Local Causal Discovery Without Causal Sufficiency

Zhaolong Ling[1], Jiale Yu[1], Yiwen Zhang[1*], Debo Cheng[2], Peng Zhou[1], Xingyu Wu[3], Bingbing Jiang[4], Kui Yu[5]

[1]School of Computer Science and Technology, Anhui University
[2]UniSA STEM, University of South Australia
[3]Department of Computing, Hong Kong Polytechnic University
[4]School of Information Science and Engineering, Hangzhou Normal University
[5]School of Computer Science and Information Technology, Hefei University of Technology
zlling@ahu.edu.cn, e23201057@stu.ahu.edu.cn, zhangyiwen@ahu.edu.cn, xingy.wu@polyu.edu.hk, debo.cheng@unisa.edu.au, jiangbb@hznu.edu.cn, yukui@hfut.edu.cn

Abstract

Local causal discovery is crucial for revealing the causal relationships between specific variables from data. Traditional local causal discovery algorithms are designed under the assumption of causal sufficiency, which states that there are no latent common causes for two or more of the observed variables in data. However, the assumption of causal sufficiency is often violated in practice. To address this issue, we first propose the local Maximal Ancestral Graph (MAG), referred to as LocalMAG, to describe the local causal relationships of the target variable in the MAG. Then, we propose a local causal discovery algorithm without the assumption of causal sufficiency, called LatentLCD, to learn the LocalMAG. Specifically, LatentLCD first uses the traditional parents and children discovery algorithm to identify the local causal skeleton that includes latent variables and verifies it theoretically. It then identifies bidirectional edges by determining whether both the target variable and its adjacent variables are colliders, thereby identifying latent variables in the local structure of the target variable. Extensive experiments on synthetic datasets validate that the proposed LatentLCD algorithm significantly outperforms the state-of-the-art methods.

Introduction

Causal discovery plays an irreplaceable role across various fields, including biology, epidemiology, medicine, economics, and computer science (Pearl 2018; Schölkopf et al. 2021; Kuang et al. 2017). Thus, recovering causal relationships from data is one of the ultimate goals in the empirical sciences (Cai, Zhang, and Hao 2013). Learning a causal model that describes the causal relationships among variables is represented by a directed acyclic graph (DAG) (Yu, Liu, and Li 2019). The existing causal discovery algorithms learn the causal relationships of a DAG from datasets and can be divided into two methods: the global discovery method, which aims to learn the entire causal relationships of the DAG (Gao, Fadnis, and Campbell 2017), and the local discovery method, which aims to uncover the causal relationships around a specific variable (Wang et al. 2023). If we are focused only on the local causal relationships of a specific variable, it is unnecessary to spend extensive time and resources learning a global causal relationships (Statnikov et al. 2015), as learning a global causal relationships is a NP-complete problem (Chickering 1996).

The traditional local causal discovery algorithms first discover the edges connected to target, then identify V-structures and apply Meek rules (Meek 1995) to find the parents and children (PC) of target (Gao and Ji 2015). For example, the PCD-by-PCD algorithm (PCD means parents, children, and descendants) (Yin et al. 2008) to identify V-structures in the PC set and orients edges using Meek rules. MB-by-MB (Wang et al. 2014), which sequentially identifies the Markov Blanket (MB) of adjacent variables of the target, then orienting the connected edges. Causal Markov Blanket (CMB) (Gao and Ji 2015) learns the local causal relationships of the target variable based on MB discovery. The Partial BN Structure Learning (PSL) algorithm (Ling et al. 2022) recursively finds Type-C and Type-NC V-structures until the local structure is determined. However, these local causal discovery algorithms are based on the assumption of causal sufficiency and do not consider learning the local causal relationships when the causal sufficiency is violated. MMB-by-MMB (MMB means MAG MB) (Xie et al. 2024) finds the PC of target even when the causal sufficiency is not satisfy. Sometimes, it discovers larger local structures, which leads to higher time costs.

The assumption of causal sufficiency implies that there are no unobserved common causes in the data generation process (Margaritis and Thrun 1999). A common cause is defined as a direct cause that influences two or more observed variables (Cheng et al. 2022a). If this common cause is not present in the observed dataset, it is referred to as a latent common cause or latent variable (Yu et al. 2018). When the data relaxes the assumption of causal sufficiency, i.e., the existence of latent variables, the traditional local causal discovery algorithms mistakenly identify the direct effects of the latent variables as direct effects or direct causes of the specific variable. For example, in Figure 1, when "Genotype" is not observed, this unobserved variable represents the latent common cause of "Attention Disorder" and

*Corresponding author: Yiwen Zhang.

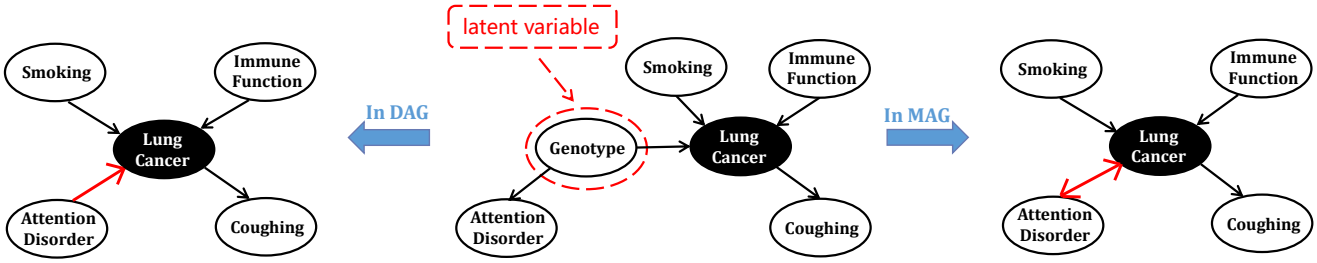

Figure 1: When the causal sufficiency is not satisfied, we consider "Genotype" as a latent variable. Discover the local causal relationships of "Lung Cancer", including {Smoking, Immune Function, Coughing, Attention Disorder}. In the latent DAG, the traditional local causal discovery algorithm falsely identifies "Attention Disorder" as the parent of "Lung Cancer". In the MAG, the correct local causal relationships include the bidirectional edge between "Attention Disorder" and "Lung Cancer". The red bidirectional edge represents the existence of latent variables between "Lung Cancer" and "Attention Disorder".

