7-2022

Identification of Linear Non-Gaussian Latent Hierarchical Structure

Feng Xie
Peking University & Beijing Technology and Business University

Biwei Huang
Carnegie Mellon University

Zhengming Chen
Guangdong University of Technology

Yangbo He
Peking University

Zhi Geng
Beijing Technology and Business University

See next page for additional authors

Identification of Linear Non-Gaussian Latent Hierarchical Structure

Feng Xie [1 2] **Biwei Huang** [3] **Zhengming Chen** [4] **Yangbo He** [1] **Zhi Geng** [2] **Kun Zhang** [5 3]

Abstract

Traditional causal discovery methods mainly focus on estimating causal relations among measured variables, but in many real-world problems, such as questionnaire-based psychometric studies, measured variables are generated by latent variables that are causally related. Accordingly, this paper investigates the problem of discovering the hidden causal variables and estimating the causal structure, including both the causal relations among latent variables and those between latent and measured variables. We relax the frequently-used measurement assumption and allow the children of latent variables to be latent as well, and hence deal with a specific type of latent hierarchical causal structure. In particular, we define a minimal latent hierarchical structure and show that for linear non-Gaussian models with the minimal latent hierarchical structure, the whole structure is identifiable from only the measured variables. Moreover, we develop a principled method to identify the structure by testing for Generalized Independent Noise (GIN) conditions in specific ways. Experimental results on both synthetic and real-world data show the effectiveness of the proposed approach.

1. Introduction

Inferring causal relationships from observational (non-experimental) data is challenging when there exist unobserved confounders. One typical strategy for handling this problem is by making use of conditional independence relations to learn the causal graph over the observed variables up to an equivalence class (Pearl, 2009; Spirtes et al., 2000). Well-known algorithms along this line include FCI (Spirtes et al., 1995), RFCI (Colombo et al., 2012), and FCI+ (Claassen et al., 2013). Another strategy is to make use of functional causal model-based approaches in the linear non-Gaussian setting (Hoyer et al., 2008; Entner & Hoyer, 2010; Chen & Chan, 2013; Tashiro et al., 2014; Wang & Drton, 2020; Salehkaleybar et al., 2020; Maeda & Shimizu, 2020; Chen et al., 2021). These works focus on estimating the causal relationships among observed variables rather than those among latent variables. However, in some real-world scenarios, researchers are usually interested in the causal relationships between latent variables—the observed variables may not necessarily be the underlying causal variables (Bartholomew et al., 2008).

A classical framework for inferring latent factors is Factor Analysis (Bartholomew et al., 2008). But with factor analysis-based approaches, the estimated factors may not be the underlying causal variables and their relations are usually not modeled (Silva et al., 2006). Silva et al. (2006) proposed a two-step approach to learn the measurement model and the causal structure among latent variables, by utilizing the vanishing Tetrad conditions (Spearman, 1928). Later, Kummerfeld & Ramsey (2016) developed the FindOneFactorClusters (FOFC) algorithm, based on the extended t-separation theorem (Sullivant et al., 2010; Spirtes, 2013), to learn the pure measurement model. Beyond the second-order statistics, Shimizu et al. (2009) leveraged non-Gaussianity and showed that a linear acyclic model for latent factors is identifiable. Cai et al. (2019) proposed a Triad condition and accordingly developed an LSTC algorithm to discover the structure over latent variables with non-Gaussian distributions. Xie et al. (2020) designed a Generalized Independent Noise (GIN) condition to address more general cases where there may be multiple latent variables behind any pair of observed variables. Other interesting developments along this line have been established (Zeng et al., 2021; Chen et al., 2022). However, the above methods assume that each latent variable set has a much larger number of observed variables as children and cannot handle the situation with latent hierarchical structure (i.e., the children of latent variables may still be latent). For instance, consider a hierarchical latent model illustrated in Figure 1, where the variables L_i, $i = 1, ..., 9$ are unobserved and X_j,

[1]Department of Probability and Statistics, Peking University, Beijing, China [2]Department of Applied Statistics, Beijing Technology and Business University, Beijing, China [3]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA [4]School of Computer Science, Guangdong University of Technology, Guangzhou, China [5]Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Feng Xie <xiefeng009@gmail.com>, Kun Zhang <kunz1@cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

$j = 1, ..., 15$, are observed. The above methods generally fail to discover the latent variable L_1.

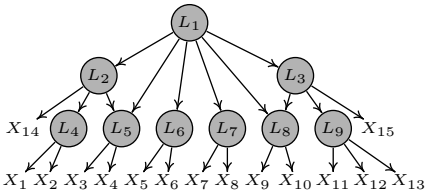


Figure 1. A hierarchical causal structure involving 9 latent variables (shaded nodes) and 15 observed variables (unshaded nodes).

Several contributions have been made to learn the latent hierarchical structure other than the measurement model. For instance, Zhang (2004) generalized the classic latent cluster models and proposed hierarchical latent class models (also known as latent tree models) for discrete variables. Poon et al. (2010) extended this model and proposed Pouch Latent Tree Models, which allow each leaf node to consist of one or more continuous observed variables. Later, Choi et al. (2011) proposed more general latent tree models for both discrete and Gaussian random variables and provided efficient estimation algorithms. Other interesting developments along this line include (Harmeling & Williams, 2010; Mourad et al., 2013; Zhang & Poon, 2017; Etesami et al., 2016; Drton et al., 2017). Although these methods have been used in a range of fields, they usually assume a tree-structured graph, i.e., there is only one path between every pair of variables in the system. However, in many settings, e.g., with the structure in Figure 1, it may be violated.

In this paper, we consider causal structure identification in a more challenge scenario where the variables form a hierarchical structure and some latent variables may have no observed variables as children. Recently, Adams et al. (2021) established necessary and sufficient conditions for structure identifiability in linear non-Gaussian setting, which is exciting. However, it does not provide a practical estimation procedure, which we aim to achieve in this work. Besides, their work need to give the number of latent factors at the beginning while our work does not need this information. Specifically, we make the following contributions:

1. We introduce a constrained causal structure involving latent variables, the *minimal latent hierarchical structure*, under which the hierarchical causal structure is identifiable without including any redundant latent nodes.
2. We propose an efficient algorithm for estimating the latent hierarchical structure by using Generalized Independent Noise (GIN) conditions. The proposed algorithm can recover the correct structure asymptotically, including both the causal relations among latent variables and those between latent and observed variables, under the linear, non-Gaussian model assumption.
3. We demonstrate the efficacy of our algorithm on both synthetic and real-word data.

2. Latent Hierarchical Causal Model

We first give notations and graph terminologies that will be used throughout the paper. Then we provide the definition of the Linear Non-Gaussian Latent Hierarchical Model and corresponding graphical constraints, under which the graph structure is identifiable.

2.1. Notation and Graph Terminology

Denote by $\mathcal{G}$ a Directed Acyclic Graph (DAG) with a set of variables $\mathbf{V}$ and a set of directed edges $\mathbf{E}$. The set of parents and children of V_i are denoted by $\mathbf{Pa}(V_i)$ and $\mathbf{Ch}(V_i)$, respectively. Furthermore, $\mathbf{Pa}(\mathbf{Y})$ denotes the set of common parents of a variable set $\mathbf{Y}$, and $|\mathbf{Y}|$ denotes the number of elements in the set $\mathbf{Y}$. Other commonly-used concepts in graphical models, such as path and d-separation, can be found in Pearl (2009); Spirtes et al. (2000).

2.2. Linear Non-Gaussian Latent Hierarchical Model

In this paper, we focus on a particular type of linear non-Gaussian causal model with variables $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$, where each observed variable $X_i \in \mathbf{X}$ and latent variable $L_j \in \mathbf{L}$ are generated according to the following linear structural equation models:

$$X_i = \sum_{L_j \in Pa(X_i)} b_{ij} L_j + \varepsilon_{X_i}, \tag{1}$$

$$L_j = \sum_{L_k \in Pa(L_j)} c_{jk} L_k + \varepsilon_{L_j}, \tag{2}$$

where b_{ij} and c_{jk} represent the causal strength from L_j to X_i and from L_k to L_j, respectively. All noise terms ε_{X_i} and ε_{L_j} are continuous random variables with non-Gaussian distributions, and are independent of each other. We assume that the generating process is recursive (Bollen, 1989). That is to say, the causal relationships over variables $\mathbf{V}$ can be represented by a DAG (Pearl, 2009; Spirtes et al., 2000).

Definition 1 (Linear Non-Gaussian Latent Hierarchical Model (LiNGLaH)). *A model is called a linear non-Gaussian latent hierarchical model, with graph structure $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, if*

- $\mathbf{V} = \mathbf{X} \bigcup \mathbf{L}$, *where $\mathbf{X}$ is the set of observed variables and $\mathbf{L}$ is the set of latent variables,*
- *each variable in $\mathbf{X}$ and $\mathbf{L}$ is generated by the structural equation models (1) and (2), respectively, and*
- *the distribution over $\mathbf{V}$ is Markov and faithful to the DAG $\mathcal{G}$.*

Here, the non-Gaussinaity of noise terms is essential to identifying causal directions between two variables in a

linear model and has been extensively studied in recent years (Shimizu et al., 2006; Shimizu, 2019). Moreover, it has been argued that the non-Gaussian distributions are ubiquitous (Spirtes & Zhang, 2016). In this paper, our goal is to establish the identifiability of the latent hierarchical structure and show how to estimate it only from observed variables $\mathbf{X}$.

2.3. Structural Conditions for Identifiability

It is noteworthy that one may not be able to determine the locations and number of the latent nodes in LiNGLaH without further assumptions. Several approaches have attempted to handle this issue under specific assumptions, e.g., the measurement assumption (each latent variable L_i has a certain number of pure measurement variables as children [1]). Representative methods along this line include BPC (Silva et al., 2006), noisy ICA-based method (Shimizu et al., 2009), FOFC (Kummerfeld & Ramsey, 2016), CFPC algorithm (Cui et al., 2018), and GIN (Xie et al., 2020). In this paper, we consider a clearly more general scenario where some latent variables have only latent variables as children (i.e., no observed children), such as the latent variable L_1 in Figure 1. In the following, we will give a sufficient graphical condition that render the causal structure of a latent hierarchical model identifiable. Specifically, with this condition the structure among latent variables does not include any "redundant" latent nodes, and we call such structure the *Minimal Latent Hierarchical Structure*.

Condition 1 (Minimal Latent Hierarchical Structure). *A structure is a minimal latent hierarchical structure if (1) each latent variable has at least three neighbors, and (2) each latent variable has at least two pure children (which can be either latent or observed).*

Note that this condition is milder than that in the minimal latent tree model (Pearl, 1988; Choi et al., 2011), where the latent tree model only allows one path between any two latent variables, while here we do not have such a restriction. Figure 2(a) shows an example of a minimal latent hierarchical structure satisfying Condition 1. In contrast, the structure in Figure 2(b) does not satisfy Condition 1 because the number of neighbor nodes of both L_6 and L_7 is fewer than 3. Intuitively, for L_6, all paths from L_6 to its observable descendants $\{X_3, X_4\}$ go through L_3 and L_6 does not have no additional and unique observable descendant relative to L_3 to help recover L_6. This implies that L_6 is a redundant variable. A similar reasoning procedure applies to L_7.

