are placed in Appendix D. Figures 5 and 6 show that EMIDGS achieves the best accuracy and efficiency when $\alpha = 0.01$ on "insurance" and "munin". In addition, combined with the figures in Appendix D, we 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, combined with the figures in Appendix D, we observe that the accuracy of different values of mT is not obvious, but the time efficiency is diminishing, with mT increasing and $mT \in \{2, 4, 6, 8, 10\}$. We also analyze the performance of different numbers of datasets in each group. Obviously, as the value of mD increases, the time also increases. The reason is that the number of datasets as a group increases. In contrast, Figure 5 shows that the best accuracy when mD is 5 on "insurance", and Figure 6 shows that the best accuracy when mD is 20 on "munin". However, combined figures in Appendix D, it shows that the accuracy of the results obtained when mD = 5 is relatively better than that others in many networks.

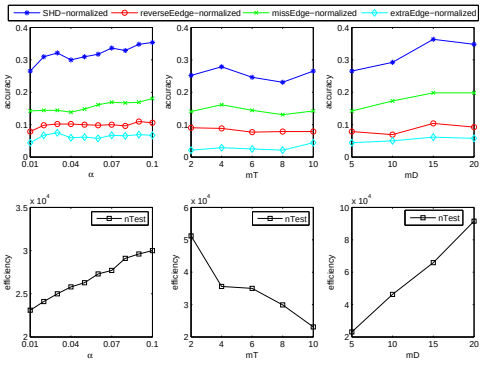

Figure 5: Comparison of performance of EMIDGS under different values of three parameters on "insurance".

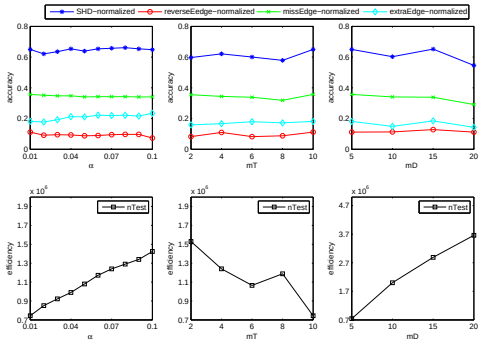

Figure 6: Comparison of performance of EMIDGS under different values of three parameters on "munin".

6 Conclusion and future work

In this paper, we proposed EMIDGS, a novel algorithm for learning causal structures from multiple manipulated datasets with unknown intervention targets. EMIDGS reconstructs the causal skeleton by the local learning algorithm, i.e. EMIDPC proposed in this paper, and orients edges as many as possible by invariance of V-structures and the unique property of perfect intervention, getting an I-MEC, which includes more causal information. We show that the learned I-MEC is theoretically correct and the combination of multiple graphs reduces the search space. In addition, experimental results validate the effectiveness of our algorithm in reducing the false discovery rate of multiple high-dimentional 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 CRISPRCCas9 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 Artifificial 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 36 th 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.