"Lung Cancer" in the DAG. However, in practical applications, it is impossible to obtain information about the positions and numbers of latent variables (Cheng et al. 2022b). Thus, the DAG model that includes latent variables fails to represent the true causal relationships among observed variables (Zhang 2008a). Using the traditional local causal discovery algorithms, learned local causal relationships of "Lung Cancer", which falsely identify "Attention Disorder" as a direct cause of "Lung Cancer".

The concept of the Maximal Ancestral Graph (MAG) has been proposed to represent latent variables in the data through bidirectional edges without knowing the number or locations of these latent variables in the graph beforehand (Richardson and Spirtes 2002). Thus, MAG depicts the causal relationships among observable variables even when the causal sufficiency is not satisfied (Cheng et al. 2024). For example, in Figure 1, the bidirectional edge between "Attention Disorder" and "Lung Cancer" represents the existence of a latent variables. In the MAG, the correct local causal relationships of the "Lung Cancer" include {Smoking, Immune Function, Coughing, Attention Disorder}, but it recognizes the bidirectional edge between "Attention Disorder" and "Lung Cancer".

When causal sufficiency is not satisfied, global causal discovery algorithms for learning an entire MAG are continuously being proposed (Nogueira et al. 2022). These algorithms use conditional independence (CI) test to find all the potential global causal relationships and apply m-separation (a graphical criterion to infer dependences/independences between variables) to identify V-structures in the orientation phase (Nogueira et al. 2022). However, learning global causal relationships is both unnecessary and a waste of resources as they have a high time-complexity (Gao and Ji 2015). Thus, in many real-word applications, we focus on the discovery of local causal relationships.

Moreover, there are some limitations when directly applying these global causal discovery algorithms for local causal discovery. For example, the Fast Causal Inference (FCI) algorithm (Spirtes 2001) introduces the Possible-D-SEP process during the skeleton establishment stage, which requires a large number of CI tests, leading to an exponential increase in time complexity (Colombo et al. 2012; Rohekar et al. 2021). The Really Fast Causal Inference (RFCI) algorithm (Colombo et al. 2012) avoids this stage, but the output of RFCI does not provide complete causal relationships (some MAGs in the equivalence class can be excluded given the data) (Rohekar et al. 2021). The Iterative Causal Discovery (ICD) algorithm (Rohekar et al. 2021) defines a PDS-tree, which reduces the size of the conditioning set for CI tests, significantly lowering time complexity. However, this also results in the loss of some edges, leading to lower accuracy in identifying the local skeleton of the target variable.

To address these challenges, this paper proposes a new algorithm based on the assumption of causal faithfulness to discover the local causal relationships of the target variable in MAG. Our main contributions are as follows:

1) We propose the concept of LocalMAG to decsribe the local causal relationships of the specific variable in the MAG, and then theoretically analyze the relationships between latent common causes and their direct effects based on LocalMAG.
2) We propose a local causal discovery algorithm without causal sufficiency, called LatentLCD. LatentLCD identify whether the target variable and its adjacent variables are all colliders to discover the bidirectional edges, uncovering the latent variables.
3) We preform extensive experiments on numerous synthetic datasets. The experimental results show that our proposed algorithm outperforms the state-of-the-art methods in the structural and skeleton accuracy.

Methodology

Given that traditional local causal discovery algorithms cannot learn correct local causal relationships when data does not satisfy causal sufficiency, we propose a local causal discovery algorithm for data containing latent variables.

Preliminaries

We use "$*$" to represent any edge mark "$-$" or "$>$".

Definition 1 (Collider and Non-Collider) (Richardson and Spirtes 2002). In an ancestral graph, given a path π and

a variable X_i, if π contains a structure $* \rightarrow X_i \leftarrow *$, X_i is called a collider; otherwise, X_i is called a non-collider.

In an ancestral graph, X_i is a collider in the following triples: $X_j \rightarrow X_i \leftarrow X_k$, $X_j \leftrightarrow X_i \leftarrow X_k$, $X_j \rightarrow X_i \leftrightarrow X_k$, $X_j \leftrightarrow X_i \leftrightarrow X_k$.

Definition 2 (m-Separation) (Richardson and Spirtes 2002). In an ancestral graph, given a path π, if the conditioning set $\mathbf{Z}$ satisfies one of the following two conditions, then π is blocked by $\mathbf{Z}$:

1) There exists a subpath $\{X_i, X_j, X_k\}$ in π, where X_j is a non-collider and $X_j \in \mathbf{Z}$.
2) In π, there exists a structure $X_i * \!\!\rightarrow X_j \leftarrow\!\! * X_k$, and there are no descendants of X_j in $\mathbf{Z}$ and $X_j \notin \mathbf{Z}$.

Definition 3 (Maximal Ancestral Graph) (Richardson and Spirtes 2002). If an ancestral graph G satisfies the condition that for every pair of non-adjacent nodes X and Y in the graph, there exists a set $\mathbf{Z}$ such that X and Y are m-separated by $\mathbf{Z}$, then the ancestral graph G is called a Maximal Ancestral Graph.

Theorem 1 (Yu et al. 2018). In a MAG, for an unshielded triplet $\{X, T, Y\}$, if X and Y are m-separated by a conditioning set $\mathbf{Z}$ but $X \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup T$, then T is a collider, indicating the presence of a V-structure $X * \!\!\rightarrow T \leftarrow\!\! * Y$. This is shown in the following formula:

$$X \perp\!\!\!\perp Y \mid \mathbf{Z} \; and \; X \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup T \tag{1}$$

Theorem 2 (Pearl 2014). Under the faithfulness assumption, $X \in \mathbf{O}$ and $Y \in \mathbf{O}$ are connected in the MAG if and only if $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ for all $\mathbf{Z} \subseteq O \setminus \{X, Y\}$.