In the next section, we will propose an algorithm to estimate the structure of LiNGLaH from observed variables $\mathbf{X}$ and show that the graph structure of LiNGLaH is fully identifiable under Condition 1.

[1] A set $\mathbf{C}$ is the set of pure children (measurement variables) of L_i if each node in $\mathbf{C}$ has only one latent parent L_i, and each node in $\mathbf{C}$ is neither the cause nor the effect of other nodes in $\mathbf{C}$.

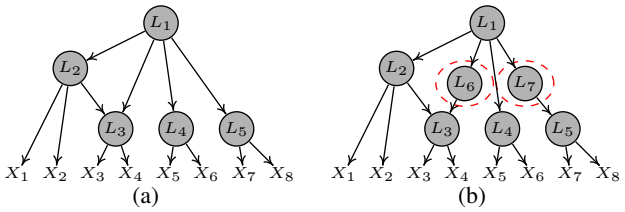


Figure 2. (a) An example of the minimal latent hierarchical structure, where the whole structure is identifiable. (b) A counter-example of the minimal latent hierarchical structure, where the structure is not identifiable because L_6 and L_7 have neighbor nodes fewer than 3.

3. Structure Identification of LiNGLaH

In this section, we propose an efficient algorithm, **La**tent **H**ierarchical **M**odel **E**stimation (LaHME), for estimating the structure of LiNGLaH. The LaHME algorithm consists of two steps. It first locates all latent variables (Step 1), and then it infers the causal structure among the identified latent variables (Step 2). The details of these two steps are described in Section 3.1 and 3.2, respectively. Furthermore, we analyze the complexity of Steps 1 and 2, respectively in Section 3.3. Finally, in Section 4, we show the soundness of the LaHME algorithm; that is, it outputs the correct causal structure asymptotically.

Before describing the identification algorithm, we first produce the GIN condition (Xie et al., 2020).

Definition 2 (GIN condition). *Let $\mathbf{Z}$ and $\mathbf{Y}$ be sets of variables in a linear non-Gaussian acyclic causal model. We say that $(\mathbf{Z}, \mathbf{Y})$ follows GIN condition if and only if $\omega^\intercal \mathbf{Y}$ are statistically independent of $\mathbf{Z}$, where ω satisfies $\omega^\intercal \mathbb{E}[\mathbf{Y}\mathbf{Z}^\intercal] = 0$ and $\omega \neq 0$.*

The GIN condition is essential to locating latent variables and identifying structure among latent variables in the LaHME algorithm.

3.1. Step 1: Locating Latent Variables

We adopt a recursive procedure to locate latent variables from observed variables. More specifically, at each iteration, it contains the following three phases:

- P1. Identifying the group of variables that share the same set of latent parents (we call such group a global causal cluster [2]) from the ***active variable set*** [3].
- P2. Determining the number of new latent variables that need to be introduced for these clusters.
- P3. Updating the active variable set.

[2] Please refer to the Section 3.1.1 for a precise description.

[3] We say a set is active if it is selected in the current iteration.

The three phases P1 $\sim$ P3 are repeated iteratively until all latent variables of the system are discovered. The complete procedure is summarized in Algorithm 1, and an illustrative example for each phase will be given immediately after introducing each phase in the following subsections, and a complete example is given in Appendix D.

Algorithm 1 LocateLatentVariables

Input: A set of observed variables $\mathbf{X}$
Output: Partial causal structure $\mathcal{G}$
Initialize active set $\mathcal{A} \coloneqq \mathbf{X}$, and let $\mathcal{G} = \emptyset$;
while $\mathcal{A} \neq \emptyset$ **do**
  ClusterList $\leftarrow$ FindGlobalCausalClusters($\mathcal{A}$); // *P1*
  $(\mathcal{L}, \mathcal{G}) \leftarrow$ DetermineLatentVariables(ClusterList, $\mathcal{A}$, $\mathcal{G}$); // *P2*
  $\mathcal{A} \leftarrow$ UpdateActiveData($\mathcal{L}$, $\mathcal{A}$, $\mathcal{G}$) // *P3*
end while
Return: $\mathcal{G}$

3.1.1. P1: Finding Global Causal Clusters

In this section, we deal with the identification of global causal clusters. We first give the definitions of causal cluster and global causal cluster that help quickly locate the latent variables, with the global causal cluster as a special kind of causal cluster .

Definition 3 (Causal Cluster & Global Causal Cluster)**.** *Let $\mathcal{A}$ be the active variable set. We say a set $\mathbf{C}_1 \subset \mathcal{A}$ is a* causal cluster *if the variables in $\mathbf{C}_1$ share the same latent parent, denoted by L_1. Furthermore, $\mathbf{C}_1$ is a* global causal cluster *if (1) $\mathbf{C}_1$ and $\mathcal{A}\backslash\mathbf{C}_1$ are d-separated by L_1, and (2) there is no proper subset $\tilde{\mathbf{C}}_1 \subset \mathbf{C}_1$ such that $\tilde{\mathbf{C}}_1$ and $\mathcal{A}\backslash\tilde{\mathbf{C}}_1$ are d-separated by L_1.*

Definition 4 (Pure Causal Cluster)**.** *Let $\mathcal{A}$ be the active variable set and that the set $\mathbf{C}_1 \subset \mathcal{A}$ be a causal cluster. We say $\mathbf{C}_1$ is a pure causal cluster if any variable in $\mathbf{C}_1$ is neither the cause nor the effect of other variables in $\mathbf{C}_1$. Otherwise, $\mathbf{C}_1$ is an impure causal cluster.*

For instance, consider the causal structure in Figure 1. 65Suppose the active variable set $\mathcal{A} = \mathbf{X}$ and $\mathbf{C}_1 = \{X_1, X_2\}$. $\mathbf{C}_1$ is a causal cluster and its common latent parent is L_1. In addition, $\mathbf{C}_1$ is a global cluster because $\mathbf{C}_1$ and $\mathcal{A}\backslash\mathbf{C}_1$ are d-separated by L_1. Suppose the active variable set $\mathcal{A} = \{L_1, L_2, L_3, L_5, ..., L_8\}$ and $\mathbf{C}_1 = \{L_2, L_6, L_7\}$. $\mathbf{C}_1$ is a causal cluster and its common latent parent is L_1. However, $\mathbf{C}_1$ is not a global causal cluster because $\mathbf{C}_1$ and $\mathcal{A}\backslash\mathbf{C}_1$ are not d-separated by L_1 (there is a directed path between L_2 and L_5). Furthermore, $\mathbf{C}_1$ is a pure cluster because any variable in $\mathbf{C}_1$ is neither the cause nor the effect of other variables in $\mathbf{C}_1$.

The global causal clusters in the current active variable set $\mathcal{A}$ can be identified by appropriately testing for the GIN condition, as formally stated in the following proposition.

Proposition 1 (Identifying Global Causal Clusters)**.** *Let $\mathcal{A}$ be the active variable set and $\mathbf{Y}$ be a proper subset of $\mathcal{A}$. Then $\mathbf{Y}$ is a global causal cluster if and only if the following two conditions hold: 1) for any subset $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$ with $|\tilde{\mathbf{Y}}| = 2$, $(\mathcal{A}\backslash\mathbf{Y}, \tilde{\mathbf{Y}})$ follows the GIN condition, and 2) no proper subset of $\mathbf{Y}$ satisfies condition 1).*

To identify global causal clusters efficiently, we start with finding clusters with size $|\mathbf{Y}| = 2$, and then increase the size of group $\mathbf{Y}$ until it satisfies condition 1 of Proposition 1 or $|\mathbf{Y}| = |\mathcal{A}| - 1$. The details of the search procedure are given in Algorithm 2, and an illustrative example is given accordingly.

Algorithm 2 FindGlobalCausalClusters

Input: A set of active variables $\mathcal{A}$
Output: ClusterList
Initialize cluster set ClusterList $= \emptyset$ and the group size GrLen $= 2$;
while $|\mathcal{A}| \geq$ GrLen $+ 1$ **do**
  repeat
    Select a subset $\mathbf{Y}$ from $\mathcal{A}$ such that $|\mathbf{Y}| =$ GrLen;
    if $(\mathcal{A}\backslash\mathbf{Y}, \tilde{\mathbf{Y}})$ follows GIN condition for any $\tilde{\mathbf{Y}} \in \mathbf{Y}$ with $|\tilde{\mathbf{Y}}| = 2$ **then**
      Add $\mathbf{Y}$ into ClusterList;
    end if
  until all subsets with size GrLen in $\mathcal{A}$ selected;
  $\mathcal{A} = \mathcal{A}\backslash$ClusterList, and GrLen $\leftarrow$ GrLen $+ 1$;
end while
Return: ClusterList

Example 1. *Consider the causal structure in Figure 1. Suppose the active variable set is $\mathcal{A} = \{X_1, ..., X_{15}\}$. Setting* GrLen $= 2$*, one can find 8 clusters: $\mathbf{C}_1 = \{X_1, X_2\}$, $\mathbf{C}_2 = \{X_3, X_4\}$, $\mathbf{C}_3 = \{X_5, X_6\}$, $\mathbf{C}_4 = \{X_7, X_8\}$, $\mathbf{C}_5 = \{X_9, X_{10}\}$,$\mathbf{C}_6 = \{X_{11}, X_{12}\}$,$\mathbf{C}_7 = \{X_{11}, X_{13}\}$, and $\mathbf{C}_8 = \{X_{12}, X_{13}\}$.*

3.1.2. P2: Determining Latent Variables

We then determine how many new latent variables need to be introduced for these clusters identified in Algorithm 2. To this end, we will deal with the following two issues:

- which clusters of variables share the common latent parent and should be merged, and
- which clusters of variables are the children of the latent variables that have been introduced in previous iterations.

We first give the following lemma on identifying pure and impure clusters with GIN conditions, which will help us address the above two issues.

Lemma 1 (Identifying Pure/Impure Clusters)**.** *Let $\mathcal{A}$ be the active variable set and $\mathbf{C}_1$ be a global causal cluster. Then the following statements hold: (1) If $|\mathbf{C}_1| > 2$, $\mathbf{C}_1$ is an*

impure cluster; (2) If $|\mathbf{C}_1| = 2$ *and for any* $V_i, V_j \in \mathbf{C}_1$, *there does not exist* $V_k, V_t \in \mathcal{A} \cup \mathbf{Ch}(\mathcal{A}) \backslash \mathbf{C}_1$ *such that* $(\{V_i, V_t\}, \{V_i, V_j, V_k\})$ *follows the GIN condition while* $(\{V_j, V_t\}, \{V_i, V_j, V_k\})$ *violates the GIN condition, then* $\mathbf{C}_1$ *is a pure cluster. Otherwise,* $\mathbf{C}_1$ *is an impure cluster.*

We now provide the conditions under which the clusters of variables share the common latent parent and should be merged to solve the first issue.