Theorem 2 indicates that under the faithfulness assumption, a MAG includes all conditional independence relationships between the observed variables, and these relationships are true in the underlying DAG.

The LatentLCD Algorithm

In this section, we describe the LatentLCD Algorithm, which discovers local causal relationships of specific variables under causal sufficiency not satisfied, as shown in Algorithm 1. We also describe definitions and theoretical analysis for ensuring the correctness of LatentLCD algorithm.

Because the local causal discovery is based on MAG, we defined the local causal relationships that specific variables should contain in the existence of latent variables in the dataset and prove their rationality and correctness.

Definition 4 (Effects of Latent causes). In a MAG, the set of Effects of Latent causes of T, denoted as EL(T), consists of variables X_i that are adjacent to T and connected by a bidirectional edge. Formally, $\forall X_i \in$ EL(T), there is a bidirectional edge between X_i and T.

Definition 5 (Local Maximal Ancestral Graph). In a MAG, the local causal relationships of a given variable T are referred to as LocalMAG(T), which consists of the children of T, the parents of T, and the set of Effects of Latent causes of T. Specifically, LocalMAG(T) in MAG includes the following components:

1) Ch(T): The children of T.
2) Pa(T): The parents of T.
3) EL(T): The set of Effects of Latent causes of T.

For example, in Figure 1, "Cancer" is defined as T. LocalMAG(T) includes {Smoking, Immune Function} (Pa(T)), {Coughing} (Ch(T)), and {Attention Disorder} (EL(T)). When there are bidirectional edges between T and its adjacent variables, meaning EL(T)$\neq \emptyset$, latent variables between bidirectional edges affect their causal relationships. EL(T) pinpoint exact locations of latent variables.

Next, we will prove the rationality and correctness of LocalMAG.

Theorem 3. Under the assumption of causal sufficiency, the LocalMAG of T in a MAG equals the local causal relationships of T in the corresponding DAG.

Proof: The LocalMAG(T) includes Pa(T), Ch(T), and EL(T). Our approach mines latent variables between T and its adjacent variables by identifying bidirectional edges, and adds these adjacent variables to EL(T). Under the assumption of causal sufficiency, EL(T)$= \emptyset$. In the corresponding DAG, the local causal relationships of T include Pa(T), Ch(T). Thus, under the assumption of causal sufficiency, we obtain complete causal relationships equivalent to the local causal relationships in a DAG.

Through Theorem 3, we have the theoretical guarantee for LocalMAG, which is important for LatentLCD.

Based on LocalMAG, the LatentLCD algorithm learns the local causal relationships of the target variable T from data containing latent variables through three steps: 1) First, the the max-min parents and children (MMPC) (Tsamardinos, Aliferis, and Statnikov 2003) algorithm is used to find the set of adjacent variables of T, which includes the PC set of T (PC(T)) and EL(T), denoted as $adj(T)$; 2) By identifying V-structures, the edges adjacent to T are oriented, and the nodes pointing to T are marked as candidate parents. Then, the FindBI algorithm (Algorithm 2) is used to traverse the candidate parent nodes to identify any potential bidirectional edges. 3) Using the R1-R4 and R8-R10 rules proposed by (Zhang 2008b) to orient the remaining edges.

Using Theorem 2, we can derive the following proposition to prove that the MMPC algorithm finds all the adjacent variables of T.

Proposition 1. Under the faithfulness assumption, if bidirectional edges exist in the LocalMAG of T, indicating the presence of latent variables, the traditional PC discovery algorithm can correctly identify the children of these latent variables as the adjacent variables of T. That is, if $T \leftrightarrow K$, $adj(T) = \{K\}$.

Proof: Let L be the latent common cause of T and K. Transform $T \leftrightarrow K$ in the MAG into $T \leftarrow L \rightarrow K$ in the DAG. Assuming that both $T \leftrightarrow K$ and $T \leftarrow L \rightarrow K$ satisfy the causal faithfulness assumption, it is known from the chain rule that:

$$P(T, L, K) = P(T \mid L)P(K \mid L)P(L) \tag{2}$$

When the variable L is observed and included in the conditioning set, it is known from Bayes' rule that:

$$P(T \mid L) = \frac{P(T, L)}{P(L)} \tag{3}$$

$$P(K \mid L) = \frac{P(K, L)}{P(L)} \tag{4}$$

Algorithm 1: LatentLCD

Input: $\mathcal{D}$: Data, T: The target variable, α: Significant level, $\mathbf{O} = \{X_1, \ldots, X_n\}$: n Variables;
Output: PAG: The Local causal relationships of T in the existence of latent;

1: **Initialize:** $\boldsymbol{W} = \emptyset, Q = \{T\}$,$PAG = (|\mathbf{O}|, |\mathbf{O}|)$
2: **repeat**
3: /*Step 1: LocalMAG-based local causal skeleton*/
4: $T_0 = Q.pop$;
5: **if** $T_0 \notin$ **W then**
6: **adj**(T_0) = MMPC($\mathcal{D}$,T_0,α);
7: $\boldsymbol{W} = \boldsymbol{W} \cup \{T_0\}$;
8: **for** each $X \in$ **adj**(T_0) **do**
9: $Q = Q.push(X)$
10: $PAG(T_0, X) = 1$;$PAG(X, T_0) = 1$;
11: **end for**
12: **end if**
13: /*Step 2: LocalMAG-based local causal orientation*/
14: **for** each X,$Y \in$ **adj**(T_0) **do**
15: **if** $X \perp\!\!\!\perp Y|\mathbf{Z}$ and $X \not\perp\!\!\!\perp Y|\mathbf{Z} \cup T_0$,$\mathbf{Z} \subseteq sep(X, Y)$ **then**
16: $PAG(X, T_0) = 2$;$PAG(Y, T_0) = 2$;
17: ($\boldsymbol{EL_1}$)=FindBI(PAG,T_0,X,$\mathcal{D}$,α);
18: ($\boldsymbol{EL_2}$)=FindBI(PAG,T_0,Y,$\mathcal{D}$,α);
19: **end if**
20: **for** each $V_i \in (\boldsymbol{EL_1} \cup \boldsymbol{EL_2})$ **do**
21: $PAG(V_i, T_0) = 2$;
22: **end for**
23: **end for**
24: Using orient-rule to orient edge orientations in PAG;
25: **until** All causal orientations of T is determined, or $Q = \emptyset$, or $\boldsymbol{W} = \mathbf{U}$, or stop criterion is satisfied
26: Return [PAG];

$$P(T, K \mid L) = \frac{P(T, L, K)}{P(L)} \tag{5}$$

Substitute Eq. (2) into Eq. (5) to obtain:

$$\begin{aligned} P(T, K \mid L) &= \frac{P(T \mid L)P(K \mid L)P(L)}{P(L)} \\ &= P(T \mid L)P(K \mid L) \end{aligned} \tag{6}$$

Thus, according to Eq. (6), when L is observed, T and K are conditionally independent; when L is not observed, T and K are dependent. Since L is a latent variable, we can determine that K is an adjacent variable of T through CI tests. Thus, under the assumption of faithfulness, using the traditional PC discovery algorithm correctly identifies the children of latent variables as the adjacent variables of T.

Through Theorem 3 and Proposition 1, the LatentLCD algorithm can correctly identify the variables in EL(T) as the adjacent variables of T using the MMPC algorithm in preparation for the orientation phase.

Theorem 4. Under the assumptions of causal faithfulness and all CI tests are reliable, the $adj(T)$ discovered by the traditional PC discovery algorithm corresponds to the true adjacent variables of T in the true graph.

Proof: According to Theorem 2, the MAG already contains all the conditional independence relationships among observed variables. Expressed using a formula, as follows:

$$X_i \not\perp\!\!\!\perp T \mid \mathbf{Z}, \mathbf{Z} \subseteq \mathbf{O} \setminus \{X_i, T\} \tag{7}$$

where $\mathbf{O}$ is the set of observed variables, $X_i \in$ PC(T) and PC(T) $\subseteq adj(T)$. Thus, the traditional PC discovery algorithm can accurately discover the PC set of T. According to Proposition 1, we have:

$$P(T, X_j \mid \mathbf{Z}) \neq P(X_j \mid \mathbf{Z})P(T \mid \mathbf{Z}), \mathbf{Z} \subseteq \mathbf{O} \setminus \{X_j, T\} \tag{8}$$

where $X_j \in$ EL(T) and EL(T) $\subseteq adj(T)$. Thus, Proposition 1 further proves that based on these reliable conditional independence relationships, the traditional PC discovery algorithm can also correctly identify EL(T). We demonstrate that Theorem 4 is correct.

Theorem 4 proves that all causal relationships contained in LocalMAG(T) can be found in the skeleton phase using the MMPC algorithm.

LatentLCD first takes as input a data $\mathcal{D}$, a target variable T, a confidence level α, and a set of variables $\mathbf{O}$. Then, a matrix PAG is defined to identify the edge marks between two variables. In the PAG matrix, if $PAG(X, Y) = 1$, it represents $X *\!\!-\!\!\circ Y$; if $PAG(X, Y) = 2$, it represents $X *\!\!\rightarrow Y$; if $PAG(X, Y) = 3$, it represents $X *\!\!-\!\!- Y$; if $PAG(X, Y) = 0$, it represents that the variables X and Y are not adjacent. LatentLCD sets up a queue Q to store variables to be processed, and $\mathbf{W}$ to store variables already processed. Additionally, LatentLCD initializes Q with the target variable T and initializes $\mathbf{W}$ as an empty set.

Finally, the LatentLCD algorithm learns the equivalence class of the MAG, known as the PAG, as shown in Algorithm 1. The detailed steps are as follows:

(1) Establishing the Local Skeleton (lines 4-12): At line 6, the MMPC algorithm is utilized to find the set of adjacent variables of T ($adj(T)$). Then, at lines 8-10, the PAG is updated to establish the undirected local skeleton of T, with both ends of the edges marked as "$\circ$". Subsequently, it adds $\{adj(T) \setminus T\}$ to the Q.

(2) Edge orientation (lines 14-21): Simultaneously identify the V-structures between the T and the variables in its adjacent variable set $adj(T)$, as well as the bidirectional edges between T and the variables in $adj(T)$.

Proposition 2. If there is a latent variable between T and X, $X \in adj(T)$, then X must be a candidate parent of T.

Proof: Assuming there is a latent variable between A and T, meaning $A \leftrightarrow T \leftarrow B$. By analyzing the unshielded triplet $\{A, T, B\}$ using Equation 1 to determine if there is a V-structure, we can confirm the relationships $A *\!\!\rightarrow T \leftarrow\!\!* B$. Thus, A is a candidate parent of T.

Thus, the LatentLCD algorithm first discovers the adjacent set of T, then identifies the candidate parents set of the target, and finally uses the FindBI algorithm to find bidirectional edges. The implementation details are as follows.

a) Identifying V-structures between T and V_i, $V_i \in adj(T)$ (lines 14-16): Traverse each pair of variables $\{X, Y\}$ in $adj(T)$, using Theorem 1 and the definition of m-separation to determine whether the unshielded triplet $\{X, T, Y\}$ forms a V-structure and T is a collider. If T is a collider, change the edge markings from X and Y pointing towards T to "$>$", for example, $X \circ\!\!\rightarrow T$. Thus, add X and Y to the set of candidate parents for T (CPC(T)). $sep(X, Y)$ denotes the separation set for X with respect to Y. b) Identifying bidirectional edges between T and V_i, $V_i \in CPC(T)$

Algorithm 2: FindBI (Find Bidirectional Edges)

Input: $\mathcal{D}$: Data, T: The target variable, α: Significant level, X: Candidate parent of T;
Output: ***EL***: The set of Effects of Latent causes of T;