Proposition 2 (Merging Rules)**.** *Let* $\mathcal{A}$ *be the active variable set and* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *be two global causal clusters.* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *share the common latent parent, if one of the following rules holds.*

- $\mathcal{R}1$. *1)* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *are both pure clusters, and 2) for any subset* $\tilde{\mathbf{C}} \subseteq \mathbf{C}_1 \cup \mathbf{C}_2$ *with* $|\tilde{\mathbf{C}}| = 2$, $(\mathcal{A} \backslash \tilde{\mathbf{C}}, \tilde{\mathbf{C}})$ *follows the GIN condition.*
- $\mathcal{R}2$. *1) One of the clusters is a pure cluster and the other is not, e.g.,* $\mathbf{C}_1$ *is pure and* $\mathbf{C}_2$ *is impure, and 2) for any variable* $V_i \in \mathbf{C}_1$ *and any variable* $V_j \in \mathbf{C}_2$, $(\mathcal{A} \backslash \{\mathbf{C}_2, V_i\}, \{V_i, V_j\})$ *follows the GIN condition.*
- $\mathcal{R}3$. *1)* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *both are impure clusters, and 2) for any subset* $\tilde{\mathbf{C}} \subseteq \mathbf{C}_1 \cup \mathbf{C}_2$ *with* $|\tilde{\mathbf{C}}| = 2$, $(\mathcal{A} \backslash \{\mathbf{C}_1 \cup \mathbf{C}_2\}, \tilde{\mathbf{C}})$ *follows the GIN condition.*

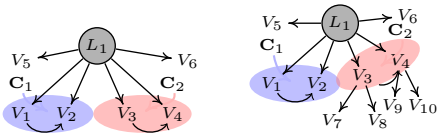
Otherwise, $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *do not share the common latent parent.*

Next, we discuss the solution for the second issue. Due to the property of hierarchical structure, we can not guarantee that all children of a latent variable are identified at the same iteration. Thus, we need to identify whether a new cluster's parents have been introduced in previous iterations. Fortunately, for any latent variable L_1 that was introduced in previous iterations, we know that all nodes in the active variable set $\mathcal{A}$ in the current iteration are causally earlier than the children of L_1 found in previous iteration. That is to say, $\mathbf{Ch}(L_1)$ are leave nodes in the subgraph with variables $\mathcal{A} \cup \mathbf{Ch}(L_1)$. This yields the following corollary derived from Proposition 2.

Corollary 1. *Let* L_1 *be a latent variable that was introduced in previous iterations,* $\mathbf{C}_2$ *be a new cluster, and* $\mathcal{A}$ *be the active variable set in the current iteration. Suppose cluster* $\mathbf{C}_1$ *was a subset of* $\mathbf{Ch}(L_1)$ *found in previous iterations. Then* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *share the common latent parent* L_1 *if setting* $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1 \backslash L_1$ *be the active set, one of the three rules in Proposition 2 holds. Otherwise,* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *do not share the common latent parent.*

Below, we give an example to illustrate rule $\mathcal{R}3$ in Proposition 2 and Corollary 1 to identify the clusters of variables that share a common latent parent. Please see Appendix C for more analyses of the three rules.

Example 2 (**Rule** 3)**.** *Consider the causal graphs in Figure 3. We first check* $\mathcal{R}3$ *of Proposition 2 in subgraph (a), where clusters* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *are identified in the same iteration and they are two impure causal clusters. Let* $\mathcal{A} = \{V_1, ..., V_6\}$. *For any subset of* $\mathbf{C}_1 \cup \mathbf{C}_2$, *e.g.,* $\tilde{\mathbf{C}}_1 = \{V_1, V_3\}$, *we have* $(\{V_5, V_6\}, \{V_1, V_3\})$ *follows the GIN condition. This implies that* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *share the same latent parent* L_1. *We next check Corollary 1 in subgraphs (b), where* L_1 *is a latent variable introduced in the first iteration,* $\mathbf{C}_1$ *is a subset of its children,* $\mathbf{C}_2$ *is a new causal cluster, and* $\mathcal{A} = \{L_1, V_3, ..., V_6\}$. *We first set* $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1 \backslash L_1 = \{V_1, ..., V_6\}$. *Then, we check* $\mathcal{R}3$ *and find that* $\mathbf{C}_1$ *and* $\mathbf{C}_2$ *share the common parent.*



(a) (b)

Figure 3. The illustrative examples for $\mathcal{R}3$ in Proposition 2 and Corollary 1.

The complete procedure of determining latent variables for the current active variable set is summarized in Algorithm 3, and an illustrative example is given in Example 3.

Algorithm 3 DetermineLatentVariables

Input: ClusterList, $\mathcal{A}$, and Partial structure $\mathcal{G}$
Output: New latent set $\mathcal{L}$ and partial structure $\mathcal{G}$

Initialize $\mathcal{G}' = \mathcal{G}$ and $\mathcal{L} = \emptyset$;
$\mathbf{C} \leftarrow$Merge clusters from ClusterList according to Rules $\mathcal{R}1 \sim \mathcal{R}3$ of Proposition 2;
for each $\mathbf{C}_i \in \mathbf{C}$ **do**
  if there exists $L_j \in \mathcal{G}'$ such that $\mathbf{C}_i$ and L_j satisfy the conditions of Corollary 1 **then**
    $\mathcal{G} = \mathcal{G} \cup \{L_j \to V_i | V_i \in \mathbf{C}_i\}$;
  else
    Introduce a new latent variable L_k into $\mathcal{L}$;
    $\mathcal{G} = \mathcal{G} \cup \{L_k \to V_i | V_i \in \mathbf{C}_i\}$;
  end if
end for
Return: $\mathcal{L}$ and $\mathcal{G}$

Example 3. *Continue to consider the structure in Figure 1, we have found 8 clusters by Algorithm 2. Now, we find that* $\mathbf{C}_6$, $\mathbf{C}_7$ *and* $\mathbf{C}_8$ *are merged base on* $\mathcal{R}1$ *of Proposition 2. For any other two clusters, we can not merge them by Proposition 2. Furthermore, because there exist no latent variables introduced in the previous iterations, we do not need to verify the rules of Corollary 1. Overall, we can determine there are six latent variables, i.e.,* $L_4, ..., L_9$.

3.1.3. P3: Updating Active Variable Set

When the active variable set $\mathcal{A}$ is the observed variable set $\mathbf{X}$, one can identify some specific latent variables that are the parents of observed variables, with Algorithm 2

and Algorithm 3. However, one may have the following concerns: for a hierarchical structure, how can we further find latent variables that are the parents of latent variables and how can we check the GIN conditions over latent variables without observing them? Thanks to the linearity assumption and the transitivity of linear causal relations, we can use their observed descendants to test for the GIN conditions. For instance, consider the structure in Figure 1. Suppose $\mathbf{Y} = \{L_4, X_{14}\}$, and then $(\mathbf{X}\backslash\{X_1, X_2\}, \{L_4, X_{14}\})$ follows the GIN condition if and only if $(\mathbf{X}\backslash\{X_1, X_2\}, \{X_1, X_{14}\})$ follows it, where the measured descendant X_1 acts as a surrogate of the latent variable L_4. Thus, we can check subsequent GIN conditions over latent variables by using their proper pure observed descendants. The following proposition shows how to update the active variable set and how to test for the GIN condition over latent variables.

Proposition 3 (Active Variable Set Update)**.** *Let $\mathcal{A}$ be the current active variable set and $\mathcal{L}$ be the latent variable sets discovered in the current iteration. Then the new active variable set $\mathcal{A}' = \mathcal{A} \cup \mathcal{L}\backslash\mathbf{Ch}(\mathcal{L})$. Moreover, the GIN conditions over variables in $\mathcal{A}'$ are equivalent to those that replace $V \in \mathcal{A}'$ by any variable in its corresponding cluster identified in the latest iteration.*

The above proposition says that when testing for the GIN condition over latent variables, we can initialize the value of the latent variable with the value of any variable in its corresponding cluster found in the latest iteration, without recovering the distribution of latent variables. The complete update procedure based on Proposition 3 is summarized in Algorithm 4, and an illustrative example is given below.

Algorithm 4 UpdateActiveData

Input: New latent set $\mathcal{L}$, $\mathcal{A}$, and partial structure $\mathcal{G}$
Output: New active data $\mathcal{A}$
if $\mathcal{L} = \emptyset$ **then**
  $\mathcal{A} = \emptyset$;
else
  for each new latent variable $L_i \in \mathcal{L}$ **do**
    Initialize L_i with the value of any variable in its corresponding cluster identified in the latest iteration.
    Add L_i into $\mathcal{A}$ and remove all children of L_i from $\mathcal{A}$;
  end for
end if
Return: $\mathcal{A}$

Example 4. *We now continue identifying the structure in Figure 1. In Phase 2, we have determined that there are six latent variable sets, including $L_4, ..., L_9$. Now, we update the active variable set $\mathcal{A}$ by Algorithm 4. For each new latent variable, e.g., L_4, we initialize L_4 with the value of any variable in $\{X_1, X_2\}$, e.g., X_1. Meanwhile, we add L_4 into $\mathcal{A}$ and remove $\{X_1, X_2\}$ from $\mathcal{A}$. The similar reasoning procedure applies to other latent variables. Thus, we obtain the new active variable set $\mathcal{A} = \{X_{14}, X_{15}, L_4, ..., L_9\}$, where the values of $X_1, X_3, X_5, X_7, X_9, X_{11}$ act as surrogate of the values of latent variables $L_4, ..., L_9$, respectively.*

3.2. Step 2: Inferring Causal Structure among Latent Variables

With step 1, we can identify latent variables, as well as the causal structure among the latent parents of pure clusters (see Lemma 2). In this section, we show how to further identify the causal structure among latent variables within an impure cluster, so that the latent hierarchical causal structure is fully identifiable. The basic idea is to first identify the causal order among latent variables and then remove redundant edges.

Below, we first show how to identify the causal order between any two latent variables by appropriately testing for GIN conditions, when their latent confounder is given.

Proposition 4 (Identifying Causal Order)**.** *Let L_p and L_q be two latent variables in an impure cluster, and denote by $\{P_1, P_2\}$ and $\{Q_1, Q_2\}$ subsets of pure children of L_p and L_q, respectively. Suppose $\mathcal{L}_t$ is the set of latent confounders of L_p and L_q. Let $\mathbf{T}_1$ and $\mathbf{T}_2$ contain one of pure children of each latent variable in $\mathcal{L}_t$, and $\mathbf{T}_1 \cap \mathbf{T}_2 = \emptyset$. Then if $(\{P_2, \mathbf{T}_2\}, \{P_1, Q_1, \mathbf{T}_1\})$ follows the GIN condition, L_p is causally earlier than L_q (denoted by $L_p \succ L_q$).*

Example 5. *Consider the causal graphs in Figure 4. $\mathbf{C}_t = \{L_2, L_3, L_4\}$ is am impure cluster. Suppose $L_p = L_2$ and $L_q = L_3$. Then the set of latent confounders $\mathcal{L}_t = \{L_1\}$. Let $\mathbf{T}_1 = \{X_7\}$, $\mathbf{T}_2 = \{X_8\}$, $\{P_1, P_2\} = \{X_1, X_2\}$ and $\{Q_1, Q_2\} = \{X_3, X_4\}$. According to Proposition 4, $(\{X_2, X_8\}, \{X_1, X_3, X_7\})$ follows the GIN condition. This implies that $L_2 \succ L_3$.*

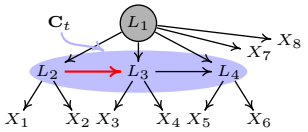


Figure 4. An illustrative example for Proposition 4 and 5

We next show how to use Proposition 4 to learn the causal order between any pair of latent variables within an impure cluster, in a recursive way. Before that, we first introduce *local root variables*, which will be used in the learning procedure.