1: **Initialize:** $\boldsymbol{EL} = \emptyset$
2: **if** $adj(X) = \emptyset$ **then**
3: $adj(X)$=MMPC($\mathcal{D}$,X,α);
4: **end if**
5: **for** each $Y \in adj(X)$ **do**
6: **if** $Y \notin adj(T)$ or $Y \neq T$ **then**
7: **if** $T \perp\!\!\!\perp Y | \mathbf{Z}$ and $T \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup X, \mathbf{Z} \subseteq sep(T, Y)$ **then**
8: $\boldsymbol{EL} = \boldsymbol{EL} \cup X$;
9: **end if**
10: **end if**
11: **end for**
12: Return [***EL***];

(lines 17-18): If an unshielded triplet $\{X, T, Y\}$ is identified as forming a V-structure and T is a collider in lines 14-16, immediately proceed to lines 17-18. Use the FindBI algorithm (Algorithm 2) to identify V-structures in the unshielded triplet $\{T, X, K\}$, $K \in adj(X)$, and confirm that X is a collide. To find existing bidirectional edges, confirming the existence of a latent variable between target T and variable X, thereby including X in EL(T).

(3) Applying orientation rules to update the PAG (line 21): The orientation rules used here are the R1-R4 and R8-R10 criteria designed by (Zhang 2008b). R8-R10 are orientation criteria for updating PAG in the existence of latent variables. The loop continues until the local causal relationships of the T are fully discovered, or the queue Q is empty, or $\mathbf{W} = \mathbf{V}$ is reached, or stop criterion is satisfied. The stop criterion is R3 of Stop Rule proposed by (Xie et al. 2024).

Find Bidirectional Edges: The FindBI Algorithm

In this section, we introduce the Find Bidirectional edges (referred to as FindBI) algorithm, which is used to discover latent variables between T and its adjacent variables, that is, the bidirectional edges connected with T, as shown in Algorithm 2. We have the following proposition for discovering bidirectional edges in the LocalMAG.

Proposition 3. Assuming $\{X, Y\} \subseteq adj(T)$ and $Z \in adj(X)$, if the unshielded triplets $\{X, T, Y\}$ and $\{T, Y, Z\}$ both satisfy the conditions of Eq. 1, and both T and Y are colliders, then $Y \in$ EL(T).

Proof: When the unshielded triplet $\{X, T, Y\}$ is identified as a V-structure using Equation 1, as shown

$$X \perp\!\!\!\perp Y \mid sep(X, Y) \; and \; X \not\perp\!\!\!\perp Y \mid (sep(X, Y) \cup T) \quad (9)$$

with T as the collider, it can be oriented as $X \ast\!\!\rightarrow T \leftarrow\!\!\ast Y$. Similarly, if another unshielded triplet $\{T, Y, Z\}$ is also recognized as a V-structure, as shown

$$T \perp\!\!\!\perp Z \mid sep(T, Z) \; and \; T \not\perp\!\!\!\perp Z \mid (sep(T, Z) \cup Y) \quad (10)$$

with Y as the collider, then it is oriented as $T \ast\!\!\rightarrow Y \leftarrow\!\!\ast Z$. Since both triplets are identified as V-structures and both T and Y are colliders, it can be inferred that $T \leftrightarrow Y$.

For example, as shown in Figure 2(b), $adj(X_1) = \{X_2, X_3\}$, the unshielded triple $\{X_2, X_1, X_3\}$ satisfies $X_2 \perp\!\!\!\perp X_3$ and $X_2 \not\perp\!\!\!\perp X_3 \mid X_1$, identifying the V-structure of $X_2 \ast\!\!\rightarrow X_1 \leftarrow\!\!\ast X_3$. In Figure 2(c), $adj(X_3) = \{X_1, X_4\}$, the unshielded triple $\{X_1, X_3, X_4\}$ satisfies $X_1 \perp\!\!\!\perp X_4$ and $X_1 \not\perp\!\!\!\perp X_4 \mid X_3$, identifying the V-structure of $X_1 \ast\!\!\rightarrow X_3 \leftarrow\!\!\ast X_4$. Thus, we identify the bidirectional edge of $X_1 \leftrightarrow X_3$.

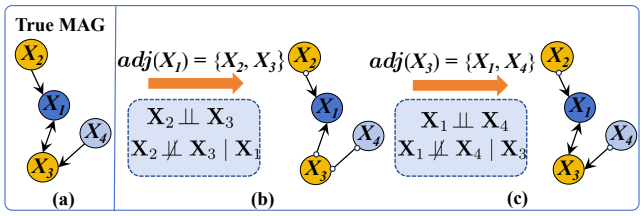

Figure 2: An example showing how Proposition 3 is used.

FindBI uses Proposition 3 to find the bidirectional edges present in LocalMAG, and thus FindBI is a key component of the proposed LatentLCD algorithm. In Algorithm 2, $\mathbf{Z}$ is a conditioning set starting as an empty set, and $sep(T, Y)$ represents the separation set that makes T conditionally independent of Y. FindBI first checks if the $adj(X)$ already exists; if not, it determines $adj(X)$ using the MMPC algorithm to reduce computational complexity. In lines 4-10 of Algorithm 2, it traverses each $Y \in adj(X)$ that is not adjacent to T and uses Eq. 1 to determine if the unshielded triplet $\{T, X, Y\}$ forms a V-structure, with X as the collider. This updates the edge mark from T to X as "$>$". Since lines 19-22 of Algorithm 1 have already identified the edge mark from X to T as "$>$", it can be confirmed that there is a bidirectional edge between T and X, denoted as $T \leftrightarrow X$.

Theorem 5. When causal sufficiency is not satisfied, the LatentLCD algorithm correctly discovers local causal relationships of the target variable.