Definition 5 (Local Root Variable)**.** *Let $\mathbf{C}_i = \{L_1, ..., L_p\}$ be an impure cluster. We say variable $L_r \in \mathbf{C}_i$ is a local root variable if there is no other latent variable in $\mathbf{C}_i$ causes it.*

For any impure cluster $\mathbf{C}_i$, due to the acyclic assumption,

the identified latent parent L_t and the identified pure children of any latent variable in $\mathbf{C}_i$ from the previous step, there always exists a local root variable L_r in $\mathbf{C}_i$ and it can be found with Proposition 4 when $\mathcal{L}_t = \{L_t\}$. After identifying the local root variable L_r in $\mathbf{C}_i$, we remove L_r from $\mathbf{C}_i$ and add it into the latent confounder set $\mathcal{L}_t$. We repeat the above procedure and recursively discover the local root variable until the casual order of the latent variables within the impure cluster is fully determined. The detailed procedure of identifying causal orders is given in Lines 2-8 of Algorithm 5.

Example 6. *Continue to consider the example in Figure 4. We have known that L_2 is a local root variable. Now, we update $\mathbf{C}_i = \{L_3, L_4\}$ and the latent confounder $\mathbf{LC} = \{L_1, L_2\}$. According to Proposition 4, we obtain that L_3 is the local root variable. Thus, we return the causal order: $L_2 \succ L_3 \succ L_4$.*

Algorithm 5 LocallyInferCausalStructure

Input: Measured variables $\mathbf{X}$ and partial structure $\mathcal{G}$
Output: Fully identified structure $\mathcal{G}$
1: **repeat**
2: Select an impure cluster $\mathbf{C}_i$ from the $\mathcal{G}$;
3: Initialize latent confounder set: $\mathbf{LC} = \emptyset$;
4: Add the common parent L_t of $\mathbf{C}_i$ into $\mathbf{LC}$;
5: **while** $|\mathbf{C}_i| > 1$ **do**
6: Find a local root variable L_r according to Proposition 4;
7: $\mathbf{C}_i = \mathbf{C}_i \backslash L_r$ and add L_r into $\mathbf{LC}$;
8: $\mathcal{G} = \mathcal{G} \cup \{L_r \to L_i | L_i \in \mathbf{C}_i\}$;
9: **end while**
10: **repeat**
11: Select an ordered pair of variables L_p and L_q in $\mathbf{C}_i$ that $L_p \succ L_q$;
12: **if** there exists set $\mathcal{L}_\mathbf{S} \subset \mathbf{C}_1$ such that each latent is causally later than L_p and is causally earlier than L_q, and L_p and L_q are d-separated by $\mathcal{L}_\mathbf{S} \cup L_t$. **then**
13: Remove the directed edge between L_p and L_q.
14: **end if**
15: **until** All ordered pairs of variables in $\mathbf{C}_i$ selected
16: **until** All impure clusters in $\mathcal{G}$ selected
17: **Return:** $\mathcal{G}$

After identifying the causal order over a set of latent variables within an impure cluster, we then remove redundant edges by using rank-deficiency test, as that in Silva et al. (2006).

Proposition 5 (Removing Redundant Edges)**.** *Let L_p and L_q be two latent variables in an impure cluster $\mathbf{C}_i$, and denote by P_1 and Q_1 pure children of L_p and L_q, respectively. Suppose L_p is causally earlier than L_q. Let L_t be the common parent of $\mathbf{C}_i$ and $\mathcal{L}_\mathbf{S}$ be the set of latent variables in $\mathbf{C}_i$ such that each latent is causally later than L_p and is causally earlier than L_q. Furthermore, let $\{T_1, T_2\}$ be two pure children of L_t, and $\mathbf{S}$ be a set of children of $\mathcal{L}_\mathbf{S}$ containing two pure children per latent. Then L_p and L_q are d-separated by $\mathcal{L}_\mathbf{S} \cup L_t$, i.e., there is no directed edge between L_p and L_q iff the rank of the correlation matrix of $\{P_1, Q_1\} \cup \{T_1, T_2\} \cup \mathbf{S}$ is less than or equal to $|L_t \cup \mathcal{L}_\mathbf{S}|$.*

Proposition 5 helps us to identify whether there is a directed edge between two latent variables by searching the d-separation set from other latent variables in sequence. The detailed procedure of removing the redundant edges is given in Lines 10-15 of Algorithm 5.

Example 7. *Continue to consider the example in Figure 4. Now, we are going to verify the directed edge between L_2 and L_4. According to Proposition 5, we obtain that the rank of the correlation matrix of $\{X_1, X_5\} \cup \{X_3, X_4, X_7, X_8\}$ is less than or equal to $|\{L_1, L_3\}| = 2$ (the d-separation set is $\{L_1, L_3\}$). This implies that L_2 and L_4 are d-separated by $\{L_1, L_3\}$, and we will remove the directed edge between L_2 and L_4.*

The complete learning procedure for identifying the causal structure among latent variables, that is based on Proposition 4 and Proposition 5, is summarized in Algorithm 5.

3.3. Complexity of LaHME Algorithm

In this section, we analyze the complexity of Steps 1 and 2 of LaHME algorithm. Denote by p the number of observed variables, by q that of latent variables, by R the maximal depth of the graph, and by S ($< q$) the maximal length of impure cluster of the graph. For Step 1, there are two dominant parts. One dominant part is to find global clusters (Algorithm 2) with worst case complexity $\mathcal{O}(Rp!)$, and the other part is to determine latent variables (Algorithm 3) with worst case complexity $\mathcal{O}(Rp^2)$. Hence, the worst case complexity is $\mathcal{O}(Rp!)$. Step 2 includes Algorithm 5 whose worst case complexity is $\mathcal{O}(TS^2)$, where T ($< q$) is the maximum number of impure clusters of the graph.

4. Identifiability of Latent Hierarchical Structure

In this section, we show that the LaHME algorithm can identify the correct causal structure asymptotically, if the data satisfies LiNGLaH and the graph structure satisfies the minimal latent hierarchical structure (Condition 1). Below, we first show that the latent variables, as well as the causal structure among the latent parents of pure clusters, are identifiable by Step 1 of the LaHME algorithm, which is stated in the following lemma.

Lemma 2. *Suppose that the input data $\mathbf{X}$ follow LiNGLaH with the minimal latent hierarchical structure. Then the*

Table 1. Performance of LaHME, GIN, FOFC, BPC, CLRG and CLNJ on learning latent hierarchical structure.

		Structure Recovery Error Rate ↓						Error in Hidden Variables ↓						Correct-Ordering Rate ↑					
Algorithm		LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ
Case 1	1k	0.1	0.2	1.0	1.0	1.0	1.0	0.1	0.1	0.5	0.6	2.0	2.0	0.96	0.92	-	-	-	-
	5k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.1	2.0	2.0	1.0	1.0	-	-	-	-
	10k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	2.0	2.0	1.0	1.0	-	-	-	-
Case 2	1k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	3.8	3.9	4.0	4.0	0.9	0.08	-	-	-	-
	5k	0.1	1.0	1.0	1.0	1.0	1.0	0.1	3.0	3.6	3.8	4.0	4.0	0.96	0.1	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.0	3.5	3.8	4.0	4.0	1.0	0.1	-	-	-	-
Case 3	1k	0.1	1.0	1.0	1.0	1.0	1.0	0.2	1.3	3.0	3.1	3.0	3.0	0.92	0.0	-	-	-	-
	5k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.2	3.0	3.2	3.0	3.0	1.0	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	3.2	3.4	3.0	3.0	1.0	0.0	-	-	-	-
Case 4	1k	0.3	1.0	1.0	1.0	1.0	1.0	0.4	3.4	7.0	7.2	8.0	8.0	0.9	0.0	-	-	-	-
	5k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	6.6	6.9	8.0	8.0	0.94	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.1	5.8	6.7	8.0	8.0	1.0	0.0	-	-	-	-

Note: The symbol '-' indicates that the current method does not output this information. ↓ means a lower value is better, and vice versa.

underlying latent variables, as well as the causal structure among the latent parents of pure clusters, are identifiable by Step 1 of the LaHME algorithm.

Lemma 2 implies that if the underlying causal structure is a tree-based graph (each latent variable only has pure children), then the underlying graph will be recovered only with Step 1 of the LaHME algorithm.

We next show that the whole hierarchical structure is identifiable with the LaHME algorithm, as stated in Theorem 1. An illustrative example of the entire procedure of LaHME is given in Appendix D.

Theorem 1 (Identifiability of Latent Hierarchical Structure)**.** *Suppose that the input data* $\mathbf{X}$ *follows LiNGLaH with the minimal latent hierarchical structure. Then the underlying causal graph* $\mathcal{G}$ *is fully identifiable with LaHME, including latent variables and their causal relationships.*

5. Experiments

In this section, we first apply the proposed method to synthetic data to demonstrate the correctness. Then, we apply our algorithm to real-world data set to show its usefulness.

5.1. Synthetic Data

In the following simulation studies, we generated data according to four typical structures that satisfy minimal latent hierarchical structure, including tree-based and measurement-based structures (see Figure 5). We considered different sample sizes $N = 1k, 5k, 10k$. The causal strength b_{ij} and c_{jk} were generated uniformly in $[-2, -0.5] \cup [0.5, 2]$ and the non-Gaussian noise terms were generated from exponential distributions to the second power. We used HSIC-based independence tests (Zhang et al., 2018) to test for the GIN condition, due to the non-Gaussianity of the data.

We compared the proposed LaHME algorithm with measurement-based methods, such as BPC (Silva et al., 2006), FOFC (Kummerfeld & Ramsey, 2016), and GIN-based approach (Xie et al., 2020). We also compared LaHME with tree-based methods, such as Chow-Liu Recursive Grouping (CLRG) and Chow-Liu Neighbor Joining (CLNJ) (Choi et al., 2011). Each experiment was repeated ten times with randomly generated data, and the results were averaged. The source code is in the Supplementary file.

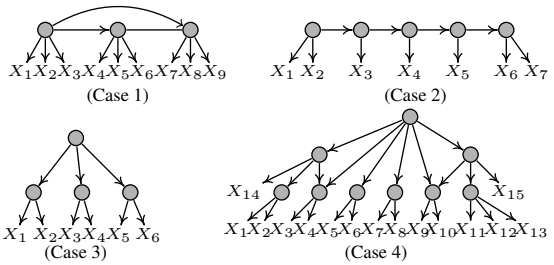


Figure 5. Latent structures used in our simulation studies.

We adapted the evaluation metrics from Choi et al. (2011) and Xie et al. (2020) to evaluate the minimal latent hierarchical structure. Specifically, we used the following three metrics:

- *Structure Recovery Error rate*: the percentage that the proposed algorithm fails to recover the ground-truth structure. Note that this is a strict measure because even a wrong latent variable or a wrong direction results in an error.
- *Error in the Number of Latent variable sets*: the absolute difference between the averaged number of latent variables estimated and the number of latent variables in the ground-truth structure.
- *Correct ordering rate*: the number of correctly inferred causal ordering divided by the total number of causal ordering in the true structure.

The experimental results were reported in Table 1. From the table, we can see that our proposed LaHME outperforms other methods with all the three evaluation metrics, in all the structures, and in all sample sizes, indicating that it can not only handle the tree-based and measurement-based structures, but also the latent hierarchical structure, including the causal directions. We found that the GIN-based algorithm does not perform well especially on hierarchical structures, so it is not appropriate to identify structures where latent variables have no observed variables as children. We further noticed that CLRG and CLNJ algorithms do not perform well on case 3, although the structure is a tree. One possible

Table 2. Results with two scenarios that violates the assumptions of our model (with sample size=10k).