Proof: Here, we assume that $\mathbf{V} = \mathbf{O} \cup \mathbf{S} \cup \mathbf{L}$ is faithful to a latent causal DAG G, and $\mathbf{S} = \emptyset$, where $\mathbf{O}$, $\mathbf{S}$, and $\mathbf{L}$ respectively represent observed variables, selection bias, and latent variables. According to Theorem 4, we know that the MMPC algorithm can correctly identify the adjacent variables of T, thus ensuring the accuracy of the local skeleton. Based on Theorem 1 and Proposition 3, by identifying V-structures, we can determine the true parents and bidirectional edges, uncovering latent common causes. Additionally, by using the orientation rules of R1-R4 and R8-R10 proposed by (Zhang 2008b), we can orient more edge markings for the local causal skeleton of T. We introduce the stop criterion (Xie et al. 2024) to satisfy consistency with the global structure. Thus, we conclude that the proposed LatentLCD algorithm correctly discovers local causal relationships of the target variable.

Experiments

In this section, we compare the LatentLCD algorithm with the state-of-the-art methods that perform global causal discovery and local causal discovery when the assumption of causal sufficiency is not satisfied. All experiments are conducted on a computer equipped with an Intel Core i7-12700 2.10GHz CPU and 16GB RAM.

Algorithm	Samples	OAcc	ske-F1	CItests	Samples	OAcc	ske-F1	CItests
FCI	500	0.22±0.09	0.77±0.11	(1.7±6.6)×10^5	1000	0.27±0.11	0.80±0.11	(1.3±3.5)×10^5
RFCI		0.33±0.09	0.82±0.08	1224±458		0.40±0.11	0.86±0.08	1444±561
ICD		0.15±0.06	0.45±0.10	**409±81**		0.17±0.06	0.46±0.10	**434±89**
LatentLCD		**0.49±0.09**	**0.86±0.07**	709±181		**0.53±0.09**	**0.88±0.06**	824±187
FCI	1500	0.29±0.13	0.81±0.11	(0.49±1.6)×10^6	2000	0.34±0.11	0.84±0.09	(2.7±9.5)×10^5
RFCI		0.42±0.11	0.87±0.07	1574±667		0.45±0.11	0.89±0.07	1616±712
ICD		0.18±0.06	0.46±0.10	textbf442±96		0.17±0.06	0.44±0.09	**432±85**
LatentLCD		**0.56±0.08**	**0.89±0.06**	917±244		**0.57±0.08**	**0.90±0.06**	934±259
FCI	2500	0.33±0.13	0.84±0.10	(0.56±1.9)×10^6	3000	0.31±0.11	0.83pm0.10	(0.73±2.6)×10^6
RFCI		0.46±0.10	0.90±0.07	1731±760		0.45±0.09	0.89±0.07	1767±837
ICD		0.18±0.06	0.46±0.10	**457±93**		0.19±0.07	0.46±0.10	**456±96**
LatentLCD		**0.58±0.09**	**0.90±0.06**	1015±336		**0.58±0.09**	**0.89±0.06**	1022±310

Table 1: Comparsion of LatentLCD, FCI, RFCI, ICD with 25 nodes in different samples

Algorithm	Samples	OAcc	ske-F1	CItests	Samples	OAcc	ske-F1	CItests
FCI	500	0.21±0.10	0.74±0.10	(0.35±1.2)×10^7	1000	0.26±0.10	0.78±0.09	(1.2±5.0)×10^7
RFCI		0.33±0.09	0.81±0.07	2152±720		0.39±0.09	0.85±0.07	2597±901
ICD		0.15±0.05	0.42±0.09	**778±128**		0.16±0.06	0.42±0.10	**837±141**
LatentLCD		**0.49±0.08**	**0.84±0.06**	1133±276		**0.54±0.08**	**0.87±0.06**	1316±274
FCI	1500	0.28±0.10	0.79±0.10	(1.3±4.3)×10^7	2000	0.30±0.11	0.82±0.09	(0.36±2.0)×10^8
RFCI		0.42±0.09	0.86±0.07	2834±1070		0.44±0.09	0.89±0.06	2858±1146
ICD		0.17±0.06	0.43±0.10	**857±151**		0.18±0.07	0.43±0.09	**854±184**
LatentLCD		**0.57±0.08**	**0.88±0.06**	1451±354		**0.58±0.09**	**0.89±0.05**	1533±318
FCI	2500	0.29±0.11	0.80±0.10	(0.47±2.2)×10^8	3000	0.31±0.11	0.81±0.10	(0.68±4.3)×10^8
RFCI		0.44±0.10	0.87±0.07	3173±1340		0.46±0.09	0.89±0.06	3278±1302
ICD		0.18±0.06	0.42±0.10	**883±168**		0.19±0.06	0.43±0.10	**894±166**
LatentLCD		**0.59±0.07**	**0.88±0.06**	1607±399		**0.60±0.08**	**0.89±0.05**	1671±424

Table 2: Comparsion of LatentLCD, FCI, RFCI, ICD with 35 nodes in different samples

Global Causal Discovery Without Causal Sufficiency

Experimental Setup In all generated datasets, we follow a process similar to RFCI (Colombo et al. 2012) to generate random DAG with latent common causes. By setting a Bernoulli($E(N)/(p'-1)$) distribution, we independently select each element in the upper triangular matrix of the DAG $\mathcal{D}(\mathbf{O}, \mathbf{L}, \mathbf{S} = \emptyset)$, thereby constructing the adjacency matrix $\mathbf{A}$ for the variable set $\mathbf{O} \cup \mathbf{L}$. For each DAG, we identify nodes with at least two children but no parents and randomly select half of them to form $\mathbf{L}$. The remaining nodes comprise $\mathbf{O}$. For subsequent experiments, we randomly generate 100 DAGs for each node count $p' \in \{15, 20, 25, 30, 35\}$, each with $E(N) = 2$. Here, $E(N)$ represents the expected number of neighboring nodes, and p' denotes the total number of nodes in the random DAG. For each DAG, we further generate datasets with sample sizes of $\{500, 1000, 1500, 2000, 2500, 3000\}$. We compare global causal discovery algorithms FCI (Spirtes 2001), RFCI (Colombo et al. 2012), and ICD (Rohekar et al. 2021). The conditional independence test used is Fisher's Z-test, with significant level α set at 0.01.