	Structure Recovery Error Rate ↓				**Error in Hidden Variables ↓**				**Correct-Ordering Rate ↑**			
Cases	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
Scenario 1	0.00	0.20	0.00	0.30	0.00	0.20	0.00	0.40	1.00	0.90	1.00	0.91
Scenario 2	1.00				2.00				0.47			

reason is that these algorithms were designed for Gaussian and discrete variables only.

In addition, we consider the following two typical scenarios where the assumptions are not fulfilled. Scenario 1) *Violation of non-Gaussianity:* we consider all four graphs in Fig. 5 with Gaussian noise terms for measured variables. Scenarios 2) *Violation of Cond. 1:* we consider the graph in Fig. 2(b), where L_6 and L_7 only have two neighbors. Experimental results are reported in Table 2. We found that in Scenario 1, the performances in cases 1&3 are almost the same, and the performance drops slightly in cases 2&4 while it is still better than by chance. In Scenario 2, the recovered graph of our method is the graph in Fig. 2(a), which indicates that Condition 1 is necessary.

In summary, these above findings show a clear advantage of our method over the comparisons. We also consider graphs with different scales (numbers of variables and depths) and report the results in Appendix E.

5.2. Real-World Data

We applied our LaHME algorithm to a multitasking behavior model, represented by a hierarchical SEM (Himi et al., 2019). In details, the multitasking behavior model contains four latent factors: *Multitasking behavior (Mb)*, *Speed (S)*, *Error (E)*, and *Question (Q)*, where factor *Mb* has no observed variables as children, and *Speed (S)*, *Error (E)*, and *Question (Q)* each has three measured variables. The detailed explanation of the data set is given in Appendix F. The data set consists of 202 samples.

Figure 6 shows the performance of different algorithms. The significance levels of LaHME, BPC, and FOFC were set to 0.001, 0.0001, and 0.000001 respectively. The reason of choosing different significance levels is that BPC and FOFC algorithms will output empty graph if the significance level is 0.001. Here, we chose the 'better' significance such that the output graphs of BPC and FOFC algorithms are closer to the ground-truth graph. The result of our output is consistent with the model given in Himi et al. (2019), which indicates the effectiveness of our method. Note that although GIN and BPC discover three other latent variables that have observed children, neither finds latent factor *Multitasking behavior*. CLRG and CLNJ output the same result and capture the latent variable *Multitasking behavior*; however, They fail to find the latent variable *Speed*. These results again indicate that our algorithm has better performance than other algorithms in learning hierarchical structure.

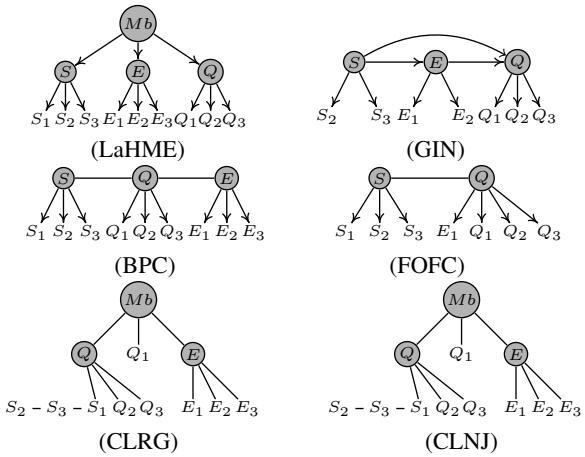


Figure 6. The output of LaHME, GIN,BPC,FOFC,CLRG and CLNJ on the multitasking behavior data. The ground-truth graph is the same as that identified by LaHME.

6. Conclusions and Further work

We investigated the problem of learning linear non-Gaussian latent hierarchical models from observational data. In particular, we allow latent variables without observed variables as children and hierarchical structures beyond a tree for the latent variables. We provided sufficient conditions for structural identifiability, and proposed a recursive clustering method, which locates the latent variables and locally infers the causal order among them. The method can output the correct causal structure asymptotically.

Currently, we assumed that each latent variable has a certain number of pure children, i.e., *1-factor model*. A future research line is to extend it to *n-factor models*, where the condition is on the number of pure children of a latent variable set. It can be achieved by extending LaHME that considers an increased size of $\mathbf{Y}$ in testing for GIN conditions. Another line of future research is to estimate the latent hierarchical structure under nonlinear causal models (e.g., the post-nonlinear model (Zhang & Hyvärinen, 2009)).

Acknowledgements

We would like to thank Clark Glymour for insightful discussions. FX, YH, and ZG would like to acknowledge the supported by the China Postdoctoral Science Foundation (020M680225) and the National Natural Science Foundation of China (NSFC 11771028, 12071015, 11971040). KZ acknowledges the support by the NIH under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award 2134901, and by a grant from Apple Inc.

References

Adams, J., Hansen, N., and Zhang, K. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34, 2021.

Bartholomew, D., Steele, F., Moustaki, I., and Galbraith, J. *The analysis and interpretation of multivariate data for social scientists*. Routledge (2 edition), 2008.

Bollen, K. A. *Structural Equations with Latent Variable*. John Wiley & Sons, 1989.

Cai, R., Xie, F., Glymour, C., Hao, Z., and Zhang, K. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12863–12872, 2019.

Chen, W., Zhang, K., Cai, R., Huang, B., Ramsey, J., Hao, Z., and Glymour, C. Fritl: A hybrid method for causal discovery in the presence of latent confounders. *arXiv preprint arXiv:2103.14238*, 2021.

Chen, Z. and Chan, L. Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders. *Neural Computation*, 25(6):1605–1641, 2013.

Chen, Z., Xie, F., Qiao, J., Hao, Z., Zhang, K., and Cai, R. Identification of linear latent variable model with arbitrary distribution. In *Proceedings 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Choi, M. J., Tan, V. Y., Anandkumar, A., and Willsky, A. S. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

Claassen, T., Mooij, J. M., and Heskes, T. Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 172. Citeseer, 2013.

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

Cui, R., Groot, P., Schauer, M., and Heskes, T. Learning the causal structure of copula models with latent variables. In *Proceedings of the 34rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 188–197. AUAI Press, 2018.

Drton, M., Lin, S., Weihs, L., and Zwiernik, P. Marginal likelihood and model selection for gaussian latent tree and forest models. *Bernoulli*, 23(2):1202–1232, 2017.

Entner, D. and Hoyer, P. O. Discovering unconfounded causal relationships using linear non-gaussian models. In *JSAI International Symposium on Artificial Intelligence*, pp. 181–195. Springer, 2010.

Etesami, J., Kiyavash, N., and Coleman, T. Learning minimal latent directed information polytrees. *Neural computation*, 28(9):1723–1768, 2016.

Harmeling, S. and Williams, C. K. Greedy learning of binary latent trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1087–1097, 2010.

Himi, S. A., Bühner, M., Schwaighofer, M., Klapetek, A., and Hilbert, S. Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition*, 189:275–298, 2019.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1655–1664. ACM, 2016.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Maeda, T. N. and Shimizu, S. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 735–745, 2020.

Mourad, R., Sinoquet, C., Zhang, N. L., Liu, T., and Leray, P. A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47(1):157–203, 2013.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

Poon, L. K., Zhang, N. L., Chen, T., and Wang, Y. Variable selection in model-based clustering: to do or to facilitate. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, pp. 887–894, 2010.

Salehkaleybar, S., Ghassami, A., Kiyavash, N., and Zhang, K. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

Shimizu, S. Non-gaussian methods for causal structure learning. *Prevention Science*, 20(3):431–441, 2019.

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

Shimizu, S., Hoyer, P. O., and Hyvärinen, A. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Silva, R., Scheine, R., Glymour, C., and Spirtes, P. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

Spearman, C. Pearson's contribution to the theory of two factors. *British Journal of Psychology. General Section*, 19(1):95–101, 1928.

Spirtes, P. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 606–615, 2013.

Spirtes, P. and Zhang, K. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp. 1–28. SpringerOpen, 2016.

Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI)*, pp. 499–506, 1995.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2000.

Sullivant, S., Talaska, K., Draisma, J., et al. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. ParceLiNGAM: a causal ordering method robust against latent confounders. *Neural Computation*, 26(1):57–83, 2014.

Wang, Y. S. and Drton, M. Causal discovery with unobserved confounding and non-gaussian data. *arXiv preprint arXiv:2007.11131*, 2020.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise conditionfor estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14891–14902, 2020.

Zeng, Y., Shimizu, S., Cai, R., Xie, F., Yamamoto, M., and Hao, Z. Causal discovery with multi-domain lingam for latent factors. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2097–2103, 2021.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 647–655. AUAI Press, 2009.

Zhang, N. L. Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, 5: 697–723, 2004.

Zhang, N. L. and Poon, L. K. Latent tree analysis. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4891–4897, 2017.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

A. Notations

Symbol	Description
$\mathcal{G}$	A directed acyclic graph
$\mathbf{V}$	The set of all variables
$\mathbf{X}$	The set of observed variables
$\mathbf{L}$	The set of latent variables
$\mathbf{Pa}(V_i)$	The set of all parents of V_i
$\mathbf{Ch}(V_i)$	The set of all children of V_i
$\mathbf{Pa}(\mathbf{Y})$	The set of common parents of a variable set $\mathbf{Y}$
$L(\mathbf{Y})$	The set of latent variables that are parents of any component of $\mathbf{Y}$.
$\lvert * \rvert$ (e.g., $\lvert\mathbf{Y}\rvert$)	The number of elements of set $*$ ($\mathbf{Y}$)
$\mathbf{\Sigma_{AB}}$	The cross-covariance matrix of set $\mathbf{A}$ and $\mathbf{B}$
$\mathrm{rank}(\mathbf{\Sigma_{AB}})$	The rank of cross-covariance matrix of set $\mathbf{A}$ and $\mathbf{B}$

Table 3. The list of main symbols used in this Appendix

B. Proofs

Before presenting the proofs of our results, we need a few more theorems and definitions.

Definition 6 (GIN condition (Xie et al., 2020))**.** *Let $\mathbf{Z}$, $\mathbf{Y}$ be sets of variables in a linear non-Gaussian acyclic causal model. We say that $(\mathbf{Z}, \mathbf{Y})$ follows GIN condition if and only if $\omega^\intercal \mathbf{Y}$ are statistically independent of $\mathbf{Z}$, where ω satisfies $\omega^\intercal \mathbb{E}[\mathbf{Y}\mathbf{Z}^\intercal] = 0$ and $\omega \neq 0$.*

In other words, $(\mathbf{Z}, \mathbf{Y})$ violates the GIN condition if and only if $E_{\mathbf{Y}||\mathbf{Z}}$ is dependent on $\mathbf{Z}$.

Below, we show the graphical implication of the GIN condition in LiNGLaH, which helps to exploit the GIN condition to discover the latent hierarchical structure, and we first give the definition of *exogenous set*, which will be used in the theorem.