Evaluation Metrics We use the following metrics to evaluate algorithm: (1) Structural Accuracy: Orientation Accuracy (OAcc) calculates the percentage of correctly oriented edge marks. (2) Skeleton Accuracy: (a) ske-pre: The number of true positive edges divided by the total number of edges. (b) ske-recall: The number of true positive edges divided by the number of true positive edges in the true graph. (c) ske-F1: ske-F1 = 2 $*$ ske-pre $*$ ske-recall/(ske-pre + ske-recall). (3) Efficiency: Since RFCI is implemented in R and LatentLCD is implemented in Matlab, we use the number of CI tests to measure the efficiency of the algorithms. Both the LatentLCD algorithm and the comparative algorithms output a PAG, and we evaluate the accuracy of the algorithms by comparing the algorithm-generated PAG with the true MAG.

Experiment Result Due to space limitations, only partial results are presented in Tables 1 and 2. The complete experimental results are in Appendix A. Based on these results, we draw the following conclusions: In terms of skeleton accuracy, LatentLCD consistently outperforms other algorithms across all sample sizes for ske-F1 and ske-recall, achieving the highest scores. In terms of structural accuracy, LatentLCD significantly outperforms other comparison algorithms in orientation accuracy across all samples. For example, in Table 1, at 3000 samples, LatentLCD increases orientation accuracy by 27%, 13%, and 38%, over FCI, RFCI, and ICD, respectively. In Table 2, at 3000 samples, LatentLCD improves orientation accuracy by 27%, 15%, and 41%, over FCI, RFCI, and ICD, respectively.

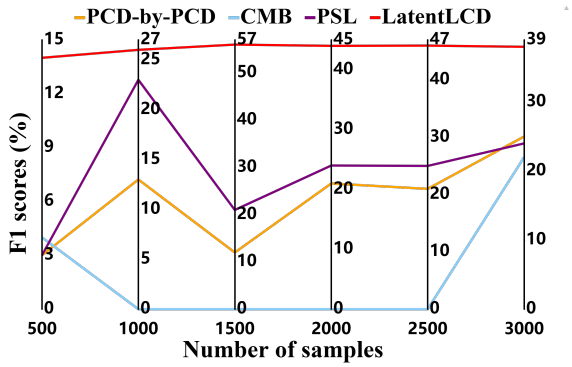

Figure 3: The F1 scores (%) of LatentLCD and its competitors when target is "DG25" on six datasets.

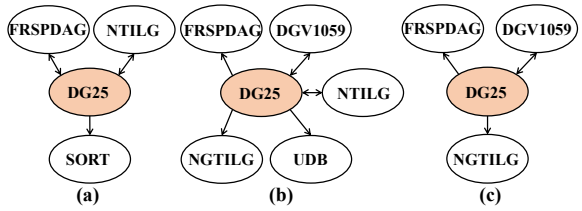

Figure 4: LocalMAG of "DG25" discoverd by LatentLCD. (a) Samples 500. (b) Samples 3000. (c) True graph.

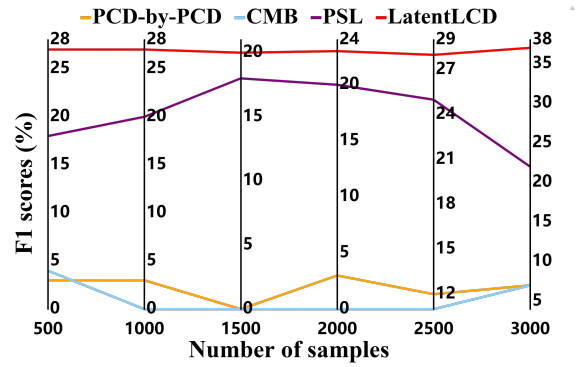

Figure 5: The F1 scores (%) of LatentLCD and its competitors when target is "FRSPDAG" on six datasets.

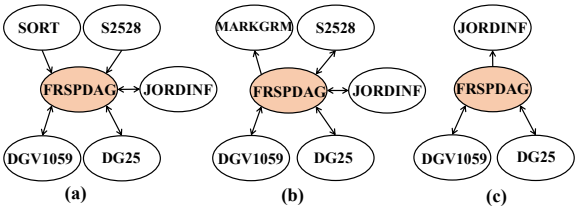

Figure 6: LocalMAG of "FRSPDAG" discoverd by LatentLCD. (a) Samples 500. (b) Samples 3000. (c) True graph.

Local Causal Discovery Without Causal Sufficiency

In this section, we compare the LatentLCD algorithm with other local causal discovery algorithms on benchmark network data. These compared local causal discovery algorithms satisfy the causal sufficiency assumption.

Experimental Setup We select a discrete Bayesian network, BARLEY, which consists of 48 nodes and 84 arcs. For this network, we treat "SAATID" as a latent variable, with all other variables considered as observed variables. Using the BARLEY network, we randomly generate six sets of data with sample sizes of {500, 1000, 1500, 2000, 2500, 3000}, each containing ten datasets. Then, we remove "SAATID" from the generated datasets. We perform local causal discovery using "DGV1059", "DG25", and "FRSPDGA" as target nodes. We compare local causal discovery algorithms PCD-by-PCD (Yin et al. 2008), CMB (Gao and Ji 2015), and PSL (Ling et al. 2022). We employ the G^2 test, setting the significant level α at 0.01. Because the output of the LatentLCD algorithm under LocalMAG is in the form of a PAG, whereas the output of the compared local causal discovery algorithms consists of the PC set of a given target variable, we convert the output of the LatentLCD algorithm into the Pa, Ch, and extra EL of the target node. If PAG(X_i,T)=2 and PAG(T,X_i)=3, $X_i \in$ Pa(T); If PAG(T,X_i)=2 and PAG(X_i,T)=3, $X_i \in$ Ch(T); If PAG(X_i,T)=2 and PAG(T,X_i)=2, $X_i \in$ EL(T).