Definition 7 (Exogenous set)**.** *We say variable set $\mathcal{S}_1$ is an exogenous set relative to variable set $\mathcal{S}_2$ if and only if 1) $\mathcal{S}_2 \subseteq \mathcal{S}_1$ or 2) for any variable V that is in $\mathcal{S}_2$ but not in $\mathcal{S}_1$, according to the causal graph over $\{V\} \cup \mathcal{S}_1$ and the ancestors of variables in $\{V\} \cup \mathcal{S}_1$, V does not cause any variable in $\mathcal{S}_1$, and the common cause for V and each variable in $\mathcal{S}_1$, if there is any, is also in $\mathcal{S}_1$ (i.e., relative to $\{V\} \cup \mathcal{S}_1$, V does not cause and is not confounded with any variable in $\mathcal{S}_1$).*

Theorem 2 (GIN Graphical Criteria in LiNGLaH)**.** *Let $\mathbf{Y}$ and $\mathbf{Z}$ be two sets of observed variables of a linear non-Gaussian latent hierarchical model (LiNGLaH). $(\mathbf{Z}, \mathbf{Y})$ satisfies the GIN condition (while with the same $\mathbf{Z}$, no proper subset of $\mathbf{Y}$ does) if and only if there exists a k-size subset of the latent variables $\mathbf{L}$, $0 \leq k \leq min(Dim(\mathbf{Y}) - 1, Dim(\mathbf{Z}))$, denoted by $\mathcal{S}_L^k$, such that 1) $\mathcal{S}_L^k$ is an exogenous set relative to $L(\mathbf{Y})$, that 2) $\mathcal{S}_L^k$ d-separates $\mathbf{Y}$ from $\mathbf{Z}$, and that 3) the covariance matrix of $\mathcal{S}_L^k$ and $\mathbf{Z}$ has rank k, and so does that of $\mathcal{S}_L^k$ and $\mathbf{Y}$.*

Roughly speaking, the conditions in this theorem can be interpreted in the following way: i.) a causally earlier subset (according to the causal order) of the common causes of $\mathbf{Y}$ d-separate $\mathbf{Y}$ from $\mathbf{Z}$, and ii.) the linear transformation from that subset of the common causes to $\mathbf{Z}$ has full column rank.

Proof. It has been shown in Xie et al. (2020) that the graphical criteria hold in linear non-Gaussian latent variable model (LiNGLaM). The key difference between LiNGLaM and LiNGLaH is that LiNGLaH allows some latent variables have no observed variables, which does not affect the graphical criteria of GIN in terms of LiNGLaH because linear causal models are transitive. Therefore, the graphical criteria also hold in LiNGLaH. □

B.1. Proof of Proposition 1

Proof. (i) Assume that $\mathbf{Y}$ is a global causal cluster. Let $\tilde{\mathbf{Y}}$ be any subset of $\mathbf{Y}$ such that $|\tilde{\mathbf{Y}}| = 2$. To prove that $(\mathcal{A} \setminus \mathbf{Y}, \tilde{\mathbf{Y}})$ follows the GIN condition, we need to verify the three conditions of GIN Graphical Criteria in Theorem 2. First, due to the

definition of global causal clusters, we know that there exists a latent parent, denoted by L_1, d-separates Y from $\mathcal{A} \setminus \mathbf{Y}$. This will imply that conditions 1 and 2 of GIN Graphical Criteria hold, i.e., L_1 is an exogenous set relative to $L(\tilde{\mathbf{Y}}) = L_1$, and L_1 d-separates $\mathbf{Y}$ from $\mathcal{A} \setminus \mathbf{Y}$. Furthermore, denote by ε the set of common components between $\mathbf{Y}$ and $\mathcal{A} \setminus \mathbf{Y}$. Because L_1 is the only parent of $\mathbf{Y}$, we have $|\varepsilon| = 1$. Therefore, the covariance matrix of L_1 and $\mathcal{A} \setminus \mathbf{Y}$ has rank 1, and so does that of L_1 and $\mathbf{Y}$. Due to the condition 2 of global causal cluster, we obtain that no proper subset $\tilde{\mathbf{Y}} \subset \mathbf{Y}$ such that $\tilde{\mathbf{Y}}$ and $\mathcal{A} \setminus \tilde{\mathbf{Y}}$ are d-separated by L_1. This will imply that there is no proper subset of $\mathbf{Y}$ satisfies the condition 1.

(ii) Assume that $\mathbf{Y}$ is not a global causal cluster. We need to consider the following two cases:

Case 1: $\mathbf{Y}$ is not a causal cluster. Since $\mathbf{Y}$ is not a causal cluster, without loss of generality, $L(\mathbf{Y})$ must contain at least two different latent parents, denoted by L_1 and L_2. Furthermore, because condition 1 holds, i.e., for any subset $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$ with $|\tilde{\mathbf{Y}}| = 2$, $(\mathcal{A} \setminus \mathbf{Y}, \tilde{\mathbf{Y}})$ follows the GIN condition. According to GIN Graphical Criteria, $\{L_1, L_2\}$ d-separates $\mathbf{Y}$ and $\mathcal{A} \setminus \mathbf{Y}$, which leads to the contradiction–the variables in $\mathbf{Y}$ share only one common parent.

Case 2: $\mathbf{Y}$ is a causal cluster but is not global. Let L_1 be the common parent of $\mathbf{Y}$. Since $\mathbf{Y}$ is not global, there are two sub-cases to be discussed (see Figure 7),

a. $\mathbf{Y}$ and $\mathcal{A} \setminus \mathbf{Y}$ are not d-separated by L_1;

b. $\mathbf{Y}$ and $\mathcal{A} \setminus \mathbf{Y}$ are d-separated by L_1, but there exist a proper subset $\tilde{\mathbf{Y}} \subset \mathbf{Y}$ such that $\tilde{\mathbf{Y}}$ and $\mathcal{A} \setminus \tilde{\mathbf{Y}}$ are d-separated by L_1.

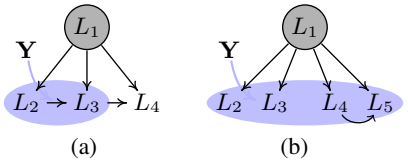


Figure 7. The illustrative examples for Case 2.

Case 2(a): Because condition 1 holds, i.e., for any subset $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$ with $|\tilde{\mathbf{Y}}| = 2$, $(\mathcal{A} \setminus \mathbf{Y}, \tilde{\mathbf{Y}})$ follows the GIN condition. According to GIN Graphical Criteria, there exist $\mathcal{S}_L^1$ such that $\mathcal{S}_L^1$ d-separates $\mathbf{Y}$ from $\mathcal{A} \setminus \mathbf{Y}$. Furthermore, since L_1 is the common parent of $\mathbf{Y}$, $\mathcal{S}_L^1 = L_1$. This will imply that L_1 d-separates $\mathbf{Y}$ from $\mathcal{A} \setminus \mathbf{Y}$, which leads to the contradiction.

Case 2(b): Because there exists a proper subset $\tilde{\mathbf{Y}} \subset \mathbf{Y}$ such that $\tilde{\mathbf{Y}}$ and $\mathcal{A} \setminus \tilde{\mathbf{Y}}$ are d-separated by L_1. Without loss of generality, denote by $\mathbf{Y}' \subset \mathbf{Y}$ the subset, i.e., $\mathbf{Y}'$ and $\mathcal{A} \setminus \mathbf{Y}'$ are d-separated by L_1. According to condition 1 and GIN Graphical Criteria, we have that for any subset $\tilde{\mathbf{Y}}'$ of $\mathbf{Y}'$ with $|\tilde{\mathbf{Y}}'| = 2$, $(\mathcal{A} \setminus \mathbf{Y}', \tilde{\mathbf{Y}}')$ follows the GIN condition, which leads to the contradiction–condition 2 does not hold.

Therefore, from (i) and (ii), the proposition is proved.

□

B.2. Proof of Lemma 1

Proof. The proof of Statement 1 is obvious: if $\mathbf{C}_1$ is a pure cluster, then there exist a proper subset $\mathbf{Y}' \subset \mathbf{Y}$ such that for any subset $\tilde{\mathbf{Y}}'$ of $\mathbf{Y}'$ with $|\tilde{\mathbf{Y}}'| = 2$, $(\mathcal{A} \setminus \mathbf{Y}', \tilde{\mathbf{Y}}')$ follows the GIN condition, which leads to the contradiction–$\mathbf{C}_1$ is a global causal cluster.

We prove Statement 2 by contradiction. Assume that $\mathbf{C}_1 = \{V_i, V_j\}$ is an impure cluster. Without loss of generality, assume that $V_i \to V_j$. Let L_1 be the common parent of $\mathbf{C}_1$. With the model assumption, L_1 has at least two pure children, denote by $V_k, V_t \in \{\mathcal{A} \cup \mathbf{Ch}(\mathcal{A}) \setminus \mathbf{C}_1\}$. Next, we will show that $(\{V_i, V_t\}, \{V_i, V_j, V_k\})$ follows the GIN condition while $(\{V_j, V_t\}, \{V_i, V_j, V_k\})$ violates the GIN condition. To do so, we need to verify the three conditions of GIN Graphical Criteria Theorem.

We first verify that $(\{V_i, V_t\}, \{V_i, V_j, V_k\})$ follows the GIN condition. Denote by ε the set of common components between $\{V_i, V_j, V_k\}$ and $\{V_i, V_t\}$. Because $\{V_i, V_j, V_k, V_t\}$ is a causal cluster and $\{V_k, V_t\}$ is a pure cluster, we know that $\varepsilon = \{\varepsilon_{L_1}, \varepsilon_{V_i}\}$. Thus, we know there exist a set $\mathcal{S}_L^2 = L_1, V_i$ such that (1) $\{L_1, V_i\}$ is an exogenous set relative to

$L(\{V_i, V_j, V_k\}) = \{L_1, V_i\}$, (2) $\{L_1, V_i\}$ d-separates $\{V_i, V_j, V_k\}$ from $\{V_i, V_t\}$, and (3) the covariance matrix of L_1 and $\mathcal{A} \setminus \mathbf{Y}$ has rank 2, and so does that of L_1 and $\mathbf{Y}$. These will imply that $(\{V_i, V_t\}, \{V_i, V_j, V_k\})$ follows the GIN condition.

We next verify that $(\{V_j, V_t\}, \{V_i, V_j, V_k\})$ violates the GIN condition. Denote by ε' the set of common components between $\{V_i, V_j, V_k\}$ and $\{V_j, V_t\}$. Because $\{V_i, V_j, V_k, V_t\}$ is a causal cluster and $\{V_k, V_t\}$ is a pure cluster, we know that $\varepsilon' = \{\varepsilon_{L_1}, \varepsilon_{V_i}, \varepsilon_{V_j}\}$. Thus, there does not exist a set $\mathcal{S}_L^k$ such that $k \leq \min(|\{V_i, V_j, V_k\}| - 1, |\{V_j, V_t\}|) = 2$. This will imply that $(\{V_j, V_t\}, \{V_i, V_j, V_k\})$ violates the GIN condition. □

B.3. Proof of Proposition 2

Proof. We will prove these three rules by contradiction, by assuming that $\mathbf{C}_1$ and $\mathbf{C}_2$ do not share the same latent parent and showing that condition (2) of each rule violates. Let L_1 and L_2 be the parents of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively.