Evaluation Metrics We evaluate the algorithms based on the following metrics. (1) F1: F1 = 2 $*$ Precision $*$ Recall/(Precision + Recall). (2) Precision: The number of variables in the output that belong to the true LocalMAG(T) divided by the total number of edges output by the algorithm. (3) Recall: The number of true positives divided by the number of true positive edges in the true graph. (4) CItests: The number of CI tests.

Experiment Result We visualize the results in Figures 3 and 5 and draw the following conclusions: When the target is "DG25", as shown in Figure 3, LatentLCD achieves higher F1 scores than other algorithms. Specifically, for a sample size of 1500, LatentLCD increases F1 scores by 44%, 56%, and 35%, over PCD-by-PCD, CMB, and PSL, respectively. When the target is "FRSPDGA", as shown in Figure 5, LatentLCD achieves the best F1 scores. Specifically, for a sample size of 3000, LatentLCD increases F1 scores by 30%, 30%, and 15%, over PCD-by-PCD, CMB, and PSL, respectively. Figure 4 and Figure 6 show the LocalMAG of the target discovered by LatentLCD. The complete experimental results are in Appendix B.

Conclusion

In this paper, we first propose the LocalMAG to describe the local causal relationships of the target variable in the MAG. Then, we propose a novel local causal discovery algorithm without causal sufficiency, LatentLCD, to determine if the target and its neighboring nodes are all colliders to identify bidirectional edges, uncovering latent common causes. The experimental results validate that LatentLCD outperforms the state-of-the-art algorithms in terms of accuracy. In future research, we will explore designing a causal discovery algorithm that transitions from a local-to-global algorithm using LatentLCD when causal sufficiency is not satisfied.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (under grant 2021ZD0111801), the National Natural Science Foundation of China (under grant 62306002, 62272001, 62176001, 62376087, and 62120106008), and the Natural Science Project of Anhui Provincial Education Department (under grant 2023AH030004), and Xunfei Zhiyuan Digital Transformation Innovation Research Special for Universities (2023ZY001).

References

Cai, R.; Zhang, Z.; and Hao, Z. 2013. Sada: A general framework to support robust causation discovery. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 208–216.

Cheng, D.; Li, J.; Liu, L.; Liu, J.; and Le, T. D. 2024. Data-driven causal effect estimation based on graphical causal modelling: A survey. *ACM Computing Surveys*, 56(5): 1–37.

Cheng, D.; Li, J.; Liu, L.; Yu, K.; Le, T. D.; and Liu, J. 2022a. Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 6108–6120.

Cheng, D.; Li, J.; Liu, L.; Yu, K.; Lee, T. D.; and Liu, J. 2022b. Discovering Ancestral Instrumental Variables for Causal Inference from Observational Data. arXiv:2206.01931.

Chickering, D. M. 1996. Learning Bayesian networks is NP-complete. *Learning from data: Artificial intelligence and statistics V*, 121–130.

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

Gao, T.; Fadnis, K.; and Campbell, M. 2017. Local-to-global Bayesian network structure learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1193–1202.

Gao, T.; and Ji, Q. 2015. Local causal discovery of direct causes and effects. In *proceedings of twenty-eighth Neural Information Processing Systems*, volume 28, 2512–2520.

Kuang, K.; Cui, P.; Li, B.; Jiang, M.; Yang, S.; and Wang, F. 2017. Treatment effect estimation with data-driven variable decomposition. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, volume 31, 140–146.

Ling, Z.; Yu, K.; Liu, L.; Li, J.; Zhang, Y.; and Wu, X. 2022. PSL: An algorithm for partial Bayesian network structure learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5): 1–25.

Margaritis, D.; and Thrun, S. 1999. Bayesian Network Induction via Local Neighborhoods. In *Proceedings of 12th Neural Information Processing Systems*, volume 12, 505–511.

Meek, C. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th conference on Uncertainty in artificial intelligence*, 403–410.

Nogueira, A. R.; Pugnana, A.; Ruggieri, S.; Pedreschi, D.; and Gama, J. 2022. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2): e1449.

Pearl, J. 2014. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pearl, J. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv:1801.04016.

Richardson, T.; and Spirtes, P. 2002. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962–1030.

Rohekar, R. Y.; Nisimov, S.; Gurwicz, Y.; and Novik, G. 2021. Iterative causal discovery in the possible presence of latent confounders and selection bias. In *Proceeding of thirty-fourth Neural Information Processing Systems*, volume 34, 2454–2465.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Spirtes, P. 2001. An anytime algorithm for causal inference. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3, 278–285.

Statnikov, A.; Ma, S.; Henaff, M.; Lytkin, N.; Efstathiadis, E.; Peskin, E. R.; and Aliferis, C. F. 2015. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *The Journal of Machine Learning Research*, 16(1): 3219–3267.

Tsamardinos, I.; Aliferis, C. F.; and Statnikov, A. 2003. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 673–678.

Wang, C.; Zhou, Y.; Zhao, Q.; and Geng, Z. 2014. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Computational statistics & data analysis*, 77: 252–266.

Wang, Y.; Cao, F.; Yu, K.; and Liang, J. 2023. Local causal discovery in multiple manipulated datasets. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10): 7235–7247.

Xie, F.; Li, Z.; Wu, P.; Zeng, Y.; Chunchen, L.; and Geng, Z. 2024. Local Causal Structure Learning in the Presence of Latent Variables. In *Forty-first International Conference on Machine Learning*, volume 235, 54511–54530.

Yin, J.; Zhou, Y.; Wang, C.; He, P.; Zheng, C.; and Geng, Z. 2008. Partial orientation and local structural learning of causal networks for prediction. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3, 93–105.

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.; and Chen, H. 2018. Mining Markov blankets without causal sufficiency. *IEEE transactions on neural networks and learning systems*, 29(12): 6333–6347.

Zhang, J. 2008a. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9(7): 1437–1474.

Zhang, J. 2008b. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896.