We first prove $\mathcal{R}1$. We consider the case where $\mathbf{C}_1$ and $\mathbf{C}_2$ are two pure clusters. With the model assumption, without loss of generality, we assume $\mathbf{C}_1 = \{V_1, V_2\}$ and $\mathbf{C}_2 = \{V_3, V_4\}$, as illustrate by Figure 8. We can show that condition (2) of $\mathcal{R}1$ is violated.

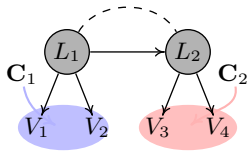


Figure 8. The illustrative example that violates $\mathcal{R}_1$.

Here, we let $\tilde{\mathbf{C}} = \{V_1, V_3\}$ be the subset of $\mathbf{C}_1 \cup \mathbf{C}_2$. Denote by ε the set of common components between $\{V_1, V_3\}$ and $\{V_2, V_4, ...\}$. Because $\mathbf{C}_1$ and $\mathbf{C}_2$ are two clusters, ε must contain $\varepsilon_{L_1}, \varepsilon_{L_2}$. That is to say, $|\varepsilon| \geq 2$. According to GIN Graphical Criteria, there does not exist a set $\mathcal{S}_L^k$ such that $k \leq \min(|\{V_1, V_3\}| - 1, |\{V_2, V_4, ...\}|) = 1$. This will imply that $(\{V_2, V_4, ...\}, \{V_1, V_3\})$ violates the GIN condition.

We next prove $\mathcal{R}2$. We need to consider the case where $\mathbf{C}_1$ is pure and $\mathbf{C}_2$ is impure. With the model assumption, without loss of generality, we assume $\mathbf{C}_1 = \{V_1, V_2\}$ and $\mathbf{C}_2 = \{V_3, V_4, V_5, ...\}$, as illustrate by Figure 9. We can show condition (2) of $\mathcal{R}2$ is violated.

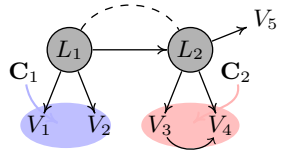


Figure 9. The illustrative example that violates $\mathcal{R}_2$.

Here, we let $V_i = V_1 \in \mathbf{C}_1$ and $V_j = V_3 \in \mathbf{C}_2$. Denote by ε the set of common components between $\{V_1, V_3\}$ and $\{V_2, V_5...\}$. Because $\mathbf{C}_1$ and $\mathbf{C}_2$ are two clusters, we know that ε must contain noise terms ε_{L_1} and ε_{L_2}. That is to say, $|\varepsilon| \geq 2$. According to GIN Graphical Criteria, there does not exist a set $\mathcal{S}_L^k$ such that $k \leq \min(|\{V_1, V_3\}| - 1, |\{V_2, V_4, ...\}|) = 1$. This will imply that $(\{V_2, V_5, ...\}, \{V_1, V_3\})$ violates the GIN condition.

Finally, we prove $\mathcal{R}3$. This proof is similar to Case 2. We need to consider the case where $\mathbf{C}_1$ and $\mathbf{C}_2$ both are impure. With the model assumption, without loss of generality, we assume $\mathbf{C}_1 = \{V_1, V_2, V_5, ...\}$ and $\mathbf{C}_2 = \{V_3, V_4, V_6, ...\}$, as illustrate by Figure 10. We can show condition (2) of $\mathcal{R}3$ is violated.

Here, we let $V_i = V_1 \in \mathbf{C}_1$ and $V_j = V_3 \in \mathbf{C}_2$. Denote by ε the set of common components between $\{V_1, V_3\}$ and $\{V_5, V_6...\}$. Because $\mathbf{C}_1$ and $\mathbf{C}_2$ are two clusters, we know that ε must contain noise terms ε_{L_1} and ε_{L_2}. That is to say, $|\varepsilon| \geq 2$. According to GIN Graphical Criteria, there does not exist a set $\mathcal{S}_L^k$ such that $k \leq \min(|\{V_1, V_3\}| - 1, |\{V_5, V_6, ...\}|) = 1$. This will imply that $(\{V_5, V_6, ...\}, \{V_1, V_3\})$ violates the GIN condition.

□

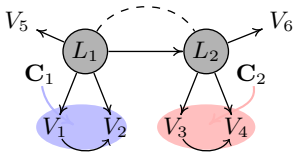


Figure 10. The illustrative example that violates $\mathcal{R}_3$.

B.4. Proof of Corollary 1

Proof. It suffices to notice that all elements in $\mathbf{C}_1$ are the children of $\mathcal{L}_1$ and do not affect the variables in $\mathcal{A}$. The corollary follow immediately from the three rules of Proposition 2 when we update $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1$. □

B.5. Proof of Proposition 3

Proof. Suppose $\mathcal{G}$ is the current graph and $\mathcal{G}'$ is the updated graph. According to Proposition 1 and 2, all elements in $\mathcal{L}$ must be the parent nodes of some nodes in $\mathcal{A}$. That is to say, all nodes in $\mathcal{L}$ must be the leaves when we remove $\mathbf{Ch}(\mathcal{L})$. Thus, the structure of the other variables in $\mathcal{A}'$ is not changed. Furthermore, because linear causal models are transitive, each latent variable L'_i in $\mathcal{A}'$ still have at least two pure children when the values of $\mathcal{L}$ are updated to their corresponding children that identified in the latest iteration. Therefore, each latent variable L'_i in $\mathcal{A}'$ satisfies the minimal latent hierarchical structure. That is to say, the GIN conditions over variables in $\mathcal{A}'$ are equivalent to those that replace $V \in \mathcal{A}'$ by any variable in its corresponding cluster identified in the latest iteration. □

B.6. Proof of Proposition 4

Proof. Without loss of generality, we assume that the ground-truth causal order is $L_p \succ L_q$. Further, we assume that $\mathcal{L}_t = \{L_1, ..., L_t\}$, $\mathbf{T}_1 = \{T_1, ..., T_t\}$ and $\mathbf{T}_2 = \{T'_1, ..., T'_t\}$. Figure 11 illustrates the case. We will prove this Proposition by leveraging the GIN Graphical Criteria. That is to say, we need to verify these conditions of GIN Graphical Criteria theorem for $\{P_1, Q_1, \mathbf{T}_1\}$ and $\{P_2, \mathbf{T}_2\}$.

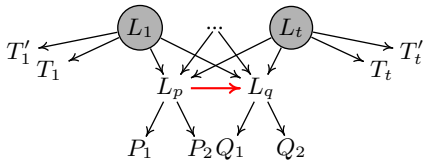


Figure 11. An illustrative example for Proposition 4.

We first verify that $(\{P_2, \mathbf{T}_2\}, \{P_1, Q_1, \mathbf{T}_1\})$ follows the GIN condition. Denote by ε the set of common components between $\{P_1, Q_1, \mathbf{T}_1\}$ and $\{P_2, \mathbf{T}_2\}$. Because $\{L_1, ..., L_t\}$ is the set of latent confounders between $\{P_1, Q_1, \mathbf{T}_1\}$ and $\{P_2, \mathbf{T}_2\}$, and $L_p \succ L_q$, we know that $\varepsilon = \{\varepsilon_{L_1}, ..., \varepsilon_{L_t}, \varepsilon_{L_p}\}$. Thus, we know there exist a set $\mathcal{S}_L^{t+1} = \{L_1, ..., L_t, L_p\}$ such that (1) $\{L_1, ..., L_t, L_p\}$ is an exogenous set relative to $L(\{P_1, Q_1, \mathbf{T}_1\}) = \{L_1, ..., L_t, L_p\}$, (2) $\{L_1, ..., L_t, L_p\}$ d-separates $\{P_1, Q_1, \mathbf{T}_1\}$ from $\{P_2, \mathbf{T}_2\}$, and (3) the covariance matrix of $\{L_1, ..., L_t, L_p\}$ and $\{P_2, \mathbf{T}_2\}$ has rank $t+1$, and so does that of $\{L_1, ..., L_t, L_p\}$ and $\{P_1, Q_1, \mathbf{T}_1\}$. These implies that $(\{P_2, \mathbf{T}_2\}, \{P_1, Q_1, \mathbf{T}_1\})$ follows the GIN condition.

We now verify that $(\{Q_2, \mathbf{T}_2\}, \{P_1, Q_1, \mathbf{T}_1\})$ violates the GIN condition. Denote by ε' the set of common components between $\{P_1, Q_1, \mathbf{T}_1\}$ and $\{Q_2, \mathbf{T}_2\}$. Because $\{L_1, ..., L_t\}$ is the set of latent confounders between $\{P_1, Q_1, \mathbf{T}_1\}$ and $\{P_2, \mathbf{T}_2\}$, and $L_p \succ L_q$, we know that $\varepsilon = \{\varepsilon_{L_1}, ..., \varepsilon_{L_t}, \varepsilon_{L_p}, \varepsilon_{L_q}\}$. Thus, we know there does not exist a set $\mathcal{S}_L^k$ such that $k \leq \min(|\{P_1, Q_1, \mathbf{T}_1\}| - 1, |\{Q_2, \mathbf{T}_2\}|) = t+1$. This will imply that $(\{V_j, V_t\}, \{V_i, V_j, V_k\})$ violates the GIN condition.

Based on the above analyses, we have $L_p \succ L_q$ if $(\{P_2, \mathbf{T}_2\}, \{P_1, Q_1, \mathbf{T}_1\})$ follows the GIN condition. □

B.7. Proof of Proposition 5

Proof. This result can be proved by following Theorem 19 in Silva et al. (2006). The key difference to Theorem 19 is that we have known the causal order of latent variables and the d-separation set of L_p and L_q can be selected in sequence. □

B.8. Proof of Lemma 2

Proof. We first show that all latent variables can be located in Step 1 of LaHME. Specifically, In the first iteration, with Proposition 1, one can detect all global causal clusters by testing for GIN conditions. Then, by Propositions 2 and Corollary 1, one can detect all latent variables. Next, by Propositions 3, the new active variable set is structurally consistent with the ground-truth one, which will ensure that the recursive search process is correct and all latent variables can be located.

Now, we show the causal structure among the latent parents of pure clusters. This result follows immediately from the definition of causal cluster—the nodes in any cluster have the common parent and there is no directed edges between them if this cluster is a pure cluster. □

B.9. Proof of Theorem 1

Proof. Based on Lemma 2, all latent variables as well as the causal structure among the latent parents of pure clusters can be identified in Step 1 of LaHME.

We now show the casual structure among latent variables within any one impure cluster can be identified in Step 2 of LaHME. For an impure cluster, $\mathbf{C}_i$, a local root set in $\mathbf{C}_i$ can be found exactly with the condition of Proposition 4. Due to the acyclic assumption, we know the order of recursive search of LocallyInferCausalStructure is the causal order of the original variables. Finally, given the causal order of the latent variables, we can use the rank-based independence tests to removing the redundant edges from the fully connected sub-graph (Proposition 5).

Based on the above analysis, one can identify all latent variables of the system and infer the causal structure among them (including the causal direction). This implies that the latent hierarchical structure of LiNGLaH is fully identifiable. □

C. Illustration of Merging Rules

Here, we give an example to illustrate the three rules in Proposition 2 and Corollary 1 to identify the clusters of variables that share a common latent parent.

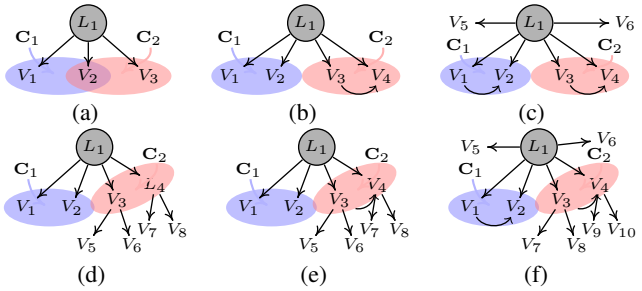


Figure 12. The illustrative examples for $\mathcal{R}1 \sim 3$ in Proposition 2 and Corollary 1.

Consider the causal graphs in Figure 12. We first analyze the subgraphs (a) $\sim$ (c) by Proposition 2, where clusters $\mathbf{C}_1$ and $\mathbf{C}_2$ are found in the same iteration.

$\mathcal{R}1$. In subgraph (a), $\mathbf{C}_1$ and $\mathbf{C}_2$ are two pure and overlapping causal clusters. Let $\mathcal{A} = \{V_1, V_2, V_3\}$. For any subset of $\mathbf{C}_1 \cup \mathbf{C}_2$, e.g., $\tilde{\mathbf{C}}_1 = \{V_1, V_3\}$, we have $(\{V_2\}, \{V_1, V_3\})$ follows the GIN condition. This implies that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the same latent parent L_1.

$\mathcal{R}2$. In subgraph (b), $\mathbf{C}_1$ is a pure cluster and $\mathbf{C}_2$ is an impure cluster. Let $\mathcal{A} = \{V_1, ..., V_4\}$. For any variable in $\mathbf{C}_1$, e.g., V_1, and for any variable in $\mathbf{C}_2$, e.g., V_3, $(\{V_2\}, \{V_1, V_3\})$ follows the GIN condition. This implies that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the common parent.

$\mathcal{R}3$. In subgraph (c), $\mathbf{C}_1$ and $\mathbf{C}_2$ are two impure cluster. Let $\mathcal{A} = \{V_1, ..., V_6\}$. For any subset of $\mathbf{C}_1 \cup \mathbf{C}_2$, e.g., $\tilde{\mathbf{C}} = \{V_1, V_3\}$, $(\{V_5, V_6\}, \{V_1, V_3\})$ follows the GIN condition. This implies that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the common parent.

We next analyze subgraphs (d) $\sim$ (f) by Corollary 1, where L_1 is a latent variable introduced in the first iteration, $\mathbf{C}_1$ is a subset of its children, $\mathbf{C}_2$ is a new causal cluster, and $\mathcal{A} = \{L_1, V_3, V_4\}$. For subgraph (d), we first set $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1 \backslash L_1 =$

$\{V_1, ..., V_4\}$. Then, we check $\mathcal{R}1$ and see that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the common parent; For subgraph (e), we first set $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1 \backslash L_1 = \{V_1, ..., V_4\}$. Then, we check $\mathcal{R}2$ and obtain that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the common parent; For subgraph (f), we first set $\mathcal{A} = \mathcal{A} \cup \mathbf{C}_1 \backslash L_1 = \{V_1, ..., V_6\}$. Then, we check $\mathcal{R}3$ and find that $\mathbf{C}_1$ and $\mathbf{C}_2$ share the common parent.

D. Illustration of LaHME Algorithm

In this section, we illustrate our LaHME algorithm with the graph in Figure 13(a).

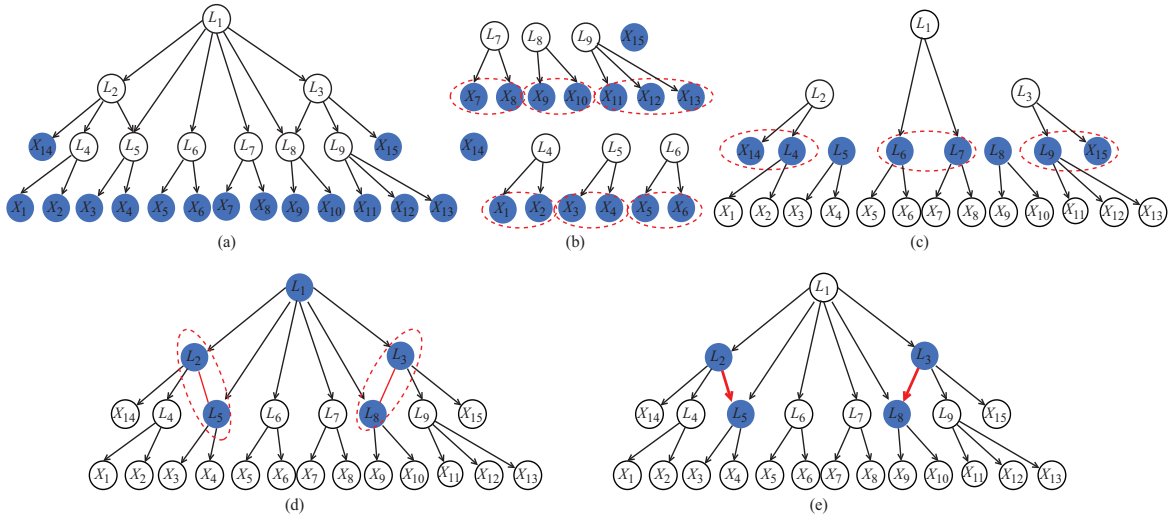


Figure 13. Illustration of the entire procedure of LaHME. Solid blue nodes indicate the active variable set $\mathcal{A}$ for the current iteration. (a) Ground-truth structure. (b) Output after the first iteration of Step 1 of LaHME. Red circles indicate the selected clusters. (c) Output after the second iteration of Step 1 of LaHME. (d) Output after the third iteration of Step 1 of LaHME. (e) Output after Step 2 of LaHME, which has recovered the true causal structure.

We assume oracle tests for GIN conditions. In this structure, $L_1, ..., L_9$ are latent variables and $X_1, ...X_{15}$ are observed variables. The estimating process is as follows:

1. LaHME first initializes active variable set $\mathcal{A} = \{X_1, ...X_{15}\}$ and graph $\mathcal{G} = \emptyset$.

2. After initialization, it runs *FindGlobalCausalClusters* (Phase 1) and learns 8 clusters, i.e., $\mathbf{C}_1 = \{X_1, X_2\}$, $\mathbf{C}_2 = \{X_3, X_4\}$, $\mathbf{C}_3 = \{X_5, X_6\}$, $\mathbf{C}_4 = \{X_7, X_8\}$, $\mathbf{C}_5 = \{X_9, X_{10}\}$, $\mathbf{C}_6 = \{X_{11}, X_{12}\}$, $\mathbf{C}_7 = \{X_{11}, X_{13}\}$, and $\mathbf{C}_8 = \{X_{12}, X_{13}\}$. Next, it runs *DetermineLatentVariables* (Phase 2) and merges $\mathbf{C}_6$, $\mathbf{C}_7$ and $\mathbf{C}_8$ into one cluster by using $\mathcal{R}1$ of Proposition 2. It obtains six clusters and introduces six latent variables for them: $\mathbf{Pa}(\{X_1, X_2\}) = L_4$, $\mathbf{Pa}(\{X_3, X_4\}) = L_5$, $\mathbf{Pa}(\{X_5, X_6\}) = L_6$, $\mathbf{Pa}(\{X_7, X_8\}) = L_7$, $\mathbf{Pa}(\{X_9, X_{10}\}) = L_8$ and $\mathbf{Pa}(\{X_{11}, X_{12}, X_{13}\}) = L_9$, as shown in Figure 13(b).

3. Next, it updates the active variable set $\mathcal{A} = \{\mathbf{X} \cup \{L_4, ..., L_9\} \setminus \{X_1, ..., X_{13}\}\} = \{X_{14}, X_{15}, L_4, ..., L_9\}$, where the values of latent variables $L_4, ..., L_9$ are the values of $X_1, X_3, X_5, X_7, X_9, X_{11}$, respectively.

4. It runs the second iteration of Step 1. Specifically, it runs *FindGlobalCausalClusters* and finds three clusters, i.e., $\mathbf{C}_1 = \{X_{14}, L_4\}$, $\mathbf{C}_2 = \{L_6, L_7\}$ and $\mathbf{C}_3 = \{L_9, X_{15}\}$. Next, it introduces three latent variables for them by running *DetermineLatentVariables*, i.e., $\mathbf{Pa}(\{X_{14}, L_4\}) = L_2$, $\mathbf{Pa}(\{L_6, L_7\}) = L_1$ and $\mathbf{Pa}(\{L_9, X_{15}\}) = L_3$, as shown in Figure 13(c).

5. It updates the active variable set $\mathcal{A} = \{L_1, L_2, L_5, L_3, L_8\}$, where the values of latent variables L_1, L_2, L_5, L_3, L_8 are the values of $X_5, X_{14}, X_3, X_{15}, X_9$, respectively.

6. Analogously, in the third iteration, it identifies two clusters $\{L_2, L_5\}$ and$\{L_3, L_8\}$. Next, it runs *DetermineLatentVariables* and finds that the introduced L_1 is the parent of $\{L_2, L_5\}$ and$\{L_3, L_8\}$, as shown in Figure 13(d).

7. Since there is no new latent variable, Step 1 of LaHME stops.

8. Finally, LaHME performs Step 2 as follows: for impure cluster $\{L_2, L_5\}$, it runs *LocallyInferCausalStructure* and finds that $\{L_2\}$ is a local root set, i.e., $L_2 \succ L_5$ and there exists the directed edge between L_2 and L_5. Similarly, for impure cluster $\{L_3, L_8\}$, it runs *LocallyInferCausalStructure* and finds that $\{L_3\}$ is a local root set, i.e., $L_3 \succ L_8$, and and there exists the directed edge between L_3 and L_8

9. Since there is no one impure cluster, Step 2 of the LaHME algorithm stops. The unknown latent structure is fully reconstructed, as given in Figure 13(e).

E. More Results of Experiments

We here consider graphs with different scales (numbers of variables and depths) and reported the results in Table 4 below. When the number of latent variables increases, it is generally harder to identify the structure, which is the typical case for causal discovery algorithms.

Table 4. Results with different numbers of latent variables (with sample size=10k).

#.Obs. (#.Lat.)	Depth	Structure Recovery Error Rate	Error in Hidden Variables	Correct-Ordering Rate	Running time
4(6)	3	0.00	0.00	1.00	35 sec.
10(12)	4	0.10	0.10	0.98	265 sec.
22(24)	5	0.40	0.60	0.95	2486 sec.

F. More Details of Real-World Data

The data set used in Himi et al. (2019) contains variables that form multitasking behavior model (including four factors), executive functions model (three factors), and predictor variables (three factors). In our experiment, we only considered the multitasking behavior model which includes four factors, as it satisfies the conditions for identifiability.

The details of the hypothesized factors in a multitasking behavior model are shown in Table 5 (Himi et al., 2019).

Latent Factors	Children (Indicators)
Speed (S)	Correctly marked Numbers (S1), Correctly marked Latters (S2), and Correctly marked Figures (S3)
Error (E)	Errors marking Numbers (E1), Errors marking Latters (E2), and Errors marking Figures (E3)
Question (Q)	Correctly answered Questions Par.1 (Q1), Correctly answered Questions Par.2 (Q2), and Correctly answered Questions Par.3 (Q3)
Multitasking behavior (Mb)	Speed, Error, and Question

Table 5. Details of the multitasking behavior data set